DOI: 10.1016/j.jeem.2017.10.003

The link between response time and preference, variance and processing heterogeneity in stated choice experiments[☆]

Danny Campbell[a,*], Morten Raun Mørkbak[b], Søren Bøye Olsen[c]

[a] *Economics Division, Stirling Management School, University of Stirling, Scotland*
[b] *Incentive, Copenhagen, Denmark*
[c] *Department of Food and Resource Economics, University of Copenhagen, Denmark*

Abstract

In this article we utilize the time respondents require to answer a self-administered online stated preference survey. While the effects of response time have been previously explored, this article proposes a different approach that explicitly recognizes the highly equivocal relationship between response time and respondents' choices. In particular, we attempt to disentangle preference, variance and processing heterogeneity and explore whether response time helps to explain these three types of heterogeneity. For this, we divide the data (ordered by response time) into approximately equal-sized subsets, and then derive different class membership probabilities for each subset. We estimate a large number of candidate models and subsequently conduct a frequentist-based model averaging approach using information criteria to derive weights of evidence for each model. Our findings show a clear link between response time and utility coefficients, error variance and processing strategies. Our results thus emphasize the importance of considering response time when modeling stated choice data.

Keywords: choice experiments, response time, preference heterogeneity, scale-adjusted latent class, independent availability logit, processing strategies, multimodel inferences, frequentist-based model averaging, willingness to pay

1. Introduction

When collecting stated preference data, online surveys have become more and more important and widespread. One reason for this is due to improvements in computer technology, but also the increased availability of the internet. Yet another reason for this increasing popularity stems from the fact that online surveys have a number of advantages over more traditional survey modes, such as mail-out paper-and-pen questionnaires, personal interviews and telephone interviews. Advantages typically mentioned in the resource economics literature (cf. Lindhjem and Navrud, 2011a,b; Fleming and Bowden, 2009; Olsen, 2009) are reduced costs, increased speed of data collection, less item non-responses, ability to adjust questionnaires according to respondent answers on-the-fly, potential for broader stimuli in terms of graphics and sound, and avoidance of manual data entry mistakes. While advantages are many, this literature has also highlighted a few important potential disadvantages, which raise concerns regarding data quality and their suitability in non-market valuation (Lindhjem and Navrud, 2011a,b). In particular, these disadvantages relate to problems concerning sample coverage and representativeness and self-selection bias. While not unique to online surveys, there can be so-called "pure survey mode effects", whereby a respondent provides different answers to otherwise identical questions only because it is administered through different survey modes.

This article focuses on one aspect of online surveys—the length of time respondents take to complete the choice experiment. The concern is that, notwithstanding the fact, as pointed out by Cook et al. (2012), that online surveys allow respondents "time to think" and reflect, an interviewer is not present to pace the respondent. For this reason, some respondents may not exert the level of cognitive effort needed to answer the questions in any meaningful way. While this concern also applies to other self-administered methods of data collection, understanding the role of response time in online surveys is especially important because of the incentives that respondents often obtain for their continued participation in such surveys. Furthermore, as respondents within pre-recruited online panels gain experience, the tendency to answer quickly may actually increase (Malhotra,

[☆] Acknowledgments to be included.

[*] Corresponding author.

Email addresses: danny.campbell@stir.ac.uk (Danny Campbell), mm@incentive.dk (Morten Raun Mørkbak), sobo@ifro.ku.dk (Søren Bøye Olsen)

Paper accepted for publication in the *Journal of Environmental Economics and Management* (23 October 2107)

2008). Consequently, with online surveys we may in fact increase the risks of panel attrition effects and surveying of experienced respondents—whose primary motivation for participating stems from the reward (either monetary or non-monetary) they receive—who answer so quickly that their choices do not reflect their actual preferences for the good in question. If this is indeed the case, there are obvious implications since it challenges the validity of any inferences made from the observed choices. Therefore, there may be reasons to be skeptical of "quick-and-dirty" responses, as coined by Schwappach and Strasmann (2006) and Olsen (2009), collected from online surveys in which participants are recruited and motivated by an incentive for completing the survey (Bonsall and Lythgoe, 2009). Similarly, concerns of reliability may be warranted for respondents who require significantly more time than would be reasonably expected. This could signal that these respondents faced distractions or were multitasking and, thus, did not give the choice decisions their utmost attention. In spite of these issues, this subject has yet to receive much attention, which gives rise to the present study.

Although the importance of response time has received considerable attention within experimental psychology, consumer research and marketing research (e.g., see Haaijer, Kamakura, and Wedel, 2000; Luce, 1986; Rubinstein, 2007), there are relatively few investigations within the stated preference literature (see Holmes et al., 1998; Haaijer, Kamakura, and Wedel, 2000; Rose and Black, 2006; Otter, Allenby, and van Zandt, 2008; Brown et al., 2008; Bonsall and Lythgoe, 2009; Vista, Rosenberger, and Collins, 2009; Hess and Stathopoulos, 2013; Börger, 2015; Börjesson and Fosgerau, 2015; Campbell, Mørkbak, and Olsen, 2016, for applications). Though, as can be seen from the above references, interest in this topic has clearly increased recently, and Bonsall and Lythgoe (2009) note that there is considerable scope for more research. Our article is intended to contribute to this area and provide a robust modeling framework for practitioners engaged in analyzing stated choice experiments. Unlike the articles mentioned above, which have established that response time has a significant bearing on the estimates of utility coefficients, error variance, model fit and predictions, we are interested in identifying the link between the length of time respondents require to answer the choice experiment and their preferences, variances and processing strategies (specifically, choice set generation) in a simultaneous estimation. While establishing the link between response time and any one of these types of heterogeneity is relatively straightforward, tackling all three simultaneously poses a challenge. Nevertheless, when only one type of heterogeneity is accounted for, there is a potential risk that the actual heterogeneity among respondents is only partially explained and, in fact, may actually be an artifact of another (unmodeled) type. For this reason, attempts to accommodate more than one type of heterogeneity at the same time would seem justified. In Campbell, Mørkbak, and Olsen (2016) we accommodate two types of heterogeneity, namely preference and variance heterogeneity, while leaving out any heterogeneity in the processing strategies adopted. With respect to response time, this latter type of heterogeneity seems highly relevant, in the sense that response time may be highly correlated with the use of any decision heuristic. Thus, in this article, while acknowledging the difficulty in separating the three types of heterogeneity simultaneously, we use latent class modeling to separately identify the different types of heterogeneity within the sample of respondents and to explore whether the class memberships differ by response time. This represents a step forward in the analysis of heterogeneity. This is the first article to explore the link between response time and processing strategies in the form of consideration set formation. Moreover, unlike Thiene, Scarpa and Louviere (2015), who also disentangle three different types of heterogeneity, this is the first attempt to investigate the connection between each type of heterogeneity and response time.

In our earlier study (Campbell, Mørkbak, and Olsen, 2016), which was based on a survey investigating recreational anglers' preferences for fishing sites, we set out to identify fast and slow respondents, but ended up concluding that identifying the thresholds is fraught with difficulty and, therefore, found no justification to drop respondents from the analysis on the basis of response time. In the present paper, we reiterate this conclusion but suggest a potential solution by approaching the issue from a different perspective. Rather than identify response time thresholds, this paper focuses on the differences between the respondents and we address the issue by not settling on an unique 'best' model, but instead 'averaging' over a bunch of models through the use of a multimodel inference approach. Our analysis is based on the division of the data (ordered by response time) into approximately equal-sized subsets. Crucially, for each subset we retrieve separate class sizes, which is fundamental for a meaningful investigation of response time. Under this framework, we are in a better position to distinguish the respondents who answered quickly and relatively inconsistently from those who also answered quickly but in a more consistent manner and, in the same vein, between respondents who took longer because they did not give the survey their full attention and those who took longer because they more fully evaluated the information presented to them. This is an important contribution of our work.

Our analysis considers 90 candidate model specifications to test for the number of preference classes, error variance classes, processing strategies and the number of different subsets based on response time. With so many competing model specifications, there is an inherent uncertainty of the true model. Given this uncertainty, and the fact that each of our different models provide different relative statistical fits, it does not seem sensible to ultimately select only one model. Instead, when considering the range of models we can use their relative statistical fits to form weights, so that weighted average estimates can be derived. For this, we conduct a frequentist-based (as opposed to Bayesian) model averaging approach (Buckland, Burnham, and Augustin, 1997; Layton and Lee, 2006). The motivation behind this multimodel inference is that it allows judgments to be made regarding the relative suitability of each of our models. This proves to be a useful exercise, since it resolves the usual uncertainty by averaging over the set of candidate models.

To test our approach we use a stated choice experiment dataset that was collected via an online survey. This was administered to a pre-recruited panel of Danes and had the aim of establishing their willingness to pay (WTP) for different product attributes of honey. We do, however, emphasize that even though our analysis is based on a food application, the impact of response time and the modeling framework introduced should be of wider interest to other fields using choice experiments. Results from our analysis provide further evidence that preferences and the variance of the observed factors are sensitive to response time. We find that for some attributes, marginal WTP estimates are smaller among respondents with a longer response time, while for other attributes we find the reverse. Opposite to what might be expected, we show that the average level of measurement error actually increases with response time. Importantly, we shed light on the fact that the processing strategies adopted by respondents in different response time quantiles are not the same, with respondents who answer quickest most likely to consider the deterministic choice set. Our analysis highlights the relevance of accommodating for the three types of heterogeneity concurrently. We especially find that the confidence set of models (i.e., the subset of models that represent the majority of evidence) include three latent segments of preferences. Nevertheless, our analysis provides strong evidence for the need to address all three types of heterogeneity simultaneously, as this leads to higher average weights of evidence.

The remainder of the article is structured as follows. In the next section, the modeling approach to investigate the role of response time on preferences, variance and processing strategies is developed. Following that, an outline of our empirical case-study is provided. In the subsequent sections we present, results from the analysis and a general discussion, and an overall conclusion.

