Accounting for Human Errors in Choice Based Conjoint Survey Data

Frédérique Schellekens, 447534

May 31, 2023

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Business Analytics and Quantative Marketing
Supervisor: Prof. dr. Dennis Fok
Second assessor: Dr. Kathrin Gruber
Supervisor from the company: Joris van Gool

Abstract

Choice Based Conjoint data is used in many marketing settings to derive and predict consumer preferences and choices. Obtaining this choice data is often done through (online) surveys. This use of surveys, where respondents state their own choices, gives rise to the risk of invalid responses. These invalid responses can have many causes, one of these is human errors. I define human errors as the error which occurs when a respondent does not pick the answer they intended to pick, had they correctly read and understood the question and paid sufficient attention when answering the question. These observations then affect the estimation results. To make the Choice Based Conjoint results more reliable, I attempt to account for these human errors. The currently used model uses a Hierarchical Bayes (HB) estimation method. I propose an extension on this model. This extension leads to an additional step in the estimation method which estimates the probability of an error occurring in the data. I implement the extended model and the general HB on 10 simulated data sets, containing various amounts of human errors, to compare the performance. In terms of the results, the extended model is able to capture some randomness or errors. The MSE shows the extended model is able to estimate the consumer preferences closer to the true values in many settings. However in simulations with small randomness or error, the likelihood of the data is worse for the extended model than the general HB model. Most importantly, the extended method shows convergence problems in the estimates. The cause of these convergence problems is currently not known. Therefore the proposed extension is not sufficient to reliably capture human errors. Further research should be implemented to evaluate the method and increase the performance.

Disclaimer: The views stated in this paper are those of the authors and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

When a consumer chooses between several products in a supermarket, or between different computers in an electronics store, the consumer identifies aspects or features of the products and uses these to compare the products. Based on this comparison, the consumer will make a decision on which, if any, product to purchase. Although the consumer makes this decision partly subconsciously, the manufacturers and retailers may be very interested in which features were considered and what weight they got in the final weigh-off before the purchasing decision. These underlying consumer preferences for certain features of products are not directly observable from the consumer decision, they are latent variables. A big step in the research related to these preferences was done by Green & Rao (1971), who introduced the concept of Conjoint Analysis. By presenting consumers with for instance certain products or other choice situations and evaluating the choices made, data is obtained about their preferences. In short, Conjoint Analysis uses hypothetical choices to infer the weigh-offs which the consumer made (Green et al., 2001).

Various forms of Conjoint Analysis (CA) exist. The method I consider is Choice Based Conjoint Analysis (CBC) (Louviere & Woodworth, 1983), where respondents choose their preferred product by choosing between several presented products descriptions. As Louviere (1988) states, a particularly powerful aspect of Conjoint Analysis is the fact that, by choosing a good design and method, the choice task can be created to behave very much alike a real life choice situation. Which leads to more reliable results as opposed to just asking a consumer to state how important they find specific product features when choosing a product. However, Conjoint Analysis results still do not always translate directly to real life decisions (Natter & Feurstein, 2002). This research has the objective of increasing the reliability of Conjoint Analysis results.

The validity of the results depends on many factors, one of these factors concerns how the questions are asked. Conjoint Analysis is oftentimes implemented through a survey. With the rise of online surveys, new research opportunities have emerged, however online surveys also bring new downsides (Pokropek et al., 2022) because the respondent is not directly supervised by an interviewer. For instance, they bring an increased risk of responses not being based on the true preferences of the respondent. The respondent could randomly select to rush through the survey or intentionally answer differently. All of these responses are invalid responses. Invalid responses will cloud the actual preferences of consumers, leading to less reliable results. Finding ways to manage these invalid responses can lead to more reliable results. A lot of research has already been done in the direction of identifying respondents who showcase careless or insufficient effort responding, which from here on will be referred to as C/IER (Johnson, 2005; Pokropek et al., 2022; Huang et al., 2012). These methods attempt to detect when a respondent should not be included in the results as the obtained choices are not valid.

However, removing a respondent from the data leads to a loss of information. When many respondents have to be removed, this may lead to a loss of reliability of the entire study. In some cases, not all choices from a respondent may be invalid, but only a subset. This leads to a less researched direction; *the question-specific validity of a respondents answers.* In short; looking for invalid answers within the answers of otherwise valid respondents. Of course many factors can cause the existence of an invalid answer, as based on Johnson (2005), we can expect intentional misrepresentation for specific questions. Furthermore a respondent may not like to answer certain questions, for instance a question which requires a lot of typing, and answer only this question invalidly. Another cause for an invalid answer can be 'human errors', which will be the main subject for research in this paper. I define a human error, as the error which occurs when a respondent does not pick the answer they intended to pick, had they correctly read and understood the question and paid sufficient attention when answering the question. A human error includes misreading a question, accidentally clicking the wrong answer or perhaps overlooking a certain answer. A respondent might oftentimes not even realise a mistake has occurred. These mistakes reduce the reliability of the obtained results when they are not observed and accounted for.

This leads to the following Research Question: *How can we recognize human errors in the decision results obtained from a Choice Based Conjoint Analysis task?*
With the subquestions:

- *How can we recognize human errors (choices which are not based on the true choice task) in the results obtained from Choice Based Conjoint Analysis, after completion of the survey of a participant?*

- *How should the analysis of the task be adjusted to account for these errors?*

This research is done following a problem description from SKIM. SKIM is a marketing agency, focused on understanding human behavior.

The information I obtain from this research can be used and implemented in various ways. First of all, the overall objective is to recognize human errors in the choices of the respondents and to find a way to estimate the consumer preferences whilst accounting for these human errors. Which 'cleans' the error effect from the true preferences. Making the estimated preferences more reliable and better fitting to real life choices of respondents. These adjusted estimates may allow for a better evaluation and prediction of consumer behavior. Secondly, the knowledge on how and where errors occur, can be used. This knowledge could for instance be used to optimize and improve questionnaires. When a high probability of errors is measured in a certain survey or the CBC part of a survey, this could indicate an error prone survey or task. Causes can be, for instance, too many repetitive tasks which make the respondents lose concentration, or an unclear and confusing task. Recognizing this problem within surveys can therefore give insight into ways to make surveys less error prone.

One of the difficulties of this research is implementing and evaluating methods when the actual occurrence of human errors in the data set is unknown. I will create a simulation for a set of respondents (Vriens et al., 1996). In this set of simulated answers, I will change several of the answers, mirroring actual human errors occurring. I then will implement the method on this simulated data set.

The next step is to use this data to estimate the preferences of the consumers, whilst accounting for the simulated errors. The currently assumed model for consumer preference, also referred to as part-worths, is extended to include the possibility of human errors occurring. Specifically, I extend the probability of a certain product being chosen when presented with several product descriptions by including the probability that an error occurred in this specific choice. I implement the probability of this error by using a parameter which stands for the probability of an error in a task. In the new model, human errors are allowed to occur. Since the model is extended, the estimation method for the part-worths should also be extended. To account for human errors in the estimation method, I focus on detecting human errors after all respondents have filled out the survey. An oftentimes used method for CBC analysis is Hierarchical Bayes (HB) (Hill, 1965). This method combines statistical models with prior knowledge to obtain predictions for the importance of specific features of a product, whilst accounting for the heterogeneity between respondents. The method recursively draws it's estimates, following certain distributions. I extend this method by implementing an additional step in the algorithm which repeatedly samples an additional parameter, used to capture the probability of an error having occurred. The additional step is also based on Bayesian econometrics. The step leads to a new estimation method. The effect of these errors is caught and the obtained part-worths are cleaned from these effects in the new method.

The results show that the extended model has potential in terms of accounting for human errors when estimating the respondent part-worths. However due to convergence problems and other performance issues, the method as currently proposed should not replace the regular Hierarchical Bayes. In more detail, the extended model is able to estimate the part-worths of the respondents closer to the true part-worths, as opposed to the model which does not account for the errors, in several simulations. This is an indication of the possibility to use the proposed method to recognize and account for human errors. Furthermore, in the case of no human errors, the results of both methods are very much alike. However, when adding human errors, the extended model estimates do not converge in all cases, the values fluctuate too much. The exact cause of these convergence problems is not known. These convergence issues cause the estimates to not be reliable. Besides these problems in convergence, the performance in terms of fitting the data and predicting choices as opposed to the regular Hierarchical Bayes, is not better in the extended model. So in general the extended model does not perform better in these measures. Which gives reason to prefer the general Hierarchical Bayes model to the extended model. However, the simulations with large human error probabilities do show a comparable or better performance for the extended model as opposed to the general Hierarchical Bayes model. So although the extended model as used in this paper should not replace the general Hierarchical Bayes model yet. Due to the performance in large error probabilities and the fact that the model improves how close the estimates are to the true part-worths, the extended model as proposed should be seen as a relevant direction of research. The extended model shows potential.

Due to this performance in large error probabilities and the fact that the model improves how close the estimates are to the true part-worths, This paper is structured as follows. Section 2 introduces Conjoint Analysis and discusses the literature on Human Errors and the currently employed methods in Choice Based Conjoint Analysis. Section 3 discusses the methods used. Section 4 describes the used data and relevant simulation settings. In Section 5 the results are discussed, after which I conclude in Section 6. Finally, Section 7 discusses possible improvements and further research.

2 Related Work

In this section, I discuss related work. Firstly, I introduce Conjoint and Choice Based Conjoint (CBC) Analysis with the used estimation method. Then I define a 'Human error' and the possible causes of such an error in Conjoint data. Finally I discuss the methods which surveys use to detect respondents who as a whole are not reliable, as well as the research which has been done on detecting human errors (answer specific errors).

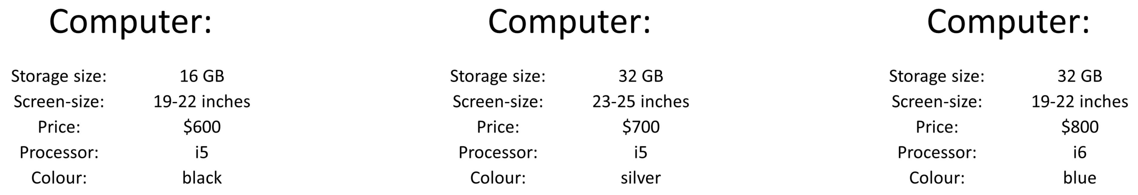
2.1 Conjoint Analysis

Conjoint Analysis, as introduced by Green & Rao (1971), is an overarching name for several methods. All methods have the objective of deriving consumer preferences for certain (product) features. Throughout the years, many have tried to create new methods or improve existing ones (Green & Srinivasan, 1978).