2. Modeling approach

The point of departure is the conventional utility specification, where respondents are indexed by n, chosen alternatives by i, choice occasions by t and the attributes by x:

$$U_{nit} = \beta x_{nit} + \eta_i + \varepsilon_{nit}, \tag{1}$$

where β are estimated marginal utility parameters for the attributes, η are alternative specific constants (subject to at least one being fixed), ε is an *iid* type I extreme value (EV1) distributed error term with variance $\pi^2/6\lambda^2$, and where λ is a scale parameter. Under these conditions the probability of respondent n's sequence of choices is given by the standard multinomial logit (MNL) model:

$$\Pr(y_n) = \prod_{t=1}^{T_n} \frac{\exp\left(\lambda\left(\beta x_{nit} + \eta_i\right)\right)}{\sum_{j=1}^{J} \exp\left(\lambda\left(\beta x_{njt} + \eta_j\right)\right)}, \tag{2}$$

where y_n is the sequence of choices for respondent n, one for each choice occasion, $i_n = [i_{n1}, i_{n2}, \dots, i_{nT_n}]$.[1]

The major advantage of the specification outlined in Eq. 2 is its simple form for choice probabilities. However, this is based on the well known assumption of preference homogeneity between respondents, which by now is widely acknowledged to be inferior to those that facilitate heterogeneity in preferences. Another issue is, for identification purposes, that the value of λ is generally set to unity meaning that it drops out of the probability calculation. But, in cases where it is believed that there is heterogeneity in the variance of the unobserved factors among respondents, this is obviously inappropriate. The estimation of separate scale parameters may, therefore, also be warranted. Finally, the above model also assumes deterministic choice sets, meaning that it is expected that all consumers consider all options presented to them. Resent evidence, however, has revealed that some respondents adopt processing strategies and other simplifying heuristics when making their choices—including considering only a subset of the available options (cf. Campbell, Hensher, and Scarpa, 2014; Thiene, Swait and Scarpa, 2016)—meaning that the deterministic assumption may also not be justified. Accommodating this type

[1] We note that accommodating the panel effect is not necessary for the MNL model, since the choice probabilities are independent. We feel that by presenting the MNL model in this manner we introduce as many of the necessary terms as early on as possible, which makes it easier to see the differences introduced in subsequent models.

of heterogeneity is also likely to be beneficial as it should help ensure that the choice models are not biased by aspects that had no bearing on respondents' choices (Campbell and Erdem, 2015). This type of heterogeneity is also very likely to be driven/associated with response time, which makes it even more important to accommodate in light of the objective of the present paper.

For these reasons, we move to specifications that accommodate heterogeneity. In this article, however, we provide the first study to explore these three specific types of heterogeneity concurrently. As a further advancement, we also explore this in the context of response time. Our rationale for this is the fact that response time may give an indication of random decision-making and/or the adoption of simplifying heuristics. The argument here is that this is likely to be exhibited in the preference structures, error variance and/or processing strategies. This motivates the present study on how to appropriately identify and accommodate these issues. In this article, we use the latent class modeling approach, which we outline below.

2.1. Preference heterogeneity

We are interested in explaining the heterogeneous nature of preferences for attributes among the sample of respondents. Such (unobserved) preference heterogeneity can be accommodated by assuming random distributions. Rather than a continuous random distribution, we opt for a finite one. The advantage of this non-parametric approach is that, compared to the commonly used continuous distributions, they are not constrained by distributional assumptions (Train, 2009). In non-parametric estimation, an approximating family of distributions is used, where the family has the property that the accuracy of the approximation rises with the number of parameters. By allowing the number of parameters to rise with sample size, nonparametric estimators are consistent for any true distribution. Finite distributions can not only provide greater flexibility but also have practical appeal as the results typically have a more intuitive meaning than the parameters and moments of the distributions retrieved from continuous parametric distributions.[2] We specify such a latent class model, as follows:

$$\Pr\left(y_n\right) = \sum_{a=1}^{A} \pi_{a_n} \prod_{t=1}^{T_n} \frac{\exp\left(\lambda\left(\beta_a x_{nit} + \eta_{a_i}\right)\right)}{\sum\limits_{j=1}^{J} \exp\left(\lambda\left(\beta_a x_{njt} + \eta_{a_j}\right)\right)}, \tag{3a}$$

where it is assumed that respondents can be identified as belonging to a specific latent class, a, each of which differs with respect to the β and η parameters, hence, denoted by β_a and η_a, and where λ is, again, constrained to unity for identification purposes.

Given our interest in response time, we also wish to explore whether the unconditional class membership probabilities differ according to response time. For this, we divide the data (ordered by total response time) into approximately equal-sized subsets.[3] Each subset is associated with different class membership parameters, thus enabling the class sizes to be different. Specifically, we define the unconditional probability of a respondent belonging to class a conditional on their response time falling within the k^{th} q-quantile as a constant only MNL model:

$$\pi_{a_n|k} = \frac{\exp\left(\omega_{a_k}\right)}{\sum\limits_{a=1}^{A} \exp\left(\omega_{a_k}\right)}, \tag{3b}$$

where ω_{a_k} denotes the constant corresponding to latent class a conditional on the k^{th} q-quantile, and where at least one constant within each k q-quantiles is constrained to be zero for identification purposes. The attraction of this specification is that, in addition to identifying latent classes of respondents based on their preferences, we can assess the influence of response time on membership to the latent segments.[4]

[2] We do note, however, that a continuous representation could have been used. However, we favor the appeal of finite distribution, but suggest that this is potentially an interesting future extension to this modeling approach.

[3] We note that we use paradata relating to the total response time of the panel of choice tasks. Of course, the time associated with each choice task could instead be used, but in our latent class models we are interested in explaining class membership at the panel (i.e., individual) level rather than cross section (i.e., observation) level. We also note that the total response time averages out idiosyncrasies unique to each task and is, arguably, a better construct of overall attention (cf. Malhotra, 2008). Since the time respondents spend on making their choices generally drops as they progress through the experiment (cf. Haaijer, Kamakura, and Wedel, 2000; Rose and Black, 2006), this also helps to disentangle the issue from the potential effects of learning and fatigue discussed in Campbell et al. (2015).

[4] We are mindful that our deterministic inclusion of response time in the class membership functions may be considered as a limitation compared to a latent variable approach. However, in this paper we are specifically interested in the link between response time (as opposed to a latent variable of survey engagement) and heterogeneity. For this reason, we choose to include response time in a deterministic fashion. Readers interested in a latent variable application of response time are directed to Hess and Stathopoulos (2013).

2.2. Variance heterogeneity

Despite the attraction of assessing the influence of response time on preferences, choices made very quickly may have a higher variance compared to those that were deliberated over a longer period, hence the potential label "quick-and-dirty" for the faster responses. In this article, we explore a specification where a distribution in the scale parameter is facilitated. Following our findings in Campbell, Mørkbak, and Olsen (2016), we recognize that the length of time required by respondents to make a well-balanced choice varies across individuals. For this reason, there are likely to be instances where the variances of the unobserved factors are not associated with response time. Therefore, a probabilistic approach for identifying heterogeneity in variances would seem justified. This is likely to offer a more flexible solution as the differences in variance for a specific response time are associated with a probability.

In an attempt to uncover and explain heterogeneity in variances, we again make use of the latent class modeling framework. Specifically, we implement a variant of the scale-adjusted latent class modeling approach outlined in Magidson and Vermunt (2008) and executed in Campbell, Hensher, and Scarpa (2011) and Campbell et al. (2015), whereby each latent class is described by a class-specific representation of scale:

$$\Pr(y_n) = \sum_{b=1}^{B} \pi_{b_n} \prod_{t=1}^{T_n} \frac{\exp(\lambda_b(\beta x_{nit} + \eta_i))}{\sum_{j=1}^{J} \exp\left(\lambda_b\left(\beta x_{njt} + \eta_j\right)\right)}, \tag{4a}$$

where it is now assumed that respondents belong to different latent classes, b, that differ with respect to the λ parameters. We note that, for identification purposes, we set $\lambda_{b=1} = 1$. Since class membership is latent, the unconditional probability of membership associated with class b, is given by:

$$\pi_{b_n|k} = \frac{\exp(\omega_{b_k})}{\sum_{b=1}^{B} \exp(\omega_{b_k})}, \tag{4b}$$

where again, ω_{b_k} are constants estimated for each of the k q-quantiles. Again, at least one constant within each k q-quantiles is subject to the same zero constraint, meaning that probabilistic estimates of the differences in variances can be uncovered for each of the k q-quantiles of response time.

2.3. Processing heterogeneity

While it is possible that respondents who answered relatively quickly processed all of the information in the choice tasks, it is also conceivable that at least some of them adopted some form of decision-making heuristic. Failing to account for this is likely to be suboptimal, and perhaps lead to misguided inferences, as the model does not reflect actual choice behavior. For this reason we extend our investigation to incorporate the heterogeneity in the processing strategies adopted by respondents. We recognize that there are many types of processing strategies, attribute non-attendance and attribute aggregation where they share a common metric (e.g., see Hensher, 2010, for a comprehensive overview), however, we choose to focus on the processing of alternatives since it is arguably the most convenient heuristic for respondents to adapt in order to eliminate or reduce cognitive burden. Specifically, rather than rely on the assumption that respondents considered all alternatives, we acknowledge that they may have considered only a subset of alternatives.

Following Manski (1977), a probabilistic model can be formulated to model this type of behavior to help distinguish between the deterministic choice set, as generated by the experimental design, and the respondent's actual consideration set. For this type of analysis we extend the independent availability logit (cf. Swait and Ben-Akiva, 1987; Swait, 2001; Frejinger, Bierlaire, and Ben-Akiva, 2009; Kaplan, Shiftan, and Bekhor, 2012; Richardson, 1982; Ben-Akiva and Boccara, 1995; Chang, Lusk, and Norwood, 2009, for examples). The probability of choice in the independent availability logit model is given by:

$$\Pr(y_n) = \sum_{c=1}^{C} \pi_{c_n} \prod_{t=1}^{T_n} \Pr(y_n|S_c), \tag{5a}$$

where $\Pr(y_n|S_c)$ is the conditional probability of the sequence of choices given the choice set is $S_c \subseteq C$, C is the set of subsets, π_c is the probability that S_c is the 'true' choice set. Since a respondent's true consideration set cannot be known with certainty, this model assumes that choice sets are latent, and the conditional choice model is MNL:

$$\Pr(y_n|S_c) = \frac{\exp(\lambda(\beta x_{nit} + \eta_i))}{\sum_{j \in S_c} \exp\left(\lambda\left(\beta x_{njt} + \eta_j\right)\right)}. \tag{5b}$$

We note that the size of C grows exponentially as a function of the number of alternatives (e.g., for a universal set with J alternatives, 2^J possible choice sets need to be taken into account (including the situation where none of the alternatives were taken into account, as would be the case under random decision-making)). As noted above, the alternatives taken into account by a respondent cannot be known with certainty. However, their observed choice behavior helps make probabilistic statements about the likelihood (i.e., π_{c_n}) of competing consideration sets being their true choice set. Moreover, since respondents who answered very quickly are unlikely to have attended to all alternatives when making their choice, response time is also likely to help identify the consideration sets. For this reason, we again retrieve class membership probabilities conditional on the k^{th} q-quantiles of response time:

$$\pi_{c_n|k} = \frac{\exp\left(\omega_{c_k}\right)}{\sum\limits_{c=1}^{C} \exp\left(\omega_{c_k}\right)}. \tag{5c}$$

2.4. *Accounting for more than one type of heterogeneity*

The above specifications provide a first step at looking towards the three types of heterogeneity as well as the role that response time plays. However, each assumes that only one type of heterogeneity is at play. This may be considered as a somewhat stringent assumption, since it is conceivable that the variations across respondents are not confined to one type. Despite this, the majority of discrete choice analysis addresses only one aspect of heterogeneity, and relatively few studies explore two concurrently. To the best of our knowledge, no article has yet addressed preference heterogeneity, variance heterogeneity and choice set formation simultaneously.[5] In this article we attempt to tackle this. To do this we expand the previous latent class models, as follows:

$$\Pr\left(y_n\right) = \sum_{z=1}^{Z} \pi_{z_n} \prod_{t=1}^{T_n} \frac{\exp\left(\lambda_z\left(\beta_z x_{nit} + \eta_{z_i}\right)\right)}{\sum\limits_{j \in S_c} \exp\left(\lambda_z\left(\beta_z x_{njt} + \eta_{z_j}\right)\right)}, \tag{6a}$$

where each of the z classes now describes a particular structure of preferences, variances and processing strategy.

We are cognizant of issues of confounding between the different types of heterogeneity, which makes it difficult to separately identify each type (Hess and Rose, 2012; Hess and Train, 2017). In this article we circumvent this by setting $Z = A \times B \times C$, using equality constraints (cf. Scarpa et al., 2009, for further details) on the class parameters and specifying the unconditional class probability for a specific combination as the product of the associated preferences, scale parameter and processing strategy unconditional probabilities obtained for each k q-quantiles of response time:[6]

$$\pi_{z_n|k} = \pi_{a_n|k}\pi_{b_n|k}\pi_{c_n|k}. \tag{6b}$$

As an example, with $A = 2$, $B = 2$ and $C = 2$, we would have $Z = 8$, and the parameters within each class would be restricted as follows:

$$Z = \begin{cases} \text{class } z_1 \text{ relates to the case } \beta_{a_1}, \eta_{a_1}, \lambda_{b_1} \text{ and } S_{c_1}; \\ \text{class } z_2 \text{ relates to the case } \beta_{a_1}, \eta_{a_1}, \lambda_{b_1} \text{ and } S_{c_2}; \\ \text{class } z_3 \text{ relates to the case } \beta_{a_1}, \eta_{a_1}, \lambda_{b_2} \text{ and } S_{c_1}; \\ \text{class } z_4 \text{ relates to the case } \beta_{a_1}, \eta_{a_1}, \lambda_{b_2} \text{ and } S_{c_2}; \\ \text{class } z_5 \text{ relates to the case } \beta_{a_2}, \eta_{a_2}, \lambda_{b_1} \text{ and } S_{c_1}; \\ \text{class } z_6 \text{ relates to the case } \beta_{a_2}, \eta_{a_2}, \lambda_{b_1} \text{ and } S_{c_2}; \\ \text{class } z_7 \text{ relates to the case } \beta_{a_2}, \eta_{a_2}, \lambda_{b_2} \text{ and } S_{c_1}; \\ \text{class } z_8 \text{ relates to the case } \beta_{a_2}, \eta_{a_2}, \lambda_{b_2} \text{ and } S_{c_2}.^{7} \end{cases} \tag{7}$$

Note that the model outlined in Eq. 6 is a fully general model for examining heterogeneity. For instance, no heterogeneity is accommodated, as in the MNL model when A, B and C are equal to 1; one type can be uncovered

[5] We recognize the work of Thiene, Scarpa and Louviere (2015), who also accounted for three types of heterogeneity. However, their study explored attribute non-attendance rather than choice set formation. We also acknowledge the potential identification issues associated with explaining the three types of heterogeneity simultaneously.

[6] While it is convenient to interpret any differences in the class-specific marginal utilities uncovered from a standard latent class model as preference heterogeneity, some of the differences may be induced by random scale. Similarly, when interpreting the within class scale heterogeneity uncovered from a scale adjusted latent class model, it is important to recognize that the estimated scale parameter incorporates not just differences in variance but also other sources of within-class correlation that might exist. See Hess and Train (2017) for a thorough discussion. However, the use of equality constraints means that we can distinguish between the different types of heterogeneity assumming all else remains constant.

in isolation when either A, B or C is greater than 1; two types can be retrieved when either A, B or C is equal to 1 and the remaining two are greater than 1; and, finally, the case where all three types of heterogeneity are tackled simultaneously is when A, B and C are all greater than 1. Therefore, for this example, given the correct equality constraints across specific classes it is possible to come up with eight model forms, depending on the assumptions of heterogeneity in respondent's preferences, or tastes, variances and processing strategies.

2.5. Model estimation and multimodel inference

All models are coded and estimated using the maxlik library in R (see Henningsen and Toome (2011) and R Core Team (2014) for further details) using maximum likelihood estimation. Since our models retrieve class probabilities, we are mindful of their vulnerability to local maxima of the simulated sample-likelihood function. Thus, in an attempt to reduce the possibility of reaching a local rather than a global maximum, we started the estimation iterations from a variety of random starting points. Specifically, we achieved this by estimating these models many times, but each time using a different vector of starting values, which are chosen randomly.

An important consideration relates to the number of classes to accommodate for preferences (i.e., A), variance (i.e., B) and processing strategies (i.e., C) as well as the number of q separate class membership probabilities to retrieve. Of course, this remains an empirical decision and should be considered on a case-by-case basis, requiring discretion and objective judgment on behalf of the analyst. Choosing a model with too few classes and/or q-quantiles of response time may involve making unrealistically simple assumptions and lead to considerable bias, poor prediction, and missed opportunities for insight. The concern is that such models may not be flexible enough to describe the sample or the population well. Yet, any increase in the number of classes and/or q-quantiles leads to a proliferation of parameters and, therefore, loss of parsimony. While this will better ensure the observed data is fitted well, it comes at the risk of being tailored too closely to the data, which compromises the ability to generalize the model beyond the existing dataset.

For our analysis we consider models that include up to three latent classes for preferences (i.e., $1 \leq A \leq 3$) and scale (i.e., $1 \leq B \leq 3$). For processing heterogeneity we consider models that accommodate up to two consideration sets (i.e., $1 \leq C \leq 2$), including: (i) all alternatives are considered; and, (ii) only non-status-quo options are considered.[8] Therefore, our models accommodate up to 18 latent classes of respondents (i.e., $Z = A \times B \times C = 3 \times 3 \times 2 = 18$), which we found to be sufficient whilst ensuring model tractability. Finally, we estimate these latent class membership probabilities for up to five equal-sized data subsets sorted by response time (i.e., $1 \leq q \leq 5$). Ultimately, we consider all 90 candidate models (i.e., 18×5) to accommodate different numbers of preference classes, error variance classes, processing strategies and response time quantiles.

With the estimation of such a large number of candidate models, model selection uncertainty becomes a concern. Each model is likely to produce information that is not captured by the best model. Given this uncertainty, and the fact that each of our different models provide different relative statistical fits, it does not seem sensible to ultimately select only one model. Indeed, as demonstrated in Layton and Lee (2006) in a stated preference study, basing inference only on the model estimated to be the best could be considered as poor practice. In this regard, penalized-likelihood information criteria, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), may serve as a useful guide for model selection. In this paper, we consider the adjusted sample size corrected BIC (BIC*), which we estimate for each respondent:

$$\text{BIC}_n^* = -2LL\left(\hat{\beta}\right)_n + \left(\ln\left(T_n + 2\right) - \ln\left(24\right)\right) K \tag{8}$$

where $LL\left(\hat{\beta}\right)_n$ is respondent n's contribution to the log-likelihood of the estimated model, K is the number of estimated parameters and T_n is the number of choice observations completed by respondent n. Note that the sample-size correction serves to reduce the sample size penalty, which should lead to better performance in our case since the number of parameters is large relative to the number of choice observations per respondent. For further background on this information criterion see Yang and Yang (2007).