Conjoint analysis allows us to derive the preferences of consumers through a series of questions and tasks. The underlying strategy of Conjoint Analysis is asking a respondent to evaluate a certain feature on its own or as a product. Such a product consists of a combination of certain features. The first Conjoint Analysis as described by Green & Rao (1971) required the respondent to rank all possible features a product could have. Although this leads to a lot of information, the task can become very extensive for large feature sets, and doesn't follow a real-life scenario. Throughout the years, many new developments and improvements have been introduced (Green & Srinivasan, 1978; Green et al., 2001). These developments oftentimes have the objective of creating a task which more closely follows real-life decisions and increasing the amount of information we can obtain from the analysis without drastically increasing the burden on the respondent. The currently used methods include asking respondents to repeatedly, for several tasks, indicate which product they would choose from a set of several alternatives. This set-up can be considered more like reality, where a consumer may stand in front of a shelf in the electronics store and have to choose from several alternative computers with differing features. The change from ranking or valuating exercises to a set of choice tasks was described by Louviere & Woodworth (1983). This choice based method is called Choice Based Conjoint (CBC) and as Louviere & Woodworth (1983) describe, the evaluation of choices can be done using for instance a Multinomial Logit Model (Chrzan & Orme, 2000). This is then called CBC Analysis, which is often used in survey level marketing research. The next section will give a more extensive explanation of CBC Analysis.

2.1.1 Choice Based Conjoint Analysis

In CBC Analysis, there is a group of individuals who are shown a set of questions, also referred to as tasks. Every task asks the respondent to make a choice from several possible combinations of features a product could consist of. Every combination they evaluate in a task is called an alternative. The features they consider within each alternative are decided by a researcher and every feature can have its own number of different options for the feature or attribute, called levels. For example, an individual answers a Conjoint survey on the topic of computers, they perform 5 consecutive tasks. Each task consists of 3 alternatives shown to them. Each alternative is a combination of the attributes: storage-size, processor, price, screen-size and colour. The attribute price then consists of 3 different levels, of these price points, one is shown in each alternative. The attribute colour however consists of 5 different colours, again only one is shown in each alternative. An example task is shown in Figure 1



Computer:

Storage size:	16 GB
Screen-size:	19-22 inches
Price:	$600
Processor:	i5
Colour:	black

(a) Alternative one

Computer:

Storage size:	32 GB
Screen-size:	23-25 inches
Price:	$700
Processor:	i5
Colour:	silver

(b) Alternative two

Computer:

Storage size:	32 GB
Screen-size:	19-22 inches
Price:	$800
Processor:	i6
Colour:	blue

(c) Alternative three

Figure 1. Example task 1

The individual will choose one of these three alternatives, the decision of which they prefer is expected to be based on the importance they assign to each of the attributes and their levels. The choice the individual makes is measured and visible. The underlying importance of each attribute-level is referred to as the part-worth of an attribute-level. The part-worths are not directly observed but Conjoint Analysis attempts to use the observed choices to derive these latent values.

As stated, the data obtained from a CBC survey consists of a certain number of respondent choices, the choice is made between the given alternatives in each choice task. The number of choice tasks will be referred to as s, the number of alternatives in each choice task is m. The number of attribute levels is k. For a certain individual, the number of part-worths to estimate (k) may be less than, equal to, or more than the number of observations for this individual. To obtain good estimates for an individual's part-worths, the observations are likely to not be sufficient (Allenby & Rossi, 1998). We will need many more observations than the k parameters to estimate, which for a set of for instance 5 attributes with each 3 levels, would already lead to more than 15 observations per individual. To not increase the burden on the respondent by having a very high level of choice tasks, a special estimation method is used for these part-worths. This method will be introduced in Section 2.1.2. This method releases the burden on the respondents by using information from other respondents to help estimate the part-worths for specific respondents (Allenby & Rossi, 1998).

Another very important part of CBC Analysis is which combinations of features the respondents see in every task, this combination is called a design. There are many different ways to design these CBC tasks. However as this is not the scope of the research, I refer to Chrzan & Orme (2000) for a more in depth explanation of the design process of CBC tasks. A relevant aspect of these designs to discuss is the fact that a CBC task can either be on a subset of the attributes (partial) or contain all attributes (full-profile). In this research full-profile is considered (Train, 2009). So every alternative is a combination of levels for all attributes considered in this research. Furthermore, not all respondents see the same design. There is a certain set of designs available, of which each respondents sees only one (Chrzan & Orme, 2000).

2.1.2 Choice Based Conjoint Analysis Estimation

The input for the estimation are design of tasks and alternatives the respondents saw as well as the choices they made in each task. The desired output are the best possible estimated part-worths. So the question is, how to go from this input, to the desired output. First the model is of interest. The model for these part-worths is a multinomial logit model (Train, 2001). Then the question is how to estimate these part-worths using this model. To answer this question, several aspects of the CBC data have to be considered. The first question to consider when deciding on an estimation method is whether the model accounts for heterogeneity. The data obtained from a CBC survey relates to many respondents and their personal choices. No two respondents are exactly the same, which means their personal part-worths are also not exactly the same. The variability in the respondents leads to heterogeneity in the data (Allenby & Rossi, 1998). This heterogeneity is of big interest in consumer marketing, as capturing the heterogeneity allows for targeting and specialised marketing methods (Allenby & Rossi, 1998). However as Allenby & Rossi (1998) state, attempting to capture the part-worths of every individual separately requires a lot of information on every individual. This information is furthermore difficult to obtain as respondents are not willing to answer high amounts of questions. In conclusion, the objective of catching heterogeneity also leads to new difficulties in the estimation. To solve these difficulties or work around them, several estimation methods are available.
In CBC Analysis, the two often times used methods are Hierarchical Bayes (HB) and Latent Class segmentation (Magidson & Vermunt, 2007). Latent Class Segmentation in CBC Analysis, as introduced by DeSarbo et al. (1995), does not allow for the heterogeneity but instead captures the heterogeneity by placing the respondents in certain groups (Bhatnagar & Ghose, 2004). This places the respondents in specific groups, within these groups heterogeneity may still occur. As the heterogeneity of the data set as a whole is an interesting aspect of the data in terms of marketing and because I intent to find respondent specific human errors, I want to allow for the heterogeneity. This leads to the method of Hierarchical Bayes (HB) (Hill, 1965), which is able to allow for the heterogeneity and enables us to estimate the part-worths of all individual respondents (Natter & Feurstein, 2002). HB attempts to solve the problem of insufficient individual data by including additional information when estimating the part-worths. This additional information is known prior to the data collection and estimation. In this case, the prior information is information about the group of respondents as whole. The exact way in which HB is able to allow for heterogeneity is explained

in Section 3.

2.2 Human Errors

To introduce the concept of human errors, I first discuss the possible causes of these errors. From a psychologist point of view, as Edmondson (2004) states, a respondents mistake could have several underlying factors. These factors include ‘perceivers’ expectations’, not paying sufficient attention or being influenced by emotions. Spiro et al. (2017) combines many papers. In this general work, Rumelhart (2017) introduces the phenomenon of ‘perceivers’ expectations’ by discussing specific types of obtaining knowledge. Adams (2017) examines this knowledge processing in reading behavior, which is relevant for survey settings. This reading behavior is examined specifically in children. Children are less-experienced readers, they showcase the effect of bottom-up and top-down reading. Top-down reading is reading the literal text as you would expect it to be, where unknown words are filled in based on the expectation of what this word should mean. Bottom-up reading allows the brain to identify the parts where the text differs from the expectation. For experienced readers these processes complement each other, for less-experienced readers, one of the two processes can dominate. In many children this will be the top-down method, which leads to readers interpreting a text based on previous knowledge but failing to differentiate between different meanings of the same word or for instance correctly interpreting the difficult structure of a sentence. This will occur mainly in difficult or less used structures and sentences. The difficulty here is that a reader will oftentimes be able to recognize the case where he/she did not understand a certain word, but will not be able to recognize the case where he/she misinterpreted the structure of a sentence. If this sentence is a question, this can lead to a human error in the answer based on the misreading.

Norman (1981) explains the concept of making a mistake because of not paying enough attention, specifically when a task has been performed many times before. This behavior can be explained as a ‘slip’. Because someone has done the same action many times, this action has been trained. When the individual then wants to switch to another action, they may accidentally end up continuing the trained action instead of starting the new action. The task someone is performing has been performed so often, switching from this task will need a certain level of attention to remember to make this change. Otherwise the individual will automatically continue and complete the task. Norman (1981) gives the example of someone riding home by car and intending to differ from the route at some point to stop at the supermarket, but because they were not actively paying attention, they find themselves not taking the turn and driving home immediately. This slip however refers to a very trained behavior, as the CBC of a survey is only a relatively short task, I do not expect this type of error to occur often. Especially not when a respondent is consciously filling out the survey. However if a respondent were, for instance, to select the most right-hand choice a few times, it could be possible he will click the right-hand answer again without thinking about it because of repeat behavior. These slips are also explained to extend beyond the trained actions, a respondent could click a certain answer because they were thinking about a certain word in the features or because someone next to them said the word ‘right’, there are various ways to fall into trained behaviour (Norman, 1981).

Aside from the before mentioned errors, it is of course also possible for a respondent to be temporarily distracted for a single question. By for instance someone talking to them, and this causing them to click a different answer then they would have clicked when they had not been distracted.

All mentioned errors will be considered in this paper.

2.3 Error Detection Methods

This research relates to recognizing errors in survey data. Although there are many papers available on recognizing unreliable respondents, as I discuss in Section 2.3.1, finding invalid answers within valid respondents has not yet been discussed widely in the literature. This direction of research seems new as on this exact topic, no previous papers are found by the author. Related research is discussed in Section 2.3.2.

2.3.1 General C/IER methods

As respondents might not answer truthfully or reliably in CBC surveys, it may be preferable to flag unreliable respondents before any analysis is performed. These respondents can be removed beforehand, as their choices may decrease the performance of the estimates (Pokropek et al., 2022; Johnson, 2005). The unreliable respondents answers may act as noise when fitting a model to the data. The recognition of these respondents relates to the detection of ‘Careless or insufficient effort responding’ (C/IER) as introduced by Johnson (2005). Pokropek et al. (2022) discusses several measures which could be evaluated and combined to find the careless respondents. These methods among others include; special questions to check whether respondents are paying attention, the time they take to answer questions, their consistency etc. (Pokropek et al., 2022). Huang et al. (2012) also discusses several of these methods, combining them into 4 categories. Namely ‘infrequency’, ‘inconsistency’, ‘pattern’ and ‘response time’. Especially the last 3 categories are relevant for this research, as at SKIM these methods are most often applied to recognize invalid respondents. These relate both to the general survey as well as the CBC. Inconsistency means answering questions not as expected based on previous answers. Pattern relates to the respondent choosing their answers following a certain pattern instead of answering truthfully. Response time may relate to answering too quickly or perhaps taking very long and being distracted. For the recognition of unreliable respondents, as Pokropek et al. (2022) state, no perfect technique is available. By combining several measures you find relevant, you can find the best method to remove invalid respondents. Still even with the discussed measures, there are contrasting arguments questioning for instance whether inconsistent answers automatically indicate invalid respondents (Kurtz & Parrish, 2001).