When considering our range of models, many of which may fit equally well, it is arguably more appropriate to derive weights of evidence for each model (which can be considered as analogous to the probability that a given model is the best approximating model, given the data and set of candidate models). This can be accomplished for each respondent by calculating the difference (Δ_{m_n}) between the information criterion value of the best model and

[8] Other consideration sets were tested, but were not found to be relevant for our empirical case-study. We are also mindful that accommodating all 2^J possible choice sets would include the situation where none of the alternatives are taken into account, which would make it impossible to distinguish between the variance and processing heterogeneity classes. For this reason, a normalisation factor is commonly included in the class membership function of an independent availability logit model to account for this (see Swait, 2001, for details). Similarly, in our paper, we do not include the situation where all alternatives are ignored, which avoids this confounding.

the information criterion value for model m:

$$\Delta_{m_n} = \text{BIC}^*_{m_n} - \text{BIC}^*_{\min_n}, \tag{9}$$

where $m = 1, 2, \ldots, m$, with M being the number of models (i.e., $M = 90$ in our case), and $\text{BIC}^*_{\min_n}$ is the smallest value of BIC^*_n in the model set, M. The term Δ_{m_n} is a calibration of model fit, using the best fitting model as the standard. The best fitting model has $\Delta_{m_n} = 0$, and all other models have $\Delta_{m_n} > 0$. Importantly, Δ_{m_n} can be used to calculate two additional measures to assess the relative strengths of each candidate model. The first of these is the evidence ratio, ER_{m_n}:

$$ER_{m_n} = \frac{\exp\left(-0.5\Delta_{\min_n}\right)}{\exp\left(-0.5\Delta_{m_n}\right)}. \tag{10}$$

This provides a measure of how much more likely the best model is compared to model m. The best fitting model has $ER_{m_n} = 1$, and all other models have $ER_{m_n} > 1$. As ER_{m_n} increases, the less confidence we have that the m^{th} model for respondent n is the best approximating model for this respondent.

The second more useful and wider use of Δ_{m_n} is the weight of evidence, WE_{m_n}, which is a probability scaling of Δ_{m_n}:

$$WE_{m_n} = \frac{\exp\left(-0.5\Delta_{m_n}\right)}{\sum\limits_{m=1}^{M} \exp\left(-0.5\Delta_{m_n}\right)}, \tag{11}$$

where the sum is over all models in the set. The scaling is convenient, as the weights range between 0 and 1 and their sum across all models in the candidate set is equal to 1. These can, therefore, be considered as analogous to the probability that a given model is the best approximating model, given the data and set of candidate models. In this paper we use these weights as the basis of frequentist-based (as opposed to Bayesian) model averaging. Specifically, we use them to derive an expected marginal WTP estimate for each respondent, $\mathbb{E}\left(\bar{\text{WTP}}_n\right)$:

$$\mathbb{E}\left(\bar{\text{WTP}}_n\right) = \sum_{m=1}^{M} WE_{m_n} \bar{\text{WTP}}_{m_n}, \tag{12a}$$

where $\bar{\text{WTP}}_{m_n}$ is the mean of the conditional (individual-specific) marginal WTP distribution for respondent n retrieved from model m. Similarly, we use these weights to derive an expected probabilistic estimate for a range of latent class membership probabilities, $\mathbb{E}\left(\bar{\pi}_{z_n}\right)$:

$$\mathbb{E}\left(\bar{\pi}_{z_n}\right) = \sum_{m=1}^{M} WE_{m_n} \bar{\pi}_{m_{z_n}}, \tag{12b}$$

where $\bar{\pi}_{m_{z_n}}$ is the mean of respondent n's conditional class membership distribution for latent class z attained from model m.

The motivation behind this multimodel inference is that it allows judgments to be made regarding the relative suitability of each of our models. Consequently, by regarding various models used in the analysis, we are in a better position to identify appropriate assumptions for accommodating heterogeneity in preferences, variance and processing strategies as well as response time conditional on the data, the set of considered models, and the inability to know the true model. For further details on multimodel inference see, for example, Buckland, Burnham, and Augustin (1997), Layton and Lee (2006) and Symonds and Moussalli (2011).

3. Case-study: willingness to pay for attributes of honey

Data for the present study were gathered using an online stated choice experiment focusing on Danish consumers' preferences and WTP for honey. The method was found to be particularly suitable because of its ability to uncover the relative weighting of the various characteristics of honey, which was confirmed during a series of focus group discussions. The attributes of honey deemed to be of greatest relevance to Danish consumers were identified as 'origin', 'method of production' and 'type of honey'. Furthermore, a cost attribute, with eight different levels, was included to denote the price increase over standard honey. An overview of all attributes and levels are presented in Table 1.

Using a D-efficiency criterion for evaluation, a Bayesian updated experimental design was employed using priors from a pilot study with 104 respondents. The final main effects experimental design consisted of 12 choice tasks, each of which comprised of two generic alternatives, labeled Option A and Option B respectively. A baseline

Table 1. Attributes and levels used in the choice experiment

Attributes	Levels
Origin	Local Denmark European Outside Europe
Method of production	Organically produced Not organically produced
Type of honey	Heather Clover Rape Mixture
Price increase per 450 g jar (DKK)	0, 3, 8, 15, 23, 30, 40 and 55

alternative, labeled as the Status-quo, was also included in the choice tasks which was described as a 450 g jar of conventional mix honey produced outside Europe and priced at DKK 25.

Respondents were sampled from a pre-recruited internet panel between June and August 2010. The sample on which the analysis in this article is based consists of a total of 592 respondents, thus leading to 7,104 observations for model estimation.

4. Results and discussion

We begin this section with summary results of response time. Following this, we present estimation results from our 90 candidate models and our multimodel inference approach.

4.1. Response time

In Table 2, we report summary statistics of response time broken down by quantiles. We find quite a range in response times. As noted by Bonsall and Lythgoe (2009), such a range can be expected since response time varies between individuals, depending on their personal decision-making styles, and is likely to reflect circumstances, such as the extent of any distraction, the time pressure they are under and their current mental and motivational state. The mean response time associated with completing all 12 choice tasks is 5 minutes (i.e., almost an average of 25 seconds per choice task). While around 80 percent of respondents completed the sequence of choice tasks within 6 minutes, a few spent nearly 30 minutes (i.e., almost an average of 2.5 minutes per choice task). While this could, of course, be interpreted as respondents spending a huge effort on answering the choice sets, we cannot rule out that this is due to measurement error in terms of respondents not focusing only on the choice experiment during that period. As response time is measured solely on a click-by-click basis, we have no way of telling whether the respondents faced any distractions or were multitasking when they were completing the choice experiment. At the other end of the spectrum almost 20 percent of respondents completed the 12 choice tasks in less than 3 minutes (which is equivalent to an average of 15 seconds per task). As we corroborate in Campbell, Mørkbak, and Olsen (2016), it is clearly not trivial to determine exactly what the minimum required response time would be in order for respondents to fully evaluate all of the information contained within the choice task. However, there is a concern that the quickest responses may, at best, provide nothing but noise to our survey or, at worst, bias our results in the case where they have not properly considered the information provided when making their choices.

While it may have been possible that respondents who answered relatively quickly processed all of the information in the choice tasks and made a utility maximizing choice, it is also conceivable that they adopted some form of decision-making heuristic, or even made completely random choices. Similarly for respondents who required a relatively long time, their choices may reflect informed decision-making or be due to the fact that they encountered a distraction or were multitasking. These findings motivate our search for flexible latent class models capable of comparing the breakdown of respondents—in terms of their preferences, error variances and processing strategies—across different response times. Ultimately, this should help us distinguish between relatively well informed and reliable choices from those that are less so and, importantly, how the likelihood of these differ according to response time.

Table 2. Response time quantiles (in minutes)

	N	Minimum	Maximum	Mean
Total	592	1.38	28.40	5.00
$q = 2$				
1st	295	1.38	4.22	3.20
2nd	297	4.23	28.40	6.78
$q = 3$				
1st	196	1.38	3.57	2.85
2nd	198	3.58	5.07	4.27
3rd	198	5.08	28.40	7.85
$q = 4$				
1st	148	1.38	3.27	2.66
2nd	147	3.28	4.22	3.75
3rd	148	4.23	5.53	4.85
4th	149	5.55	28.40	8.70
$q = 5$				
1st	119	1.38	3.08	2.53
2nd	113	3.10	3.85	3.45
3rd	122	3.87	4.68	4.24
4th	119	4.72	6.02	5.27
5th	119	6.03	28.40	9.43

4.2. Estimation results

As a point of reference, in Table 3 we present the MNL model (i.e., where $A = B = C = 1$) along with more recognizable heterogeneity models (i.e., where one type of heterogeneity is explored in isolation). From the MNL model we see, as expected, that the cost coefficient is negative and significant, indicating that, all else held constant, respondents are more likely to choose a cheaper honey product compared to one that is more expensive. The marginal utility parameters for the three origin levels are positive and significant, implying that this sample of Danish consumers, on average, prefer honey that originates in Denmark (both locally and nationally) or elsewhere in Europe compared to honey produced outside of Europe. Similarly, the marginal utilities for honey produced organically and from heather are found to be positive and significant. The status-quo alternative specific constant—whose coefficient can be interpreted as the marginal (dis-)utilities relative to the experimentally designed alternatives in the choice task—is negative and significant revealing that, on average, this sample of consumers dislike the baseline honey product.

Moving to the results obtained for Models 2 and 3, which are familiar latent class models where the latent segments are characterized only in terms of preference differences (i.e., where $A > 1$ and $B = C = 1$), we draw attention to the huge improvements in model fit. However, we do acknowledge that this improvement is partly due to the fact that the panel nature of the data is being accounted for, making it difficult to truly corroborate this. In both models, we see that the signs and significance of the marginal utility parameters in the modal class are relatively similar to those retrieved under the MNL model. In the two preference class model (i.e., where $A = 2$) we highlight the differences in sign of the local and heather marginal utilities in the second latent class. For the three preference class model (i.e., where $A = 3$) we observe that the region of origin is of relatively lesser importance to the second latent class compared to the third latent class. In Models 4 and 7 the latent classes are differentiated only on the basis of the scale (i.e., where $B > 1$ and $A = C = 1$). Importantly, both models reveal a share of respondents (in the region of 25–30 percent) with relative scale parameters indistinguishable from zero, which essentially implies random decision-making. Accounting for this leads to improvements in model fit. The final model in Table 3 is an independent availability logit (i.e., where $C > 1$ and $A = B = 1$) that shows that perhaps over one-quarter of respondents excluded the status-quo alternative from their consideration set. Comparing across all models that address one type of heterogeneity, we can see that, other things being held constant, higher improvements are attained when preference heterogeneity is accommodated, followed by processing heterogeneity.