2.3.2 Answer Specific Error Detection

Unreliable respondents and human errors differ, so their detection methods may also have to differ. Within a survey, a respondent can make a mistake, but a respondent can also just not fill out the survey actively or purposely answer randomly. As stated, when researching specifically human errors in Choice Based Conjoint data, no previous research has been found. A direction of research which might relate to the topic of finding errors in otherwise valid respondents is for instance user fatigue (Bradley & Daly, 1994), where a respondent becomes more likely to make an error as the survey progresses due to fatigue. Fatigue is shown to occur in online surveys by Savage & Waldman (2008). However, Hess et al. (2012) show this worry may be larger than necessary. In their research the fit of the estimates does not differ significantly over various parts of the choice tasks. The estimations do not seem to perform significantly worse or better at the beginning or end of the survey. Furthermore, outside of conjoint data sets, Chandola et al. (2009) discusses, in a very broad setting, which unexpected data or specific error instances in a data set can be detected. Following this paper, since the occurrence of an human error depends on the preferences of this specific human, the human errors can be seen as a 'contextual anomaly'. In this case, the context is the individual. Furthermore, it is not known if an error occurred, the error detection therefore has to be unsupervised. For this type of error, Chandola et al. (2009) proposes for instance reverting the individual error to an error which can be recognized in the entire data set, so one might use the context of an individual to derive which outliers for an individual may also be an outlier in the overall data set. Furthermore, one can implement a neural network or clustering technique or use other statistical methods to express the likelihood of a specific data point being an error. Although these ideas are very general, in my methods I will attempt to use statistical methods to express the likelihood of an error having occurred.

3 Methodology

In this section I firstly introduce Hierarchical Bayes. Then I state the general model and estimation method. I then extend the model and state the estimation method for the extended case. Finally I introduce the evaluation metrics.

3.1 Hierarchical Bayes

The methodology I use for the estimation of the multinomial logit model as introduced in Section 2.1.2, is Hierarchical Bayes (HB). The HB used follows known literature, primarily Train (2009); Regier et al. (2009); Sawtooth Software (2021); Paap (2021b). As stated before, HB uses the observed choices of all individuals and the tasks they saw to estimate the part-worths on an individual level. I will refer to part-worths as β, where β_i is the vector of part-worths for individual i. The choices of the individuals are seen as Y and the design is X. The problem which arises when attempting to estimate on an individual level is the fact that we need a high amount of data for every individual respondent. HB attempts to solve this problem by using Bayesian econometrics. This type of mathematics is based on different assumptions than we usually follow in statistical methods (Greenberg, 2012). More specifically, Bayesian econometrics assumes uncertainty for parameters. Although generally used statistical methods cannot overcome the problem of insufficient data, this new statistical direction is able to propose a solution. The next section explains this statistical direction in more detail. In Section 3.1.2 I show how Bayesian econometrics can be used when working with CBC data.

3.1.1 Bayesian Econometrics

In the most used (frequentist) statistics, the parameter we attempt to estimate is assumed to be fixed (Bolstad & Curran, 2016). As Greenberg (2012) explains, we can then estimate the interval that covers the true parameter with a certain probability. So the interval captures uncertainty. We can approach this true parameter by repeatedly sampling from the data and obtaining the probabilities related to the interval. Bayesian econometrics does not agree with the concept of repeatedly sampling from a model to estimate the parameter, as the world and models around us are as we know them. We can not resample the world, so instead of resampling the world we accept that with the current knowledge of the world that we have, we may be uncertain about the parameter itself (Bolstad & Curran, 2016). The parameter is not fixed but has randomness.

Because we do not try to estimate the fixed parameter but work with probability, we can use new information to update our beliefs. We start from a certain belief about the parameter, then because of the data at hand, new or additional information may lead to a different belief. Using this principle, we can also use information we already know about the parameter as additional information before estimating. We then simply use the actual data afterwards to update our beliefs. For instance, when estimating the parameter for the price-elasticity of a product, a professional may tell us the value is likely to be between 0 and 0.5. We can start from this value and then use the obtained data to update the estimate. In Bayesian terms, the information from which we start is called a *prior*. A prior should be stated before seeing the actual data. When the data is then used to update the prior we obtain the *posterior*. Because of this prior, we can add information to the known data, thus allowing for estimations in sparse data sets. It should be noted that more information included in the prior means more information available for the estimation, but also a bigger influence of the prior information on the final estimate. So a wrong prior can lead to a bad estimate.

Putting this in mathematics,we are interested in the probability distribution of the parameter, given the observed data. We want to know $P(\theta|y)$. To obtain this distribution, Bayes'theorem is used as can be seen in Equation 1 (Greenberg, 2012; Train, 2009).

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{1}$$

With $P(\theta)$ as the pre-specified prior on the parameters and $P(y|\theta)$ as the conditional likelihood of the data, given the parameter. $P(\theta|y)$ is here the conditional posterior distribution. Equation 1 can be reduced to Equation 2

$$P(\theta|y) \propto P(\theta)P(y|\theta) \tag{2}$$

This is the basis of Bayesian econometrics, Hierarchical Bayes takes the usage of priors a step further. When we implement a prior on a parameter (for instance on θ) there may also be prior knowledge available on the prior $P(\theta)$ itself. θ may also depend on an unknown parameter (Greenberg, 2012). Take λ as the parameter on which θ depends, we can then state the prior as $P(\theta|\lambda)$. If we see λ as a unknown parameter, λ can also have a prior specification. This concept of priors on priors is called Hierarchical and leads to Hierarchical Bayes.

These formulas and ideas are the basis of Bayesian econometrics. The problem of the sparse data when allowing for heterogeneity in the data set is solved by the use of a prior. For survey data sets specifically, the prior can use the information about other respondents within the survey to increase the amount of information available for the estimation of one specific respondent. The next section further explains how the concept of Hierarchical Bayes is used to estimate the part-worths for all individuals in a CBC data set.

3.1.2 The Bayesian approach in CBC data

To explain the estimation method in CBC data, I firstly specify the model. For this specification and the estimation method, I follow Train (2009); Regier et al. (2009); Gelman et al. (2013); Sawtooth Software (2021) and Paap (2021b). The choice an individual makes in a task is expected to be the alternative with the highest utility. For a single individual i, the utility of a certain alternative j in task s is U_{ij}, with

$$U_{ij} = x'_{sj}\beta_i + \epsilon_{isj} \tag{3}$$

Where ϵ follows an extreme value distribution, namely the Gumbel distribution. β_i contains the part-worths for individual i and x_{sj} is the design of alternative j in task s. The probability of an individual choosing a certain alternative within a task is then a combination of all alternatives in this task. Because of the Gumbell error, for an unknown part-worth, this probability follows a multinomial logistic distribution. Which makes the overall choice behavior of the individuals, given their part-worths, a multinomial logistic distribution.

$$Y_i \sim MNL(\beta_i) \tag{4}$$

Where Y includes all choices of all individuals and β are the part-worths. The probability that a certain alternative j is chosen within choice task s by individual i is

$$Pr(Y_{is} = j) = Pr(U_{isj} \geq U_{ism}, \forall m) = p_{isj} = \frac{exp(x'_{isj}\beta_i)}{\sum_m exp(x'_{ism}\beta_i)} \tag{5}$$

Where n is the total number of respondents. S is the number of choice tasks presented to each respondent Furthermore, m is the number of alternatives included in each choice task and k is the amount of features included in the alternatives. The choices are caught in Y, this is a nxS matrix. For every row Y_i, each value of the row can take a value of 1 to k. Indicating which alternative was chosen in this specific task.

The objective is to estimate the part-worths, β. I attempt to estimate nxk part-worths, and I have (N x S) observations. The number of observations may not be sufficient to accurately estimate the part-worths as is. Using priors can help solve this problem. As I attempt to estimate β, the first action would be to put a prior on this. As it is not known what the exact distribution is of β, I assume a certain distribution. This prior for β_i is:

$$\beta_i | b, \Sigma \sim N(b, \Sigma) \tag{6}$$

With b and Σ as the mean and variance. For the b and Σ I also assume prior distributions, which contain information about all respondents combined. The priors for b and Σ are as follows:

$$\Sigma \sim IW(vI_k, v) \tag{7}$$

$$b \sim N(0, \sigma_b^2 I_k) \tag{8}$$

where, setting the amount of part-worths to estimate as p, the following has to hold;

$$v > p + 1 \tag{9}$$

When setting these priors for the estimation, a smaller variance in the distribution will lead to the prior influencing the posterior more. I attempt to set the variance in the priors high, this ensures little information is included in the prior and the results are not influenced by the priors (Train, 2001). The v should however also not be too large, as this leads to a possibly incorrect estimation (Rossi & Allenby, 2003). I therefore set v as,

$$v = p + 2 \tag{10}$$

Following the same reasoning, I set σ_b to 1.5 since this leads to a relatively large variance and a spread out prior. The used documentation does not implement a prior on b, I differentiate from the stated literature by using a prior which is informative. I use this prior to restrict the b in the case of many errors or otherwise random data.

From the likelihood $P(y|\theta)$ as given in Equation 4 and the prior specifications, I can derive the posterior distribution, using Equation 2. There are known conditional posterior distributions for both b and Σ, conditional on $\beta_i, i = 1, ..., N$. For β_i the posterior distribution is not known. The following conditional posterior distributions is derived from Train (2009),

$$p(\Sigma | \beta, b, Y) \sim IW(p + 2 + N, I_k + N * S_1) \tag{11}$$

$$S_1 = \frac{1}{N} \sum_{i}^{n} (\beta_n - b)(\beta_n - b)' \tag{12}$$

and the conditional posterior distribution for the mean can be derived from Train (2009) as

$$p(b | \beta, \Sigma, Y, \sigma_b) \sim N(\sum_{n=1}^{N} \beta_i * ((\sigma_b^2 I_k)^{-1} \Sigma + N)^{-1}, (N\Sigma^{-1} + (\sigma_b^2 I_k)^{-1})^{-1}) \tag{13}$$

To then estimate the part-worths, using these conditional posterior distribution, it should first be noted that the posterior distributions are dependent upon each other. To sample from these distributions, a recursive method called the Gibbs method is implemented. Recursively for the data set, b is first sampled from Equation 13 using the previous estimates for Σ and β. Then Σ is sampled from Equation 11, using the new estimate for b and the previous estimate for β. For the sampling of Σ I use a package from Statisticat & LLC. (2021). Finally β has to be sampled, using the new b and Σ. If this is iterated, one in the end obtains draws from the true posterior distribution. However, as the conditional posterior distribution for β_i is not known, it is not possible to directly sample from this distribution. For this sampling, an additional step is implemented.