Model fit summary results for the 90 candidate models are presented in Table 4. This includes models that accommodate up to three latent classes for preferences and scale as well as specifications that allow for not only the deterministic choice set but also the consideration set that includes only the non-status-quo alternatives. Each

Table 3. Estimation results for benchmark heterogeneity models (whole sample)

	Model 1	Model 2	Model 3	Model 4	Model 7	Model 10
A	1	2	3	1	1	1
B	1	1	1	2	3	1
C	1	1	1	1	1	2
LL$\left(\hat{\beta}\right)$	-6,483.35	-5,205.68	-4,857.80	-6,163.53	-6,150.34	-6,110.42
BIC*	13,017.91	10,519.48	9,880.63	12,389.66	12,374.65	12,277.75
Preference class 1						
Price	-0.05** (0.00)	-0.06** (0.00)	-0.08** (0.00)	-0.07** (0.00)	-0.22** (0.06)	-0.06** (0.00)
Origin: Local	1.08** (0.06)	2.24** (0.11)	1.63** (0.14)	1.97** (0.13)	5.74** (1.73)	1.28** (0.08)
Origin: Denmark	1.39** (0.06)	2.53** (0.12)	1.90** (0.13)	2.45** (0.14)	7.30** (2.20)	1.62** (0.08)
Origin: European	0.58** (0.05)	0.75** (0.09)	0.90** (0.10)	0.92** (0.07)	2.69** (0.93)	0.63** (0.06)
Organic	0.49** (0.04)	0.58** (0.07)	0.79** (0.08)	0.54** (0.06)	1.42** (0.34)	0.47** (0.04)
Type: Heather	0.24** (0.05)	0.14* (0.08)	-0.01 (0.11)	0.38** (0.07)	1.05** (0.39)	0.26** (0.05)
Type: Clover	-0.22** (0.05)	-0.05 (0.06)	-0.22** (0.08)	-0.13* (0.06)	-0.55** (0.27)	-0.19** (0.04)
Type: Rape	-0.09** (0.05)	-0.30** (0.06)	-0.34** (0.08)	-0.16** (0.05)	-0.56** (0.19)	-0.12** (0.04)
SQ ASC	-0.52** (0.07)	-1.63** (0.10)	-1.74** (0.13)	-1.38** (0.10)	-4.38** (0.98)	-0.10 (0.08)
Preference class 2						
Price		-0.16** (0.01)	-0.16** (0.02)			
Origin: Local		-0.15 (0.27)	-0.71* (0.43)			
Origin: Denmark		0.78** (0.23)	0.22 (0.39)			
Origin: European		0.12 (0.12)	-0.13 (0.18)			
Organic		0.40** (0.11)	0.26* (0.15)			
Type: Heather		-0.12 (0.12)	-0.04 (0.18)			
Type: Clover		-0.51** (0.12)	-0.58** (0.15)			
Type: Rape		-0.19* (0.12)	-0.26 (0.17)			
SQ ASC		-1.01** (0.15)	-0.97** (0.21)			
Preference class 3						
Price			-0.04** (0.01)			
Origin: Local			3.52** (0.45)			
Origin: Denmark			3.64** (0.52)			
Origin: European			0.46* (0.22)			
Organic			0.42* (0.20)			
Type: Heather			0.20 (0.21)			
Type: Clover			0.37 (0.27)			
Type: Rape			-0.33 (0.23)			
SQ ASC			-2.00** (0.37)			
Scale parameters						
λ_2				0.00 (0.07)	0.31** (0.08)	
λ_3					0.00 (0.00)	
Unconditional preference class probabilities						
$\Pr(\beta = \beta_1)$		0.61** (0.03)	0.52** (0.03)			
$\Pr(\beta = \beta_2)$		0.39** (0.03)	0.30** (0.03)			
$\Pr(\beta = \beta_3)$			0.18** (0.02)			
Unconditional variance class probabilities						
$\Pr(\lambda = 1)$				0.73** (0.02)	0.08** (0.02)	
$\Pr(\lambda = \lambda_2)$				0.27** (0.02)	0.67** (0.03)	
$\Pr(\lambda = \lambda_3)$					0.25** (0.02)	
Unconditional processing class probabilities						
$\Pr(S_c = \text{A}, \text{B}, \text{SQ})$						0.74** (0.02)
$\Pr(S_c = \text{A}, \text{B})$						0.26** (0.02)

Notes: All estimated standard errors (in parentheses) are robust and clustered at the respondent level. * and ** indicate statistical significance at the 5 and 1 percent level respectively using the p-value of a one-sided test. The associated p-values for the estimated scale parameters test: $H_0 : \hat{\lambda}_b = 1$. The estimated parameter and its standard error for λ_3 in Model 7 are both positive but are less than 1×10^{-2}.

Table 4. Model fit summary results

Model	*A*	*B*	*C*		$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$
1	1	1	1	$\text{LL}\left(\hat{\beta}\right)$	-6,483.35	-6,483.35	-6,483.35	-6,483.35	-6,483.35
				BIC*	13,017.91	13,017.91	13,017.91	13,017.91	13,017.91
2	2	1	1	$\text{LL}\left(\hat{\beta}\right)$	-5,205.67	-5,204.91	-5,204.15	-5,204.20	-5,203.37
				BIC*	10,519.46	10,523.63	10,527.80	10,533.59	10,537.63
3	3	1	1	$\text{LL}\left(\hat{\beta}\right)$	-4,857.80	-4,847.75	-4,839.54	-4,836.26	-4,836.43
				BIC*	9,880.63	9,871.91	9,866.87	9,871.69	9,883.42
4	1	2	1	$\text{LL}\left(\hat{\beta}\right)$	-6,163.53	-6,158.89	-6,158.83	-6,155.46	-6,155.63
				BIC*	12,389.66	12,386.07	12,391.64	12,390.60	12,396.62
5	2	2	1	$\text{LL}\left(\hat{\beta}\right)$	-5,166.19	-5,165.42	-5,166.09	-5,163.83	-5,166.10
				BIC*	10,451.89	10,461.72	10,474.45	10,481.30	10,497.23
6	3	2	1	$\text{LL}\left(\hat{\beta}\right)$	-4,820.68	-4,802.95	-4,789.88	-4,788.10	-4,788.70
				BIC*	9,817.77	9,799.38	9,790.31	9,803.83	9,822.11
7	1	3	1	$\text{LL}\left(\hat{\beta}\right)$	-6,150.34	-6,145.24	-6,144.51	-6,140.73	-6,139.87
				BIC*	12,374.65	12,375.85	12,385.77	12,389.58	12,399.24
8	2	3	1	$\text{LL}\left(\hat{\beta}\right)$	-5,166.19	-5,165.19	-5,166.04	-5,163.77	-5,162.28
				BIC*	10,463.27	10,478.33	10,497.11	10,509.64	10,523.73
9	3	3	1	$\text{LL}\left(\hat{\beta}\right)$	-4,777.29	-4,775.02	-4,770.79	-4,766.18	-4,763.41
				BIC*	9,742.37	9,760.60	9,774.90	9,788.43	9,805.66
10	1	1	2	$\text{LL}\left(\hat{\beta}\right)$	-6,110.42	-6,109.36	-6,107.31	-6,107.64	-6,106.66
				BIC*	12,277.75	12,281.32	12,282.92	12,289.26	12,292.99
11	2	1	2	$\text{LL}\left(\hat{\beta}\right)$	-5,178.87	-5,177.70	-5,176.18	-5,174.84	-5,174.41
				BIC*	10,471.55	10,480.59	10,488.93	10,497.63	10,508.15
12	3	1	2	$\text{LL}\left(\hat{\beta}\right)$	-4,825.23	-4,816.37	-4,806.42	-4,799.76	-4,798.76
				BIC*	9,821.18	9,820.52	9,817.69	9,821.46	9,836.52
13	1	2	2	$\text{LL}\left(\hat{\beta}\right)$	-5,950.07	-5,930.23	-5,924.35	-5,917.21	-5,916.87
				BIC*	11,968.42	11,940.14	11,939.74	11,936.84	11,947.55
14	2	2	2	$\text{LL}\left(\hat{\beta}\right)$	-5,138.34	-5,138.29	-5,138.18	-5,138.12	-5,138.16
				BIC*	10,401.86	10,418.84	10,435.69	10,452.65	10,469.80
15	3	2	2	$\text{LL}\left(\hat{\beta}\right)$	-4,786.31	-4,767.98	-4,758.97	-4,756.24	-4,770.46
				BIC*	9,754.72	9,740.83	9,745.57	9,762.86	9,814.06
16	1	3	2	$\text{LL}\left(\hat{\beta}\right)$	-5,949.33	-5,927.11	-5,920.19	-5,916.18	-5,909.53
				BIC*	11,978.33	11,950.96	11,954.20	11,963.25	11,967.01
17	2	3	2	$\text{LL}\left(\hat{\beta}\right)$	-5,137.23	-5,136.12	-5,133.57	-5,131.29	-5,133.85
				BIC*	10,411.04	10,431.59	10,449.25	10,467.44	10,495.32
18	3	3	2	$\text{LL}\left(\hat{\beta}\right)$	-4,748.73	-4,720.09	-4,715.97	-4,711.11	-4,720.46
				BIC*	9,690.94	9,662.12	9,682.32	9,701.06	9,748.21

of the 18 combinations of preference, scale and processing class structures are evaluated for cases where separate latent class membership probabilities are attained for up to five equal-sized data subsets sorted by response time, leading to $M = 18 \times 5 = 90$ candidate models.

The column headed by $q = 1$ is the entire sample and, thus, does not retrieve separate latent class probabilities on the basis of response time. We can see from this column that compared to the MNL model (the first model listed), all other models, which account for at least one type of heterogeneity, offer an improvement in model fit. As expected, we also find that higher log-likelihoods are obtained under more flexible specifications, which accommodate more heterogeneity. This is also supported by the BIC* values, which account for the increase in estimated parameters.

When different class membership probabilities are permitted for respondents who are below and above the median response time we witness improvements in model fit. This is an important finding since it supports our conjecture that preference structures, scale and/or processing strategies may be different among ‘fast’ and ‘slow’ respondents, thus motivating this line of inquiry. For instance, we find an increase of over 10 log-likelihood units for model 3, implying, all else held constant, that the quickest and slowest respondents most likely have different preferences. A difference in log-likelihood is also found for model 7, which, again, suggests that different proportions of respondents below and above the median response time are associated with the latent classes used to explain differences in scale. Interestingly, there is only a modest increase in log-likelihood for model 10, indicating that the processing strategies adopted by the relatively fast respondents are unlikely to differ from those used by the respondents who took relatively more time.