Sampling β_i has a problem of insufficient data as introduced before. However since β_i is dependent on the other posterior distributions, which use data about all respondents in their estimation, the amount of data available for the estimation of β_i is increased. The exact conditional posterior distribution for β is still not known. To be able to estimate β without this known distribution, I implement a Metropolis-Hastings (MH) step. In this algorithm, the β for every individual will be updated following a random walk, and accepted with a certain probability.

The entire Gibbs estimation is repeated recursively until the draws converge to a stable distribution. The obtained draws from β are the part-worths for every individual. These β_i's can be seen as a sample of the true posterior distribution.

As stated before, within the Gibbs algorithm the MH step is present. This step proposes a new β_i vector for every i, independently of each other. Then it decides for each of the proposed β_i's, whether it is accepted or not. If the new draw is not accepted, the β_i remains the previous draw. This particular MH algorithm uses a random-walk principle. So given the current estimate β_i^o, a new estimate β_i^n is proposed.

$$\beta_i^n = \beta_i^o + \rho * d \tag{14}$$

Where d is a random sample from the Normal distribution with a mean of 0 and a variance proportional to Σ. And ρ is an hyperparameter, which influences the size of the steps taken in the random-walk. The ρ has to be set to a certain value. Based on literature, this value should lead to about 23% acceptance of the new beta (Train, 2001). I set ρ to 0.70.

To decide whether to maintain the previous estimate of β_i or to update to β_i^n, a comparison is made based on the conditional posterior distribution, which is the likelihood of the found parameters occurring, given the data. The conditional posterior distribution equals

$$P(\beta_i|b, \Sigma, Y_i) \propto P(Y_i|\beta_i, b, \Sigma)P(\beta_i|b, \Sigma), \tag{15}$$

where $P(Y_i|\beta_i, b, \Sigma)$ is the likelihood of the data given β_i and $P(\beta_i|b, \Sigma)$ is the relative density of the distribution of the parameters of interest. The comparison, which determines whether to update, is done by using a proportion which can be seen in Equation 16.

$$r = \frac{\frac{p_n}{d_n}}{\frac{p_o}{d_o}} \tag{16}$$

With p_o the current likelihood of the data, and p_n the likelihood of the data given the proposed β_i. Furthermore, d_o is the density of the current β_i and d_n that of the proposed β. The likelihood of the data, p, given β_i is calculated by multiplying the probabilities of all decisions occurring. The relative density is calculated through Equation 17.

$$d \propto exp[-\frac{1}{2}(\beta - b)^{'}\Sigma^{-1}(\beta - b)] \tag{17}$$

Finally, β_i^n replaces the old β_i^o with a probability of r. So the new β_i^n is more likely to be accepted when the probability of this parameter occurring, given the data, is better or not much lower than that of the previous β_i^o.

All valid responses of individuals are expected to follow the specified model, which I refer to as the general model. The general model also leads to the estimation method as explained. However, when a human error occurs, the model is not exactly followed. The proposed extension on this model to allow for human errors is explained in Section 3.2.1. The extension on the estimation method, I introduce in Section 3.2.2.

For the general model and the extended model, the estimation algorithms itself are programmed in R. For the initialisation I set Σ as an identity matrix and β as a matrix of vectors following $\beta_i^{initial} \sim N(0,1)$. b is then the average of this matrix over all respondents. These initialisation values I only set once, for every separate simulation I implement the same initialisation values.

3.2 Model extension

In this section I first introduce the extended model, after which I explain the proposed extended estimation method.

3.2.1 The extended model

To allow for the effect of a human error, I extend the model with the probability that an error has occurred in a specific choice task s for individual i. The extension of the model therefore takes place in the probability with which a respondent chooses a certain alternative. I extend Equation 5. To extend this probability, I need to know what the probability of an error occurring is. This error probability is not known and has to be estimated. The probability of an error occurring in task s for respondent i, I define as q_{is}. To maintain a proper probability, with values between 0 and 1, I specify this probability as a logit function, dependent on a certain γ. The γ is a parameter which has to be estimated. This leads to the definition of the probability of an error

$$q_{is} = \frac{e^{\gamma}}{1 + e^{\gamma}} \tag{18}$$

An error can either occur or not occur. For both cases I need to set the probability that alternative j is chosen. When no error occurs, this probability is already given by Equation 5. This conditional probability can be seen in Equation 19. When an error occurs this probability differs. For simplicity I assume the respondent to have an equal probability of choosing all alternatives in a task when an error occurs. This can be seen in Equation 20. This assumption is more likely to hold for smaller choice tasks, where choosing an answer near the intended answer leads to nearly the same effect as choosing a random answer. For this research it holds that $s = 4$, which is relatively small. It should be noted that repeat errors behave differently and are not specifically accounted for in this model.

$$Pr(Y_i s = j | error = 0) = p_{isj} \tag{19}$$

$$Pr(Y_i s = j | error = 1) = 1/m \tag{20}$$

Combining the probabilities, leads to an updated probability of alternative j being chosen. This probability is

$$Pr(Y_{is} = j) = p_{isj} * (1 - q_{is}) + \frac{1}{m} * q_{is} \tag{21}$$

The probability of an error occurring is here set over all choice sets and individuals, since γ does not depend on i, s or j.

The extension on the model, mentioned in Section 3.2, also requires an extension on the estimation method.

3.2.2 Human error recognition extension

To capture the effect of a human error in the algorithm, the HB method is extended. As introduced in Section 3.2.1, the parameter γ has to be estimated. This estimation is an additional step in the recursive Gibbs algorithm since β depends on γ for the likelihood in the acceptance ratio. After estimating b and Σ, I first estimate γ before estimating β. The addition of γ in the model leads to the extension of two parts of the Gibbs algorithm. Firstly, within the β estimation, the extended probability leads to the ratio as shown in Equation 16 to be adjusted. Equation 15 becomes;

$$P(\beta_i | b, \Sigma, \gamma, Y_i) \tag{22}$$

Additionally, equation 5 in this setting becomes Equation 21.

Furthermore, the estimation of γ itself requires an additional step in the Gibbs algorithm. As the posterior distribution of γ is not known, the sampling also follows a MH step. I set the prior on γ,

$$\gamma | \beta, \Sigma, b \sim N(b_\gamma, \Sigma_\gamma) \tag{23}$$

I set $b_\gamma = -2.5$ and $\Sigma_\gamma = 1$ as this restricts the probability of an error occurring to reasonable values. As an effect, the γ is pulled slightly towards 2.5, preventing the γ from going to extreme negative values. In these large negative values, the q_{is} stays relatively the same for large differences in γ so these values are not informative and prevent convergence. This leads to a random walk algorithm, where the proposed gamma (γ^n) is

$$\gamma^n = \gamma^o + \delta * d \tag{24}$$

Where d is a random sample from the Normal distribution with a mean of 0 and a variance 1. And δ is an hyperparameter, which influences the size of the steps, taken in the random-walk. The ratio with which this proposed γ is accepted can be calculated as explained in Section 3.1.2. However instead of a ratio per respondent, the likelihoods are multiplied over all respondents, leading to one general likelihood.

I initialise γ on -2 to start from a low probability or error. δ is set to obtain a acceptance probability of 30%, as is proposed for the acceptance in Metropolis-Hastings steps by Calderhead (2014). To obtain this acceptance probability, δ is set to 0.25.

3.3 Performance evaluation

This section first discusses how I test for convergence of the estimation. Then I introduce the metrics I use to compare the performance of all models.

3.3.1 Convergence

For all models, I take 20.000 draws. the first 10.000 are considered burn-in and not retained and next, only one in every 5 draws is saved. So the thin-factor is 5. This results in 10.000 draws and 2.000 data points for every part-worth. To test if the part-worths have converged, I refer to the methods discussed in Paap (2021a); Roy (2020). The first method is visual inspection. The draws should move around a horizontal line. If the visual inspection seems to hold, I use Geweke's metric (Plummer et al., 2006). This metric assumes that for enough draws, the mean of the first part of the samples should be about the same as the mean of the last part of the samples. The first part is here generally 10% of the draws and the last part the final 50%. However, since I thin my values before applying the diagnostic and therefore may expect a less smooth distribution, I take 20% for the first part. This takes more information into account and lessens the chance of rejecting the hypothesis of convergence because of a small bump in the estimates. The metric is

$$CD = \frac{\hat{m}_1 - \hat{m}_2}{\sqrt{\hat{S}_1^2 + \hat{S}_2^2}} \sim N(0,1) \tag{25}$$

Where $\hat{m}_1$ is the mean of the first part and $\hat{m}_2$ is the mean of the second part. And $\hat{S}_1^2$ is the variance of the first part, $\hat{S}_2^2$ is the variance of the second part.

For the implementation I use the package as provided by Plummer et al. (2006), which accounts for the autocorrelation within the draws.

3.3.2 Evaluation Metrics

To compare the general model and the extended model in terms of their performance, I first have to define the performance measures. The performance of the algorithm is evaluated on simulated data. The methods both have the objective of estimating the part-worths for the individuals as well as possible. This can be measured in several ways. The first measure is the hit-rate as done by Moore (2004). Given the parameter estimates, the choices can be predicted. The hit-rate then measures how accurately the model predicts the true choices. For this metric, I first simulate new choices without human errors. For the simulation I use the β as used in the original simulation to ensure the respondents behave comparably. I then calculate the hitrate through

$$hitrate = \sum_{i=1}^{N} \sum_{s=1}^{S} Q_{i,s} \tag{26}$$

Where $Q_{i,s}$ is a dummy, which is 1 when the predicted choice for the task s of the respondent i matches the true choice in this task. This hit rate is of great interest, as it indicates which model is better able to predict the choices a respondent makes. If the hit rate goes up when the part-worths were obtained whilst accounting for errors, this may indicate the model is indeed able to catch these human errors. The part-worths then predict the choices of the respondents better than in the case where part-worths are obtained with a method in which the human errors are not accounted for.

Furthermore, the likelihood of the data, given the estimated part-worths and γ is of interest. This likelihood can be calculated using Equations 5 and 21.

$$LogLikelihood = \sum_{i=1}^{N} \sum_{s=1}^{S} log(Pr(Y_{is} = j)) \tag{27}$$

where j is the actually chosen alternative in choice task s for individual i. A higher likelihood indicates the true data, given the estimated part-worths (and γ) is more likely. Moore (2004) also use the likelihood as a measure of comparison, in their case it is used in the log marginal density. I compare the likelihood on newly simulated data sets with human errors using the same human error settings as the simulation from which the estimates are obtained.

Finally since I not only know the true choices, but also the true underlying part-worths, I can compare these. This is done by using the Mean Squared Error (MSE), following Otter et al. (2004).

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\beta_i - \hat{\beta}_i)'(\beta_i - \hat{\beta}_i) \tag{28}$$

where β_i are the true part-worths for respondent i and $\hat{\beta}_i$ the obtained estimate for this respondent. Optimally, the MSE would be 0 as I then estimate the part-worths perfectly. In general it holds that a lower MSE is preferred.

Of these three evaluation metrics, I give the most importance to the MSE. Since the overall objective of this paper is to clean the estimates from the effect of human errors. The MSE shows how close the estimated parameters are to the true parameters and therefore gives a good indication of this performance. However in a more practical setting, the hitrate could be considered as most important since the objective of, for instance, marketeers can be to predict the choices which respondents make as good as possible.

4 Data

To evaluate the effect of human errors, I need a 'clean' data set to include these errors in. As discussed, this data set is generated in a simulation. I include several simulations to compare the performance of the methods. The basis of these simulations is a design which the respondents saw and a general assumption for what the respondents part-worths look like. I first discuss this data generation and then discuss the various parameters I alter throughout the simulations.

4.1 Data generation

I simulate the decisions of n respondents for a set of s given choice tasks, given a certain design. Firstly, these responses are simulated following exact distributions. Then, I alter certain responses to simulate human errors.

4.1.1 Design

In terms of the design, I use the design generator of SKIM (Sawtooth Software, 2021) to create the combinations of choice tasks and alternatives. I create 50 designs, each design consists of 25 choice tasks with 4 alternatives within each choice task. The alternatives in these designs consist of a combination of 4 attributes, divided over 10 levels. The n respondents all see one of the 50 designs when answering the survey questions.

4.1.2 General simulation

I assume known distributions for the utilities of respondents. As shown in Section 3.1.2, under the assumption of no errors occurring, the probability an individual chooses a certain alternative in a choice task follows a multinomial logistic distribution, dependent on the part-worths of this individual. These part-worths are assumed to have a normal distribution, based on a certain mean and variance. So, given a certain mean and variance, I can generate part-worths and subsequently simulate choices. I assume

$$\beta_i \sim N(\alpha * \mu, \alpha^2 * \phi * \Sigma) \tag{29}$$

As I want to adjust some settings in the simulations of the part-worths and human errors, I rewrite this setting to;

$$\beta_i = \alpha * \mu + \alpha * \sqrt{\phi}\Sigma^{\frac{1}{2}} * \epsilon_i \tag{30}$$

with

$$\epsilon_i \sim N(0, I) \tag{31}$$

Here α and ϕ can be adjusted over the part-worths. α influences the absolute size of the part-worths. A larger part-worth means the choices of the respondents are more likely to follow the part-worths. This is the case because choices are based on a utility as seen in Equation 3. When the part-worth is relatively big as compared to the gumbell error, the error will have less influence on the choices. So there is then also relatively less randomness in the choices of the respondents. ϕ influences the variance of the part-worths. When ϕ becomes larger, the variance within the part-worths for an individual also becomes larger. I will discuss these settings in Section 4.2.1. The μ and ϵ_i however, will not be adjust over the simulations. To be able to compare as well as possible between the results, I set these vectors to set values for all individuals. I simulate the part-worths for 1000 respondents.

Firstly the μ or the 'true', average of the part-worths is obtained by taking an normal distribution with a mean of 0 and variance of 1 for the $10 - 4$ part-worths. This mean of the part-worths can be seen in the following matrix.

$$\begin{bmatrix} Level1.1 & Level1.2 & Level2.1 & Level3.1 & Level3.2 & Level4.1 \\ -0.27657708 & -1.26975977 & 0.05473915 & -0.10121495 & -1.16606773 & 2.06502502 \end{bmatrix}$$

For the ϵ I take a normal distribution with a mean of 0 and a variance of 1 for the 6 part-worths as well. I however do this for every respondent and combine these vectors in one matrix. For every simulation, I can now use these set inputs.

I then for every choice task calculate the probability p_{isj} that a certain alternative j is chosen in task s by individual i. This probability is calculated by firstly obtaining the utility of each alternative following Equation 3, where I sample ϵ from the Gumbel(0,1) distribution. The alternative with the highest utility in each choice set is then set to 'chosen'. This leads to the simulated y.

4.1.3 Implementing Human Errors

As based on Section 2, I assume three main causes of human errors. 'perceivers' expectations', not paying enough attention and being unknowingly influenced by emotions. I follow these causes to implement human errors in the data set. Firstly in terms of lack of attention, this can result in randomly clicking another answer cause you were not paying attention. This is simulated by, for every answer, having a very small chance of choosing a randomized answer (where this could include the simulated answer). This probability will increase slightly as more questions have been answered, as we assume a respondent may lose some focus due to user fatigue. Furthermore, a respondent can mis-click by lack of attention as opposed to the one he wanted to click. I assume this behavior to lead to the same effect as randomly clicking, as in the used designs the choice tasks only contain 4 alternatives. Finally lack of attention can also result in a 'slip' of behavior. Where a certain action is repeated because one has performed this action oftentimes before. This is simulated by a probability of choosing the last given answer again.

In terms of emotions, this cannot be simulated so I expect these mistakes to be caught in the lack of attention or random choice instance. Finally for the misunderstanding of a question, I assume this probability to be very small in CBC tasks as it is mainly choice based. I therefore do not simulate this human error.

Given the simulated y, I adjust the true choices to simulate the human errors. I simulate both random choices and repeat choices with a certain probability. The probability of such an error occurring increases over the survey. This increase happens with a factor of $1.05 * Prob_{previous}$ for every choice task. I alter the probabilities to analyze the influence of each error on the performance of the model. This leads to a total of 10 simulations which are discussed in the next section.

4.2 Simulation settings

The general objective of the methods is to capture the part-worths, however these part-worths may differ for different types of respondents. To be able to compare the implemented methods for several types of response behavior, I alter the data for four hyperparameters. As introduced in Section 4.1, these hyperparameters include the size of α, ϕ, the probability of a repeat error and the probability of a random error. For every hyperparameter, I compare two values against a base case.

4.2.1 Adjusting α and ϕ

As stated before, larger α leads to less randomness in the respondent choices. The choices therefore are more predictable, which leads to easier to estimate values. For a smaller α and therefore smaller part-worth, the opposite holds. The choices become more random and the part-worths more difficult to estimate as they explain a smaller part of the choices. The part-worth estimates may be around zero for large error. For my simulation I want a data set where the part-worths can be estimated well, but the choices still contain some randomness. So the effect of the part-worths compared to the error should be bigger. But the error should still have an influence. I take a base size of 0.7 for α, furthermore I test 1.2 and 0.2.

Where α influences both the value and variance of β_i, ϕ only influences the variance. So a larger value for ϕ indicates a bigger variance in the part-worth values. Which can be seen as more differing preferences within respondents. I set the base value of ϕ to 1 and differ with 1.2 and 0.8.

4.2.2 Probability of the random error

A random error refers to the respondent choosing an alternatively randomly instead of choosing the intended one. This could result in the intended answer being chosen, but also another alternative. The standard probability of this error occurring is expected to be small, especially for shorter surveys. The probability that an error occurs however may increase as the amount of tasks increases due to for instance user fatigue. This increasing probability is simulated in the data. The initial probability of a random error is set at 0.025 and the alternative simulations include 0.01 and 0.05.

4.2.3 Probability of the repeat error

A repeat error refers to the respondent choosing the alternative not based on their intended answer, but by clicking the alternative in the same position as the previous choice task. Again this probability is expected to be small and may increase. The initial probability is again set at 0.025 and the alternative simulations include 0.01 and 0.05.

I combine the mentioned levels of hyperparameters to obtain the following simulations:

Table 1. Hyperparameter settings

simulation	specification	α	ϕ	random error	repeat error
1	No human error	0.7	1	0	0
2	Base case	0.7	1	0.025	0.025
3	random error larger	0.7	1	0.05	0.025
4	random error smaller	0.7	1	0.01	0.025
5	repeat error larger	0.7	1	0.025	0.05
6	repeat error smaller	0.7	1	0.025	0.01
7	α larger	1.2	1	0.025	0.025
8	α smaller	0.2	1	0.025	0.025
9	ϕ larger	0.7	1.2	0.025	0.025
10	ϕ smaller	0.7	0.8	0.025	0.025

5 Results

This section discusses the results as obtained from the proposed simulations and methods. In Section 5.1 I discuss the data obtained from the simulations. In Section 5.2 I then introduce the results from the estimations.

All methods are implemented in R (RStudio Team, 2021). The methods are implemented from scratch where possible, leading to less efficient methods as compared to the packages available. However the coding allows for more control over all factors of the estimations.

5.1 Simulations

The simulations take in a set value for the mean of the part-worths and the errors, based on which the simulation is performed. The simulations therefore contain the same part-worths for all cases where this part-worth sampling is not altered. Only where the α or ϕ is altered, do the part-worths differ. The exact mean part-worths can be seen in Table 2 and Table 3.

5.2 Estimation results

To answer my research question, I am interested in the performance of the extended model as opposed to the general (HB) model. As well as the performance of both models on their own. The results I obtain for the estimation are evaluated on their values themselves firstly, I also discuss the convergence of the estimates and the evaluation metrics. I then evaluate the obtained acceptance probabilities and finally compare the run times of both estimation methods.

5.2.1 Estimates

The estimates of the part-worths result in $(N\text{x}k)$ estimates. I do not evaluate these all separately, instead I use the estimation of the posterior mean of the part-worths as the estimate to compare on. This mean contains k estimates to analyse for every simulation and model. The posterior standard deviation of these means is taken over the last 10.000 iterations of the models. The estimate is the mean of the means over these last 10.000 iterations

For the general model, the estimates are shown in Table 2. The standard deviations are relatively constant for all part-worths. Furthermore, introducing human errors seems to pull the estimated part-worths to zero, compared to the part-worths in the simulation without errors. The effect is seen in the estimates for β_1, β_2, β_4 and β_6. This effect is as expected and may be caused by the fact that the model is not able to estimate as extreme values as possible when there are no errors. Since the errors influence the probability of certain estimates. Expanding on this effect, it holds that for almost all simulations and parameters, including errors pushes the estimates further away from the true value as opposed to the simulation without errors. It can also be seen that the effect of a higher probability of random errors, as opposed to a higher probability of repeat errors, also leads to different estimates. So it can be concluded that the effect of these errors is not the same on the model and its estimates.

For the estimates of the extended model, the results are shown in Table 3. In this table the estimate of q is also included. The posterior standard deviation shows there is more variability in the estimations when errors are added, this difference is however not very large. Furthermore, certain estimates have a higher variance in general than others. For instance, β_6 as compared to β_3. Also, some estimates now seem to actually be pulled towards the true value as opposed to the case of no human errors. This possibly indicates the q is able to catch some of the randomness occurring from the errors. This allows the estimates to be closer to the true value. The estimates for q behave as expected, indicating a very low probability of error when there are no human errors implemented. Furthermore the probability of an error is higher when the implemented error probability was larger and lower when the implemented error probability was smaller. Lastly, the estimate of q for small α is high. The probability of an error is over 3 times larger in this case as opposed to the case where there were more errors added. This indicates the q does not only capture human errors but also other errors or randomness in the data. In this case, a small α leads to higher randomness in the respondent choices as the gumbell error is relatively larger as compared to the true part-worths. The extended model seems to capture some of this randomness in the q estimation.