While further improvements in model fit are achieved when the respondents are grouped into tertiles, quartiles and quintiles, we remark that the magnitude of the improvements diminish. Moreover, on the basis of the BIC* values, which penalizes for the proliferation of parameters needed to obtain separate class probabilities for each k^{th} q-quantile, there is generally little support to move beyond segmenting into tertiles. Notwithstanding this, all models considered, the model associated with the highest log-likelihood value is model 18 on the basis that respondents are separated into quartiles, whereas the model with the lowest BIC* value is the equivalent model where respondents are split by the median response time.

Table 5 presents a more detailed overview of the estimation results. To begin with, the mean of the model weights for the k q-quantiles (i.e., $\sum_{n_k}^{N_k} \mathbb{E}\left(\bar{\pi}_{z_{n_k}}\right)/N_k$) are given. These can be interpreted as the average probability that each model is the best approximating model (given the data, the set of models, and the unknowable true model) for respondents within each quantile. The larger the value, the more confidence we have that the model is the best approximating model. From the table, the confidence set of models (i.e., the subset that represent the majority of evidence) comprises of models 9, 12, 15 and 18. Interestingly, these models all allow for three latent segments of preferences, which implies that we can be fairly confident that a model accommodating for this is the best approximating model. Irrespective of response time quantile, model 12, which concurrently permits three classes of preferences and two consideration sets, is generally found to have the highest average weight of evidence. This bolsters the status of model 12 in terms of its candidacy as the best approximating model. Therefore, there is a relatively high confidence that if a different sample was drawn from the population, model 12 will, again, be judged, on average, as the best fitting. We also draw attention to the fact that the average weight of evidence for model 12 appears to be higher among faster respondents.

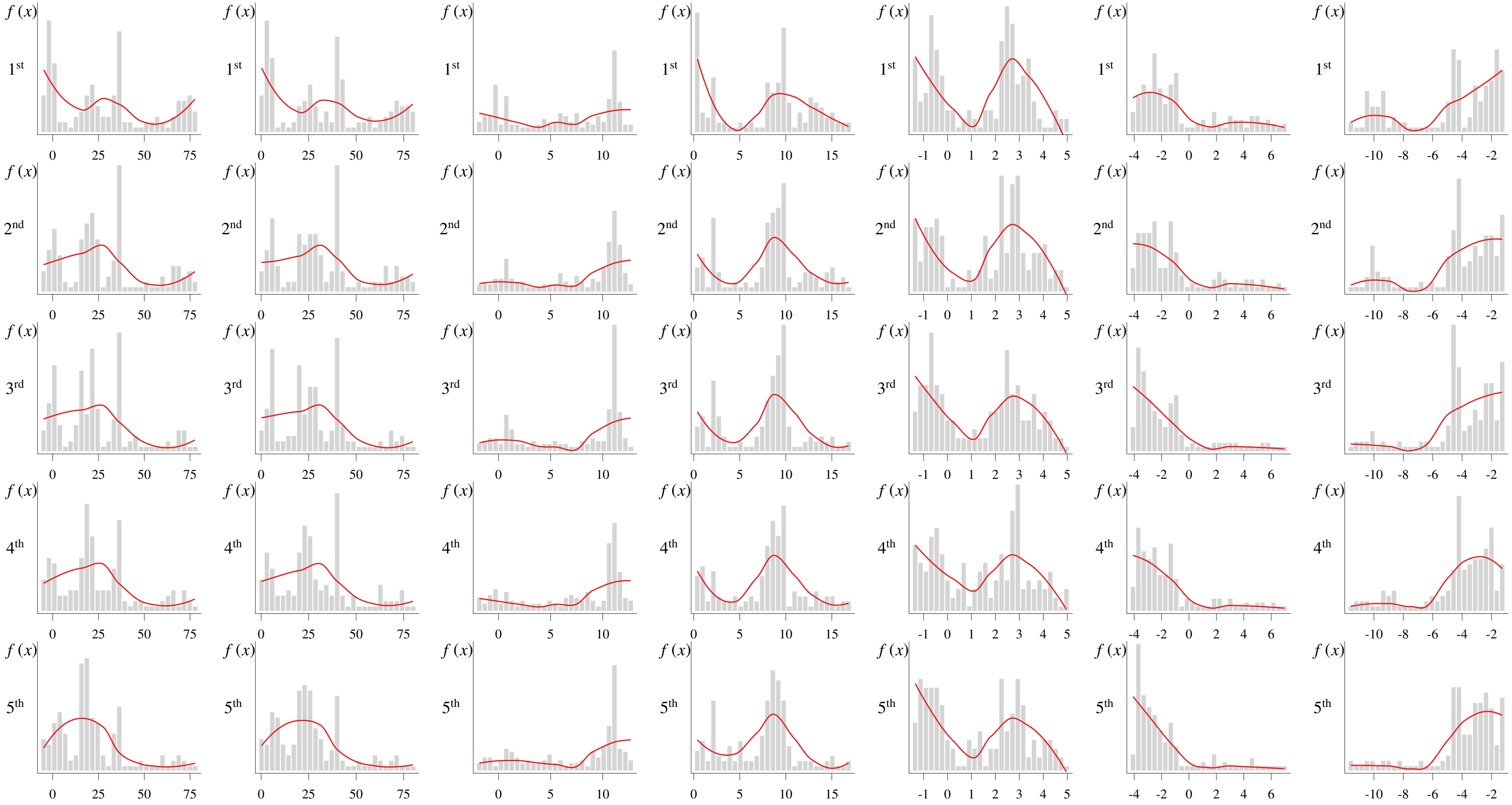
Table 5 also compares the mean of the expected marginal WTP estimate for each respondent (Eq. 12a) against response time quantile. This reveals that this sample of respondents are, on average, willing to pay extra for honey that originates in Denmark (both locally and nationally) or elsewhere in Europe as well as for honey organically produced. Of the different types of honey, on average they are willing to pay most for honey produced from heather and mixed vegetation. Importantly, in the context of this paper, an investigation of the marginal WTP estimates reveals that they are notably different by quantile. On average, respondents who took a longer time to complete the choice experiment had lower marginal WTP estimates for local and Danish honey (relative to non-European honey) and also for honey originating from heather and clover vegetation (relative to mixed vegetation). In fact, if we compare the fastest and slowest quintiles, the average marginal WTP estimates for a jar of Danish honey drops from over DKK 30 to below DKK 20. In contrast, the marginal WTP for honey produced elsewhere in Europe (with respect to non-European honey) for the slowest quintile is over 30 percent higher than the respective value for the fastest quintile. Marginal WTP for organically produced honey does not appear to relate to response time.

To further assess marginal WTP, in Fig. 1 we plot their distributions by response time quintile. Note that the distributions are not conditional on a single model but on the whole model set. For each respondent, the mean of the conditional distribution is retrieved from every model and then weighted using its weight of evidence for that respondent. These more clearly illustrate the differences in the average marginal WTP estimates discussed above. However, in addition, they highlight that there is a higher degree of heterogeneity among quicker respondents compared to slower respondents. Whereas there is considerable variability in the values estimated for respondents in the first and second quintiles, the distributions for the fourth and fifth quintiles generally exhibit a more pronounced peak and, therefore, less variability.

This article employed a probabilistic approach for identifying heterogeneity in scales. To facilitate interpretation, in Table 5 we report the mean of the expected class memberships (Eq. 12b) for classes where $\lambda < 1$, $\lambda = 1$ and $\lambda > 1$ against response time quantile. While we recognize the difficultly involved with comparing (and, therefore, weighting) the scale parameters across the 18 candidate models, this does not prevent us from drawing some conclusions regarding how the average scale differs depending on the length of time respondents take to answer the choice experiment. Opposite to what might be expected, we find that the average share of respondents belonging in classes where the scale parameter is estimated to be above 1 decreases with response time. Recall that the scale parameter is inversely proportional to the variance of the error term, meaning that, on average, the level of noise actually increases with response time. While this alleviates our concern that the relatively quick responses provide nothing but noise, it does seem somewhat counterintuitive—respondents who spent more time deliberating and processing the information did not, on average, eventually make more consistent choices. One explanation for this could be that the long response times may be an artifact of respondents not focusing solely on the task at hand, perhaps because they faced distractions or were multitasking. If so, it is plausible that their choices are associated with a higher degree of variability.

Table 5. Main model results compared against response latency quantiles

	$q = 1$	$q = 2$			$q = 3$				$q = 4$					$q = 5$					
	Total	1st	2nd	Total	1st	2nd	3rd	Total	1st	2nd	3rd	4th	Total	1st	2nd	3rd	4th	5th	Total
Mean of model weights																			
Model 1	0.03	0.04	0.03	0.04	0.03	0.03	0.04	0.03	0.02	0.05	0.02	0.04	0.03	0.03	0.03	0.04	0.04	0.03	0.03
Model 2	0.05	0.05	0.05	0.05	0.06	0.06	0.04	0.05	0.04	0.07	0.05	0.04	0.05	0.07	0.04	0.05	0.04	0.05	0.05
Model 3	0.06	0.07	0.05	0.06	0.07	0.06	0.07	0.06	0.07	0.07	0.07	0.05	0.06	0.07	0.05	0.04	0.07	0.06	0.06
Model 4	0.03	0.03	0.05	0.04	0.04	0.03	0.05	0.04	0.02	0.04	0.03	0.05	0.04	0.02	0.05	0.03	0.05	0.04	0.04
Model 5	0.05	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.07	0.04	0.06	0.04	0.05	0.03	0.05	0.07	0.04	0.05	0.05
Model 6	0.06	0.05	0.06	0.06	0.06	0.06	0.07	0.06	0.07	0.04	0.06	0.07	0.06	0.07	0.06	0.06	0.07	0.07	0.07
Model 7	0.05	0.05	0.04	0.04	0.03	0.04	0.04	0.04	0.04	0.05	0.03	0.05	0.04	0.01	0.03	0.04	0.04	0.04	0.03
Model 8	0.05	0.06	0.06	0.06	0.07	0.04	0.05	0.06	0.05	0.05	0.06	0.06	0.05	0.07	0.07	0.05	0.06	0.06	0.06
Model 9	0.07	0.08	0.06	0.07	0.10	0.08	0.04	0.07	0.10	0.05	0.07	0.08	0.08	0.10	0.08	0.06	0.06	0.04	0.07
Model 10	0.05	0.04	0.05	0.04	0.02	0.05	0.06	0.04	0.03	0.05	0.04	0.06	0.05	0.03	0.05	0.05	0.04	0.06	0.05
Model 11	0.06	0.07	0.06	0.06	0.06	0.09	0.05	0.07	0.06	0.08	0.06	0.05	0.06	0.06	0.07	0.10	0.05	0.05	0.06
Model 12	0.08	0.07	0.08	0.08	0.10	0.06	0.06	0.07	0.09	0.07	0.07	0.07	0.08	0.09	0.08	0.06	0.09	0.07	0.08
Model 13	0.05	0.04	0.06	0.05	0.04	0.05	0.06	0.05	0.02	0.05	0.07	0.05	0.05	0.04	0.06	0.04	0.07	0.07	0.05
Model 14	0.06	0.06	0.06	0.06	0.05	0.06	0.05	0.06	0.05	0.06	0.07	0.06	0.06	0.04	0.05	0.06	0.08	0.05	0.06
Model 15	0.07	0.07	0.08	0.08	0.07	0.05	0.08	0.07	0.08	0.05	0.07	0.06	0.07	0.08	0.05	0.07	0.09	0.08	0.07
Model 16	0.05	0.04	0.06	0.05	0.03	0.06	0.07	0.05	0.03	0.05	0.05	0.05	0.04	0.03	0.05	0.05	0.04	0.08	0.05
Model 17	0.05	0.07	0.05	0.06	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.05	0.06	0.05	0.10	0.04	0.04	0.06	0.06
Model 18	0.08	0.08	0.05	0.07	0.09	0.06	0.06	0.07	0.10	0.06	0.05	0.05	0.07	0.12	0.06	0.08	0.05	0.05	0.07
Mean of weighted conditional means of marginal WTP estimates (DKK per jar)																			
Origin: Local	24.05	25.10	20.67	22.90	27.24	21.43	21.00	23.25	26.82	24.05	22.23	20.32	23.37	27.20	26.61	21.18	22.92	18.83	23.37
Origin: Denmark	28.52	29.25	25.21	27.24	31.35	26.03	25.63	27.70	30.87	28.57	26.72	24.92	27.78	31.01	30.74	25.79	27.35	23.54	27.71
Origin: European	7.88	7.51	8.12	7.82	6.67	7.79	8.03	7.49	6.34	8.02	7.93	7.86	7.54	6.04	7.94	7.96	8.03	7.90	7.57
Organic	7.40	7.82	7.51	7.67	7.71	7.46	7.51	7.56	7.32	7.83	7.59	7.27	7.50	7.73	8.32	7.49	7.80	7.24	7.72
Type: Heather	1.50	1.42	1.15	1.29	1.39	1.34	1.33	1.35	1.39	1.57	1.32	1.32	1.40	1.55	1.65	1.28	1.45	1.07	1.40
Type: Clover	-1.61	-1.77	-2.39	-2.08	-1.33	-2.20	-2.34	-1.95	-1.18	-1.96	-2.10	-2.29	-1.88	-1.07	-1.58	-2.24	-2.01	-2.48	-1.87
Type: Rape	-3.81	-4.09	-3.63	-3.86	-4.52	-3.69	-3.60	-3.94	-4.42	-3.86	-3.74	-3.44	-3.86	-4.71	-4.31	-3.68	-3.93	-3.50	-4.03
Mean of weighted conditional means of λ class memberships																			
$\Pr(\lambda < 1)$	0.13	0.18	0.17	0.17	0.15	0.20	0.17	0.17	0.18	0.16	0.20	0.24	0.19	0.19	0.24	0.19	0.23	0.28	0.23
$\Pr(\lambda = 1)$	0.75	0.71	0.73	0.72	0.74	0.70	0.74	0.73	0.74	0.77	0.70	0.69	0.73	0.71	0.71	0.76	0.69	0.68	0.71
$\Pr(\lambda > 1)$	0.12	0.12	0.10	0.11	0.11	0.10	0.09	0.10	0.08	0.08	0.10	0.07	0.08	0.10	0.04	0.05	0.08	0.04	0.06
Mean of weighted conditional means of processing strategy class memberships																			
$\Pr(S_c = \text{A, B, SQ})$	0.93	0.94	0.91	0.93	0.93	0.92	0.93	0.93	0.92	0.95	0.92	0.92	0.93	0.94	0.92	0.94	0.94	0.92	0.93
$\Pr(S_c = \text{A, B})$	0.07	0.06	0.09	0.07	0.07	0.08	0.07	0.07	0.08	0.05	0.08	0.08	0.07	0.06	0.08	0.06	0.06	0.08	0.07



(a) Origin: Local (b) Origin: Denmark (c) Origin: European (d) Organic (e) Type: Heather (f) Type: Clover (g) Type: Rape

Figure 1. Distributions of the weighted conditional means (over all models) of estimated marginal WTP (DKK per jar) by response time quintile

The final set of results listed in Table 5 refers to the mean weighted average class memberships for the two processing strategies considered in our models. Reassuringly, we find strong evidence that the majority of respondents comply with the standard assumption of considering all alternatives within the deterministic choice set. Of especial relevance in this article is the noticeable link between the processing strategies adopted and response time. For instance, while, on average, only 6 percent of respondents in the first quintile excluded the status-quo alternative from their consideration set, the respective figure for the fifth quintile is 8 percent. Again, the direction of change is not as anticipated—even though an elimination-by-aspects decision-making mechanism is adopted (which simplifies the choice) response time is not reduced on average. Nevertheless, it does explain some of the increase in choice variability and further supports the concern that respondents who required a longer time faced distractions and did not treat the exercise as serious as others.

5. Conclusions

While the relationship between the length of time that respondents take to answer stated preference questions and the reliability of their choices might seem obvious, in a previous paper (Campbell, Mørkbak, and Olsen, 2016) we show that the relationship between response time and data quality is ambiguous. In this follow up paper based on a new dataset, we contribute further evidence to underpin the need to appreciate the differences inherent in 'fast' and 'slow' choices. Here we propose a novel approach to assess the effects of response time on the estimates of utility coefficients, scale and processing strategies. While it is relatively straightforward to establish each of these in turn, this article sets out to address them simultaneously, which, as discussed in Hess and Train (2017), is considerably more challenging. Our motivation stems from a concern that accommodating only one type of heterogeneity may provide an incomplete picture and may even be distorted by the other (unmodeled) types. Our analysis is based on the latent class modeling framework and is aimed at separately identifying the heterogeneity in preferences, scale and the processing strategies across respondents. We estimate separate class membership probabilities for different subsets of respondents, classified on the basis of how long they required to complete the survey. Our analysis considers 90 candidate model specifications and uses multimodel inference to form weights of evidence. Specifically, we use these weights to rank models in this candidate set, and to derive weighted average results so that they are not conditional on a single model but on the entire model set. We test our approach through an empirical dataset collected via an online survey to establish the value that Danes are willing to pay for various attributes associated with honey.

Our article raises a number of methodological issues. Importantly, in accordance with earlier studies (e.g., Haaijer, Kamakura, and Wedel, 2000; Rose and Black, 2006; Otter, Allenby, and van Zandt, 2008; Vista, Rosenberger, and Collins, 2009; Haaijer, Kamakura, and Wedel, 2000) and our preceding study (Campbell, Mørkbak, and Olsen, 2016), we show that response time is an important consideration when modeling discrete choice data. We find that while marginal WTP estimates for some attributes are higher among respondents with long response times, for other attributes it is the reverse. Opposite to what might be expected, we find that measurement error is highest among those who took the longest time to complete the choice experiment. Importantly, we also shed light on the fact that the processing strategies adopted also differ by response time quantile. Furthermore, our model inference analysis gives a strong signal that models that address different types of heterogeneity simultaneously are leading candidates for the best approximating model (given the data, the set of models, and the unknowable true model).

From data quality and, as we have seen, from welfare analysis standpoints, it may be tempting to 'clean' datasets from respondents below or above a certain response time as suggested by Bonsall and Lythgoe (2009). However, unlike our earlier paper (Campbell, Mørkbak, and Olsen, 2016), we stress that the analytical framework proposed in the present paper does not, and was not intended to, identify what these thresholds might be. Instead, our analysis is intended to provide analysts with a modeling framework to assess the link between survey results and response time. We have striven for a model that recognizes the highly equivocal link between response time and data quality without the need to 'clean' out any of the data.