When comparing the models, the estimates for the part-worths in the 'no error' case are alike. When errors are introduced, the estimates of the models begin to differ. This is as expected, as the general model does not account for the errors and sees them as true choices. Furthermore, most Standard Deviations are larger in the case of the extended model, indicating more uncertainty about the estimates. Finally, although the direction of the estimates as compared to zero is the same in both models, the estimates in the extended model are more often pulled towards the true parameter values in all simulations with errors.

Table 2. The mean of the estimated part-worths in the general model, for all simulations. As compared to their true values

simulation	β_1 (SE)	β_2 (SE)	β_3 (SE)	β_4 (SE)	β_5 (SE)	β_6 (SE)
'no error'	-0.2169 (0.0289)	-0.8477 (0.0298)	0.0123 (0.0276)	-0.0764 (0.0290)	-0.9052 (0.0306)	1.4274 (0.0283)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*
'base model'	-0.2059 (0.0260)	-0.7582 (0.0257)	0.0331 (0.0232)	-0.0440 (0.0245)	-0.6799 (0.0281)	1.2177 (0.0244)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*
'random error larger'	-0.1911 (0.0236)	-0.6559 (0.0234)	0.0445 (0.0219)	-0.0403 (0.0238)	-0.6449 (0.0261)	1.0736 (0.0229)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*
'random error smaller'	-0.2121 (0.0273)	-0.7431 (0.0265)	0.0013 (0.0235)	-0.0374 (0.0285)	-0.7253 (0.0289)	1.2663 (0.0249)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*
'repeat error larger'	-0.1569 (0.0246)	-0.6374 (0.0243)	0.0037 (0.0219)	-0.0641 (0.0250)	-0.6653 (0.0266)	1.1198 (0.0237)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*
'repeat error smaller'	-0.2095 (0.0267)	-0.7771 (0.0271)	0.0210 (0.0242)	-0.0662 (0.0270)	-0.7261 (0.0272)	1.2308 (0.0249)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*
'alpha larger'	-0.3027 (0.0354)	-1.1242 (0.0335)	0.0173 (0.0322)	-0.1245 (0.0343)	-1.0658 (0.0353)	1.8053 (0.0325)
	-0.3973	*-1.5095*	*0.0269*	*-0.1180*	*-1.4341*	*2.4259*
'alpha smaller'	-0.0466 (0.0181)	-0.2227 (0.0159)	0.0308 (0.0161)	-0.0213 (0.0190)	-0.1977(0.0171)	0.3557(0.0160)
	-0.0662	*-0.2516*	*0.0045*	*-0.0197*	*-0.2390*	*0.4043*
'phi larger'	-0.2059 (0.0280)	-0.7599 (0.0273)	0.0181 (0.0248)	-0.0784 (0.0264)	-0.6968 (0.0277)	1.2148 (0.0259)
	-0.2354	*-0.8798*	*0.0135*	*-0.0686*	*-0.8385*	*1.4122*
'phi smaller'	-0.2200 (0.0242)	-0.7297 (0.0248)	-0.0069 (0.0207)	-0.1010 (0.0238)	-0.7103 (0.0253)	1.2028 (0.0231)
	-0.2277	*-0.8814*	*0.0181*	*-0.0690*	*-0.8344*	*1.4183*

Table 3. The mean of the estimated part-worths and γ in the extended model, for all simulations. As compared to their true values

simulation	β_1 (SE)	β_2 (SE)	β_3 (SE)	β_4 (SE)	β_5 (SE)	β_6 (SE)	q (SE)
‘no error’	-0.2161 (0.0284)	-0.8452 (0.0286)	0.0121 (0.0282)	-0.0763 (0.0294)	-0.9072 (0.0303)	1.4289 (0.0270)	0.0000 (0.0002)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*	
‘base model’	-0.2384 (0.0309)	-0.8724 (0.0402)	0.0445 (0.0280)	-0.0518 (0.0296)	-0.7880 (0.0398)	1.3981 (0.0507)	0.0754 (0.0147)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*	
‘random error larger’	-0.2480 (0.0327)	-0.8851 (0.0415)	0.0594 (0.0289)	-0.0488 (0.0318)	-0.8586 (0.0445)	1.4246 (0.0538)	0.1556 (0.0150)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*	
‘random error smaller’	-0.2345 (0.0306)	-0.8183 (0.0365)	0.0041 (0.0270)	-0.0434 (0.0309)	-0.8092 (0.0379)	1.3972 (0.0455)	0.0539(0.0136)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*	
‘repeat error larger’	-0.1841 (0.0290)	-0.7574 (0.0358)	0.0023 (0.0267)	-0.0762 (0.02907)	-0.7941 (0.0377)	1.3242 (0.0440)	0.0986(0.0151)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*	
‘repeat error smaller’	-0.2293 (0.0308)	-0.8660 (0.0360)	0.0254 (0.0285)	-0.0747 (0.0302)	-0.8123 (0.0397)	1.3646 (0.0444)	0.0582 (0.0142)
	-0.2317	*-0.8806*	*0.0157*	*-0.0688*	*-0.8366*	*1.4151*	
‘alpha larger’	-0.3785 (0.0449)	-1.4606 (0.0479)	0.0309 (0.0405)	-0.1431 (0.0435)	-1.3857 (0.0499)	2.3218 (0.0528)	0.0857 (0.0055)
	-0.3973	*-1.5095*	*0.0269*	*-0.1180*	*-1.4341*	*2.4259*	
‘alpha smaller’	-0.0957 (0.0343)	-0.4247 (0.0588)	0.0548 (0.0281)	-0.0366 (0.0321)	-0.3710 (0.0400)	0.6659 (0.0644)	0.4319 (0.0644)
	-0.0662	*-0.2516*	*0.0045*	*-0.0197*	*-0.2390*	*0.4043*	
‘phi larger’	-0.2366 (0.0322)	-0.8825 (0.0359)	0.0248 (0.0287)	-0.0849 (0.0307)	-0.8185 (0.0388)	1.4103 (0.0466)	0.0792 (0.0122)
	-0.2354	*-0.8798*	*0.0135*	*-0.0686*	*-0.8385*	*1.4122*	
‘phi smaller’	-0.2636 (0.0302)	-0.8773 (0.0385)	0.0007 (0.0249)	-0.1193 (0.0285)	-0.8640 (0.0412)	1.4496 (0.0495)	0.1059 (0.0152)
	-0.2277	*-0.8814*	*0.0181*	*-0.0690*	*-0.8344*	*1.4183*	

5.2.2 Convergence

The objective of the estimation methods is to converge to the true posterior distribution of the parameters. If a parameter has not converged, this indicates a strong uncertainty in the estimate and a possibly bad performance of the sampler. I evaluate the convergence of the mean parameters over all individuals. To test for convergence, I first look at the plot of the estimates. I then implement a test for statistical convergence. In terms of statistical convergence I used Geweke's metric (Plummer et al., 2006). This leads to a Z-test, where I take the two-tailed statistic for a 95% certainty. If the Z-statistic is larger than 1.96 or smaller than -1.96, I have statistical reason to reject the hypothesis of convergence. The statistics for which I haven reason to reject the hypothesis are shown as **bold**. In Table 4, 5, 6 the convergence statistics are shown for the general model, extended model and the γ of the extended model. In these tables Geweke's metrics are visible for every simulation. The statistic is implemented over the sampled mean of the last 10.000 iterations. One non-convergence could occur in a set of six parameters. However when the amount of rejected convergences becomes more than 2, there is reasonable doubt to question the performance of the model.

For the general model, the test statistics can be seen in Table 4. No simulations in this model raise suspicions about the convergence of the estimates.

Table 4. The Geweke diagnostic, testing for the convergence of the mean of the part-worths in the general model, for all simulations

simulation	β_1	β_2	β_3	β_4	β_5	β_6
'no error'	0.6673	0.5228	-0.3019	1.3360	0.9128	0.2403
'base model'	**-3.3200**	-1.4660	0.6214	-0.0924	1.8860	-1.0660
'random error larger'	0.5497	0.5140	-1.1940	0.0274	-0.5948	-0.3971
'random error smaller'	-0.0115	0.4299	0.9775	-0.4036	-0.0771	1.4850
'repeat error larger'	**-2.0890**	-0.1458	-0.8528	-1.5360	-1.5880	-1.8120
'repeat error smaller'	-0.0259	1.9180	-0.8773	0.1161	1.5500	-0.2391
'alpha larger'	0.5085	1.0820	-0.4768	1.5010	-0.1825	**-2.0360**
'alpha smaller'	-0.5076	-0.3349	0.1248	0.3173	-1.6680	1.2930
'phi larger'	0.2031	1.0770	-1.1390	0.8439	0.5438	-1.6870
'phi smaller'	**-2.371**	-1.4820	-0.7173	1.1360	**2.0260**	-1.1290

For the extended model, both the estimates of the general parameters and the convergence of γ are of interest. The convergence of β can be seen in Table 5 and the convergence of γ in Table 6. As can be seen, there is statistical reason to reject hypothesis of convergence for several estimates. For the simulation setting 'repeat error smaller', 'alpha larger' and 'alpha smaller', I have reason to doubt the performance of the model. Smaller α lead to poor convergence, this can be explained by the higher randomness in this simulation. Since the convergence was present in the simulation without errors, this may indicate that the extended model

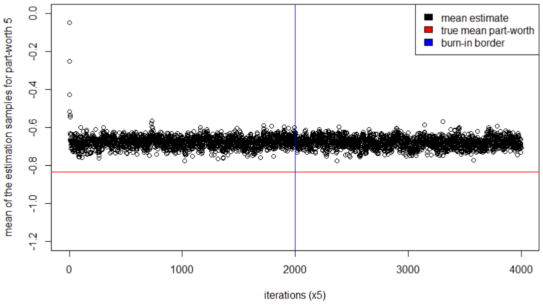


Figure 2. The samples of the general model for 20.000 iterations in the 'base case' simulation for β_5

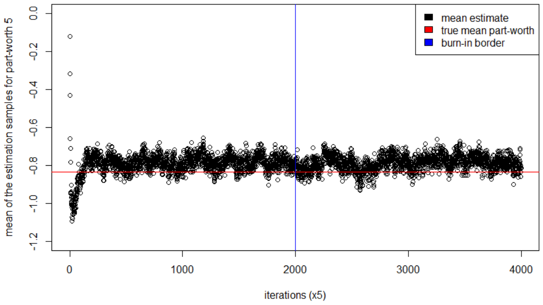


Figure 3. The samples of the extended model for 20.000 iterations in the 'base case' simulation for β_5

has a problem with estimating in this setting. To test the hypothesis of high randomness, I simulated 10 data sets based on the same part-worths as 'alpha smaller'. By comparing these new choice sets against the used choice set for 'alpha smaller', I found that for an average of 31% of the choices, the choices were the same. I compare this value against a setting where all part-worths are equal to 0, where the percentage of similar choices is 28%. The percentage of similar choices in the 'alpha smaller' setting is very close to random choices. So a small amount of information is still included, but this simulation does indeed contain a high amount of randomness. The percentage of repeats for the 'base case' is approximately 44%.

For the convergence problem in the other simulations, probable cause is the fact that the additional γ is included, influencing the β estimates as well. As opposed to the general setting, where the location of b is primarily based on the data and current estimates through β, now the γ location also is sampled using the data directly in the acceptance step. So if in one step, γ is updated and accepted, in the next step β_i has different probabilities of acceptance than for the previous γ. For different γ, different β_i will become the estimation. For different β_i, different γ will be the estimation. These different possible results may be the explanation of the lack of convergence.

The convergence problems are also visible in the comparison of the estimation as seen in Figures 2 and 3. Both figures show the samples for the conditional posterior mean of the part-worths. Specifically the figures show the samples for the 'base case' simulation for β_5. Both samplers converged, however there is a difference visible for the extended model as opposed to the general model. the extended model seems to follow more of a 'path', where the samples more away and then back to the mean. This path becomes more clear when disregarding the first 10.000 iterations as seen in Figures 4 and 5. These figures do show the extended model samples closer to the true value as opposed to the general model.

Figures 6 and 7 show the samples for β_5 in a simulation where the general model did converge and the extended model did not. Here the extended model is also further away from the true value and shows a somewhat descending path over the last 10.000 iterations. The general model does not show this path, making a strong case for a better performance in the general model.

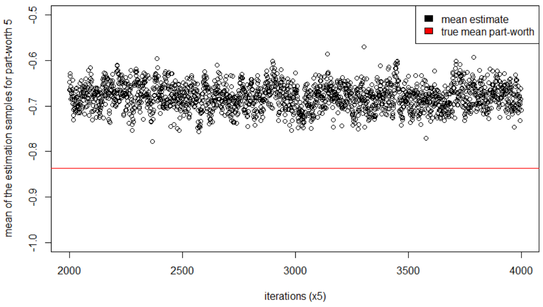


Figure 4. The samples of the general model for 10.000 iterations after burn-in in the 'base case' simulation for β_5

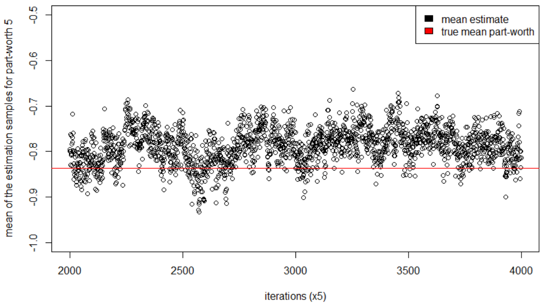


Figure 5. The samples of the extended model for 10.000 iterations after burn-in in the 'base case' simulation for β_5

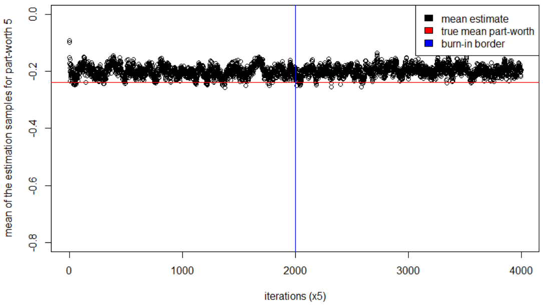


Figure 6. The samples of the general model for 20.000 iterations in the 'alpha smaller' simulation for β_5

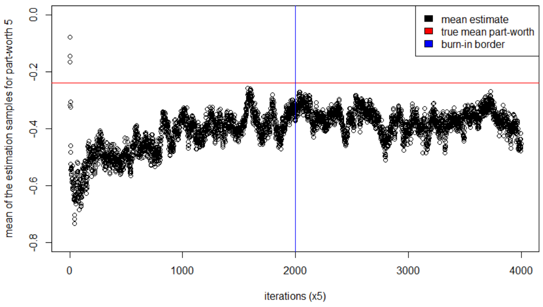


Figure 7. The samples of the extended model for 20.000 iterations in the 'alpha smaller' simulation for β_5

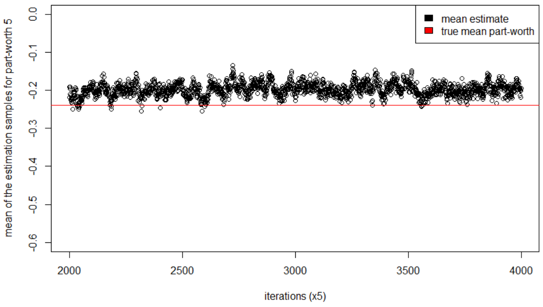


Figure 8. The samples of the general model for 10.000 iterations after burn-in in the 'alpha smaller' simulation for β_5

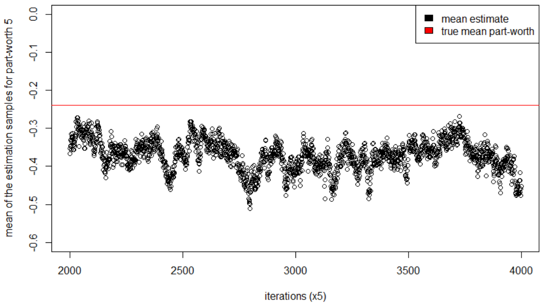


Figure 9. The samples of the extended model for 10.000 iterations after burn-in in the 'alpha smaller' simulation for β_5

Table 5. The Geweke diagnostic, testing for the convergence of the mean of the part-worths in the extended model, for all simulations

simulation	β_1	β_2	β_3	β_4	β_5	β_6
'no error'	-0.0926	-1.2030	-0.0966	-0.8690	0.2688	-0.5222
'base model'	-0.7395	-0.4697	-1.1840	-1.4110	-0.7724	0.5169
'random error larger'	0.0265	-1.879	0.4133	-0.6869	**-2.3970**	**2.9460**
'random error smaller'	**2.2500**	1.5740	-0.4741	-0.6494	1.3940	-1.8830
'repeat error larger'	-0.5460	0.1121	-0.8015	0.9861	0.5922	-0.3209
'repeat error smaller'	1.6280	-1.8990	1.4480	**-3.6860**	**-2.3660**	**2.0030**
'alpha larger'	1.8920	**2.2970**	-1.2370	1,8520	**3.5390**	**-2.4370**
'alpha smaller'	**3.4180**	**5.1450**	-1.5200	-0.7339	**2.4640**	**-2.8020**
'phi larger'	-0.3815	-0.5690	-0.5733	**-2.3100**	-0.8872	0.5886
'phi smaller'	-0.5156	-1.5070	**2.1930**	-1.8130	-1.7830	0.9819

As can be seen in Table 6, γ does not converge for several simulations. I first discuss the simulation without human errors. Where after I discuss the simulations with human errors. In the case of no human errors, γ attempts to catch a probability which equals 0. Because γ in this case may become very negative and therefore the probability q will become very small, as discussed in Section 5.2.1, a higher or lower γ will not lead to a real difference in the probability. It would be expected for γ to freely move around in these large values without being pulled towards a certain value. It can then be expected that there is no convergence in this case. In this specific setting, the prior on γ restricts these extreme values, causing γ to not completely not converge. For the other simulations, γ is not expected to take a non-zero value where freely moving around. The fact that γ does not converge for several simulations raises suspicions of general convergence problems for the extended model.

The problems in convergence for γ are also visible in Figures 10 and 11. Where the 'base case' in fact did converge and the 'alpha smaller' sampler did not converge. The plot in Figure 11 does not show a clear overall trend for this amount of iterations, but it does show the plot is not varying around a certain value yet. This leads to no convergence. Since the plot does not seem to be ascending or descending, I have reason to doubt whether the estimate would converge after more iterations. Both the 'base case' and 'alpha smaller' estimations show a high variability, this shows uncertainty about the estimate for the γ. The plot in Figure 11 raises serious suspicions regarding the extended models performance in terms of convergence.

Because both γ and the part-worths show a lack of convergence for several simulations, I doubt the performance of the extended model.

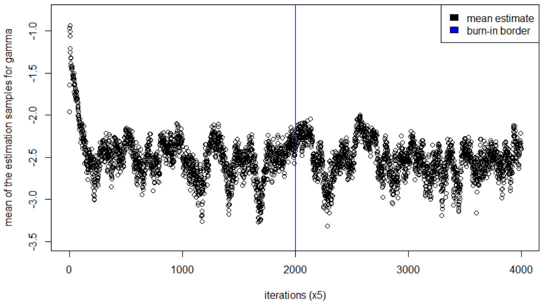


Figure 10. The samples of the general model for 20.000 iterations in the 'base case' simulation for γ

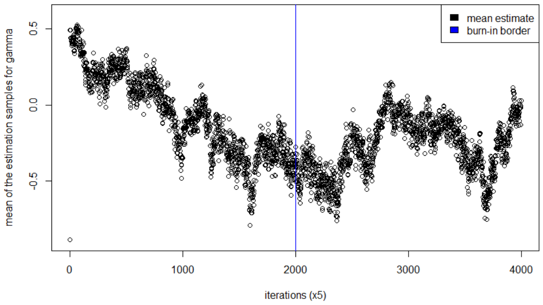


Figure 11. The samples of the extended model for 20.000 iterations in the 'alpha smaller' simulation for γ

Table 6. The Geweke diagnostic, testing for the convergence of γ for all simulations

simulation	γ
'no error'	**-3.0810**
'base model'	0.3309
'random error larger'	1.9610
'random error smaller'	**-2.3840**
'repeat error larger'	-0.2964
'repeat error smaller'	1.8630
'alpha larger'	**-2.7020**
'alpha smaller'	**-3.1930**
'phi larger'	-0.0367
'phi smaller'	1.2370

5.2.3 Evaluation metrics

The evaluation metrics include the hitrate, MSE and likelihood. I first discuss these for the general model, then I discuss them for the extended model. Finally I compare the findings. All findings are based on the results of the last 10.000 iterations of the models. The hitrate and likelihood are out of sample metrics, MSE is in sample. For the hitrate, no errors are included in the out of sample estimation. The likelihood does have errors included in the out of sample data. For both models, the results are shown in Table 7. For every simulation and evaluation metric, the metric of the model which performs better is made **bold**.