With respect to our findings, these are most likely not isolated to online surveys alone, meaning that there is scope for further research in uncovering the role of response time in other modes—especially given the evidence that response time also is likely to vary across modes (e.g., Börjesson and Algers, 2011). This would provide us with additional criteria that we can evaluate when designing stated choice surveys. Even though our analysis is based on a food application, the impact of response time and the modeling framework introduced should be of interest to a broader audience—especially in settings where the choice is even more abstract or difficult than what is the case in the present situation. Thus, we encourage ongoing research and applications within our proposed modeling framework. Another string of future research lies in how the paradata is measured. In specific relation to response time, as also suggested by Lindhjem and Navrud (2011b), the development of more stringent measures could be beneficial—for instance, accounting for webpage load times as well as respondent multitasking would seem appropriate. With regard to the modeling framework presented here, there is also scope for further research in terms of developing even more flexible models. Even though we have done far more than what is usually done with respect to heterogeneity (where only one type is taken into account at a time), one could easily imagine an extension of even more latent classes within all three different kinds of heterogeneity. In the current analysis, we have limited the number of classes, due to computational and sample size issues, but we acknowledge that our classes may not provide sufficient flexibility to fully accommodate the heterogeneity in preferences, error variance and processing strategies. Moreover, the econometric model could be extended to accommodate other decision heuristics such as attribute non-attendance or other simplifying heuristic decision rules not captured in our specification. Thus, as with other research focusing on heterogeneity, the reader has to bear in mind that the results regarding heterogeneity are still at risk of being confounded with other unmodeled types of heterogeneity. However, there may be limits on how much heterogeneity can be accounted for due to potential increasing problems of identification and convergence. Indeed, the ability to correctly identify more than one type of heterogeneity at the same time is complicated by the fact that scale heterogeneity is not identified separately from other sources of heterogeneity (e.g., see Hess and Train, 2017). The issue of utilizing choice task time rather than choice sequence time could be a further interesting issue to examine, in terms of both measurement-wise as well as econometrically (although one would have to deal with the risk of this being confounded with learning or fatigue effects, as addressed in Campbell et al. (2015), and consider the potential implications it may pose for the maintenance of the panel structure). Finally, while model comparison and averaging are routinely performed when analyzing stated preference data within the Bayesian framework (e.g., Leon-Gonzalez and Scarpa, 2008; Balcombe, Chalak and Fraser, 2009), with the notable exception of Layton and Lee (2006), the practice appears to be seldom considered within the classical framework. Analysts engaged in estimating stated choice experiments using classical models should benefit from using such an approach when comparing and ranking multiple competing models. Indeed, this should become a recommended course of action in practice.

In conclusion, irrespective of response time, our article clearly demonstrates the need for researchers to consider the variations among respondents. Despite the significant gains that have been made in this regard, there is a need for models that are better equipped to accommodate this variation. We should recognize that heterogeneity is not restricted to preferences, nor is it limited to either variances or processing strategies. The factors influencing choice outcomes are complex, so it should not come as a surprise that accommodating as many types of heterogeneity as possible will lead to better choice models.

References

Balcombe, K., A. Chalak, and I. Fraser. 2009. "Model selection for the mixed logit with Bayesian estimation." *Journal of Environmental Economics and Management* 57:226–237.

Ben-Akiva, M., and B. Boccara. 1995. "Discrete choice models with latent choice sets." *International Journal of Research in Marketing* 12:9–24.

Bonsall, P., and B. Lythgoe. 2009. "Factors affecting the amount of effort expended in responding to questions in behavioural choice experiments." *Journal of Choice Modelling* 2:216–236.

Börger, T. 2015. "Are fast responses more random? Testing the effect of response time on scale in an online choice experiment." *Environmental and Resource Economics in press*.

Börjesson, M., and S. Algers. 2011. "Properties of internet and telephone data collection methods in a stated choice value of time study context." *Journal of Choice Modelling* 4:1–19.

Börjesson, M., and M. Fosgerau. 2015. "Response time patterns in a stated choice experiment." *Journal of Choice Modelling* 14:48–58.

Brown, T.C., D. Kingsley, G.L. Peterson, N.E. Flores, A. Clarke, and A. Birjulin. 2008. "Reliability of individual valuations of public and private goods: choice consistency, response time, and preference refinement." *Journal of Public Economics* 92:1595–1606.

Buckland, S.T., K.P. Burnham, and N.H. Augustin. 1997. "Model selection: an integral part of inference." *Biometrics* 53:603–618.

Campbell, D., M. Boeri, E. Doherty, and W.G. Hutchinson. 2015. "Learning, fatigue and preference formation in discrete choice experiments." *Journal of Economic Behavior and Organization* 119:345–363.

Campbell, D., and S. Erdem. 2015. “Position bias in best-worst scaling surveys: a case study on trust in institutions.” *American Journal of Agricultural Economics* 97:526–545.

Campbell, D., D.A. Hensher, and R. Scarpa. 2011. “Non-attendance to attributes in environmental choice analysis: a latent class specification.” *Journal of Environmental Planning and Management* 54:1061–1076.

Campbell, D., D.A. Hensher, and R. Scarpa. 2014. “Bounding WTP distributions to reflect the ‘actual’ consideration set.” *Journal of Choice Modelling* 11:4–15.

Campbell, D., M.R. Mørkbak, and S.B. Olsen. 2016. “Response time in online stated choice experiments: the non-triviality of identifying fast and slow respondents.” *Journal of Environmental Economics and Policy in press*.

Chang, J.B., J.L. Lusk, and F.B. Norwood. 2009. “How closely do hypothetical surveys and laboratory experiments predict field behavior?” *American Journal of Agricultural Economics* 91:518–534.

Cook, J., M. Jeuland, B. Maskery, and D. Whittington. 2012. “Giving stated preference respondents “time to think”: results from four countries.” *Environmental and Resource Economics* 51:473–496.

Fleming, C.M., and M. Bowden. 2009. “Web-based surveys as an alternative to traditional mail methods.” *Journal of Environmental Management* 90:284–292.

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

Haaijer, R., W. Kamakura, and M. Wedel. 2000. “Response latencies in the analysis of conjoint choice experiments.” *Journal of Marketing Research* 37:376–382.

Henningsen, A. and O. Toome. 2011. “Maxlik: A Package for Maximum Likelihood Estimation in R.” *Computational Statistics* 26:443–458.

Hensher, D.A. 2010. “Attribute processing, heuristics and preference construction in choice analysis.” In S. Hess and A. Daly, eds. *Choice Modelling The State-of-the-art and the State-of-practice - Proceedings from the Inaugural International Choice Modelling Conference*. Emerald Press, UK, pp. 35–70.

Hess, S., and J.M. Rose. 2012. “Can scale and coefficient heterogeneity be separated in random coefficients models?” *Transportation* 39:1225–1239.

Hess, S., and A. Stathopoulos. 2013. “Linking response quality to survey engagement: a combined random scale and latent variable approach.” *Journal of Choice Modelling* 7:1–12.

Hess, S., and K. Train. 2017. “Correlation and scale in mixed logit models.” *Journal of Choice Modelling* 23:1–8.

Holmes, T., K. Alger, C. Zinkhan, and E. Mercer. 1998. “The effect of response time on conjoint analysis estimates of rainforest protection values.” *Journal of Forest Economics* 4:7–28.

Kaplan, S., Y. Shiftan, and S. Bekhor. 2012. “Development and estimation of a semi-compensatory model with a flexible error structure.” *Transportation Research Part B: Methodological* 46:291–304.

Layton, D.F., and S.T. Lee. 2006. “Embracing model uncertainty: strategies for response pooling and model averaging.” *Environmental and Resource Economics* 34:51–85.

Leon-Gonzalez, R. and R. Scarpa. 2008. “Improving multi-site benefit functions via Bayesian model averaging: a new approach to benefit transfer.” *Journal of Environmental Economics and Management* 56:50–68.

Lindhjem, H., and S. Navrud. 2011a. “Are Internet surveys an alternative to face-to-face interviews in contingent valuation?” *Ecological Economics* 70:1628–1637.

Lindhjem, H., and S. Navrud. 2011b. “Using internet in stated preference surveys: a review and comparison of survey modes.” *International Review of Environmental and Resource Economics* 5:309–351.

Luce, R.D. 1986. *Response Times: Their Role in Inferring Elementary Mental Organization*. New York: Oxford University Press.

Magidson, J., and J.K. Vermunt. 2008. “Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference.” In *Sawtooth Software Conference*.

Malhotra, N. 2008. “Completion time and response order effects in web surveys.” *Public Opinion Quarterly* 72:914.

Manski, C.F. 1977. “The structure of random utility models.” *Theory and Decision* 8:229–254.

Olsen, S.B. 2009. “Choosing between internet and mail survey modes for choice experiment surveys considering non-market goods.” *Environmental and Resource Economics* 44:591–610.

Otter, T., G.M. Allenby, and T. van Zandt. 2008. “An integrated model of discrete choice and response time.” *Journal of Marketing Research* 45:593–607.

R Core Team. 2014. *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Richardson, A. 1982. “Search models and choice set generation.” *Transportation Research Part A: General* 16:403–419.

Rose, J., and I. Black. 2006. “Means matter, but variance matter too: decomposing response latency influences on variance heterogeneity in stated preference experiments.” *Marketing Letters* 17:295–310.

Rubinstein, A. 2007. “Instinctive and Cognitive Reasoning: A Study of Response Times*.” *The Economic Journal* 117:1243–1259.

Scarpa, R., T.J. Gilbride, D. Campbell, and D.A. Hensher. 2009. “Modelling attribute non-attendance in choice experiments for rural landscape valuation.” *European Review of Agricultural Economics* 36:151–174.

Schwappach, D.L.B., and T.J. Strasmann. 2006. ““Quick and dirty numbers”? The reliability of a stated-preference technique for the measurement of preferences for resource allocation.” *Journal of Health Economics* 25:432–448.

Swait, J. 2001. “Choice set generation within the generalized extreme value family of discrete choice models.” *Transportation Research Part B: Methodological* 35:643–666.

Swait, J., and M. Ben-Akiva. 1987. “Incorporating random constraints in discrete models of choice set generation.” *Transportation Research Part B: Methodological* 21:91–102.

Symonds, M.R.E., and A. Moussalli. 2011. “A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike’s information criterion.” *Behavioral Ecology and Sociobiology* 65:13–21.

Thiene, M., and R. Scarpa, J.J. Louviere. 2015. “Addressing preference heterogeneity, multiple scales and attribute attendance with a correlated finite mixing model of tap water choice.” *Environmental and Resource Economics* 62:637–656.

Thiene, M., and J. Swait, and R. Scarpa. 2016. “Choice set formation for outdoor destinations: the role of motivations and preference discrimination in site selection for the management of public expenditures on protected areas.” *Journal of Environmental Economics and Management in press*.

Train, K.E. 2009. *Discrete choice methods with simulation*. Cambridge University Press.

Vista, A.B., R.S. Rosenberger, and A.R. Collins. 2009. “If you provide it, will they read it? Response time effects in a choice experiment.” *Canadian Journal of Agricultural Economics* 57:365–377.

Yang, C.C., and C.C. Yang. 2007. “Separating latent classes by information criteria.” *Journal of Classification* 24:1432–1343.