Firstly I discuss the general model. As can be seen, the *hitrate* is in general higher for 'alpha larger' than for the case without errors. This can be explained by the fact that a large α makes the parameters easier to estimate. So the estimates may be better in general in this case. Smaller α also leads to the lowest hitrate which is as expected due to the higher randomness of the choices in this case. Smaller error probabilities lead to higher hitrates as opposed to larger error probabilities, which is as expected. The inclusion of human errors does seem to lead to a certain loss in performance in the hitrate. So when human errors are present (specifically with a larger probability of occurence), the general model is worse at predicting the choices of respondents (purely based on their part-worths) as opposed to the case of no errors. The *likelihood* of the data is highest when there are no errors present. This is as expected, since in the case of no errors, there is also less randomness present in the respondent choices. Furthermore, again the simulations with smaller error probability have higher likelihood as opposed to simulations with larger error probabilities. Bigger α also again has a higher likelihood than smaller α, however in this case the larger α does not outperform the simulation without errors. Finally the *MSE* again shows the same findings. An important observation in the MSE evaluation is the fact that the true β values are more varied or of a different order in the last 4 simulations as opposed to each other and the first six. So especially the α variations cannot be directly compared.

For the extended model, the hitrate, likelihood and MSE show the same findings as for the general model. The hitrate is about the same for large random error and large repeat error.

Finally I compare the general model with the extended model. The first notable observation is the fact that in terms of MSE, likelihood and hitrate, the no error simulation results do not differ greatly. So the extended model is also able to estimate the data well in case of no errors. This is supported by the fact that the γ estimate becomes very small.

The hitrate shows small differences in performance of the models. In the case of 'repeat error larger' they have in fact the same hitrate. No clear conclusions can be made about which model outperforms the other in terms of hitrate because of the small differences. In some cases the extended model performs slightly better, in others the general model. There does not seem to be a clear structure in which models performs better in which setting. I expect this to be partly random since the difference is often not more than 0.0010, which translates to around 25 more answers predicted correctly in a total of 25000 predictions. The only case where the performance is larger is the 'alpha smaller' simulation. This simulation shows a higher hitrate for

Table 7. The evaluation metrics for the general (gen.) model and extended (ext.) model for all simulations

simulation	$hitrate_{gen.}$	$hitrate_{ext.}$	$MSE_{gen.}$	$MSE_{ext.}$	$likelihood_{gen.}$	$likelihood_{ext.}$
‘no error’	**0.5317**	0.5316	**1.1729**	1.1733	-27214.02	**-27207.81**
‘base model’	**0.5248**	0.5244	1.4865	**1.3767**	**-27547.01**	-27550.30
‘random error larger’	0.5245	**0.5250**	1.7598	**1.4413**	-27690.29	**-27675.06**
‘random error smaller’	0.5269	**0.5276**	1.3733	**1.2987**	**-27350.69**	-27356.41
‘repeat error larger’	0.5250	0.5250	1.7294	**1.4863**	-27733.70	**-27711.55**
‘repeat error smaller’	**0.5297**	0.5294	1.3648	**1.2876**	**-27343.72**	-27361.50
‘alpha larger’	0.5388	**0.5397**	4.1894	**3.0051**	**-27510.30**	-27766.26
‘alpha smaller’	0.4819	**0.4868**	**0.2169**	0.3728	-31636.58	**-31605.08**
‘phi larger’	**0.5296**	0.5292	1.6521	**1.5043**	**-27340.02**	-27343.89
‘phi smaller’	0.5194	**0.5209**	1.3122	**1.1866**	-27723.17	**-27601.36**

the extended model. This slightly better performance may be explained by the fact that the ‘alpha smaller’ simulation has more randomness in the true part-worths. If the extended model is able to catch a part of this randomness, the estimates may predict the true choices better. It should be noted that the difference in performance for the ‘alpha smaller’ simulation is still small, so no certain conclusions can be based on this difference. Overall the hitrate shows a comparable performance for the two models, this does not lead to reason to doubt the performance of the extended model. It however also does not show an increased performance in the extended model.

When looking at the MSE, a bigger difference is visible in the performance, as opposed to the hitrate. Except for the simulation without errors and the smaller α, the estimates in the extended model are closer to the true part-worths in all simulations. The ’no error’ simulation having a better performance in the general model can be expected since this model does not have the influence of a γ here. The fact that the ‘alpha smaller’ simulation has a higher fit in the general model as compared to the extended model can be explained by the fact that a smaller α leads to higher randomness. In the extended model this randomness may have been caught in the γ. This is indeed seen in the fact that the estimate for the probability of an error in this simulation is over 40%. So because the γ caught the randomness of the choices of the respondents, the estimates of the part-worths have also been scaled. This causes the estimates to be further away from the true values. For data sets with high randomness or small part-worths for the respondents, the extended model may not perform as well as compared to the general model. For all other simulations with errors, the MSE of the extended model is lower. This indicates that the γ indeed is able to clean some of the human error effects from the estimates.

When looking at the likelihood, the difference of the simulations with larger probability of errors as compared to the lower probability becomes more apparent. The likelihood of the general model is higher for the base model, ‘random error smaller’, ‘repeat error smaller’, ‘alpha larger’ and ‘phi larger’. Again all values do not differ greatly. The likelihood shows that in the case of more errors, the extended model fits the data better. However when there are less errors present in the data. The fit of the general model is indeed better, even when fitting to a data set with errors present.

So the MSE and hitrate indicate a good ability to estimate the parameters in the extended model, but the likelihood indicates worse performance whenever there isn’t a large amount of errors present. Also, the randomness of the choices of respondents for small α has a negative effect on how close the estimated parameters are to the true parameters in the extended model.

To conclude, the extended model is capable of adjusting the estimates of both the part-worths and γ to account for human errors, as can be seen in the MSE. However the performance on the likelihood indicates the extended model performs better only when there is a large amount of errors present. Finally, the extended model has the risk of accounting for more randomness than just the randomness caused by human errors.

5.2.4 Acceptance probabilities

Since the ρ and δ which influence the acceptance probabilities are set over all simulations, it is relevant to look at the probabilities which are obtained. These probabilities are shown in Table 8. The first observation is the fact that the acceptance probability of γ for the ‘no error’ simulation is very large. Especially since I intended to have an acceptance probability of approximately 0.30 for γ. This can be explained by the fact that the estimate of γ in this case is also very far away from the initialisation value. For this γ, the probability of an error becomes very small and a slightly different value for γ does not influence this probability significantly. So γ may move around more freely. For the part-worths I intended to have an acceptance probability around 0.23. This is reached in most simulations and the obtained acceptance probabilities are within acceptable regions. The risk of having a differing acceptance probability is first that having a lower acceptance rate could lead to slower convergence. With 20.000 iterations I do not expect the problem of slow convergence to have occurred. However, a large acceptance probability can lead to more random movement in the estimates, which could also cause a lack of convergence. To analyse if this problem occurred, I first look at ‘alpha smaller’ with probabilities of 0.3708 and 0.3839 and ‘random errors smaller’ which has relatively high γ acceptance. For the general model, ‘alpha smaller’ does not show low convergence. For the extended model ‘alpha smaller’ does show a convergence performance different from most other simulations. I can not know for certain if this low convergence was caused by the high acceptance rate or the randomness in the respondent choices. It may have influenced the convergence and explain this low performance partly. However since the general model did not show this problem, this does not raise immediate suspicions of convergence problems due to this acceptance rate. When looking at ‘random errors smaller’, there is no indication of low convergence. I conclude therefore, the difference in acceptance probabilities can have a slight influence on the convergence in the models. This difference in the probabilities can however not be the primary cause and most likely has only a small influence, if any.

Table 8. The acceptance probabilities for the part-worths in both the general model and the extended model with γ, for all simulations

simulation	$\beta_{general}$	$\beta_{extended}$	γ
'no error'	0.2167	0.2165	0.9777
'base model'	0.2448	0.2429	0.3507
'random error larger'	0.2695	0.2638	0.2247
'random error smaller'	0.2378	0.2371	0.4213
'repeat error larger'	0.2675	0.2657	0.3171
'repeat error smaller'	0.2373	0.2367	0.3973
'alpha larger'	0.1804	0.1700	0.2183
'alpha smaller'	0.3708	0.3839	0.2333
'phi larger'	0.2322	0.2290	0.3189
'phi smaller'	0.2759	0.2735	0.2841

5.2.5 Run time

To compare the run time of the models, I first note that the run time is highly dependent on the application used. The results can be seen in Table 9. The run time between simulations can not be compared since this is highly dependent on the application used. The comparison between the models indicates the general model is a lot faster than the extended model. This is as expected when adding an additional step. It takes approximately 52.82% longer for the extended model in the case of 20.000 iterations. When planning to do many more iterations, this difference in run time could pose a problem.

Table 9. The run times of the extended and general model in hours for all simulations

simulation	extended model	general model
'no error'	2.67	1.72
'base model'	2.56	1.67
'random error larger'	2.58	1.68
'random error smaller'	2.56	1.66
'repeat error larger'	2.59	1.68
'repeat error smaller'	2.55	1.70
'alpha larger'	2.54	1.67
'alpha smaller'	2.56	1.69
'phi larger'	2.55	1.67
'phi smaller'	2.56	1.69

6 Conclusion

CBC analysis is an often used method in marketing research, however the research into human errors influencing the results is not extensive yet. This paper attempts to provide an additional direction in this research by accounting for the human errors when estimating part-worths. The main Research Question is *How can we recognize human errors in the decision results obtained from a Choice Based Conjoint Analysis task?*. To answer this question, the first subquestion is how to recognize these human errors in the data. With the next step and goal of adjusting the analysis to account for these errors. This all of course had the objective of improving the reliability of the estimated part-worths.

The two subquestions both relate to the research and implementation done before the results analysis. In Section 2 I introduce the currently used estimation methods for Choice Based Conjoint Analysis, namely Hierarchical Bayes. As stated, there is no extension on this method currently available with the exact objective of accounting for human errors. The ways in which human errors can occur in the data then have been discussed in Section 2.2. This includes not paying attention and selecting randomly or for instance accidentally repeating the previous answer. To then answer the question how we can adjust the analysis to recognize and account for these errors, I extend the Hierarchical Bayes algorithm with an additional Metropolis-Hastings step. This additional step attempts to estimate a parameter which catches the probability of an error occurring in the results. The question then is whether this additional step is indeed able to adjust the analysis to account for the errors. To evaluate and compare the performance of both methods, I implement them on 10 different simulations. These results are available in Section 5.

The results firstly show an influence of the human errors on the estimates of the part-worths, as well as the performance of the method. The likelihood and the hitrate of the model decrease, whilst the MSE increases when errors are introduced. This supports the objective of extending the currently used method to decrease the influence of these errors and sustain the performance.

To answer the main question, I discuss whether the extended model is indeed able to account for these human errors. I do this by comparing the general model, which does not specifically account for the errors, with the extended model. When I compare the results with the extended model, the MSE of the extended model is smaller in almost all simulations. The fact that the MSE is smaller indicates that the estimates are indeed closer to the true part-worths as opposed to the estimates of the general model. Furthermore, the hitrate shows no clear difference between the two models and the likelihood of the extended model performs better in several simulations. This indicates the extended model is indeed in some way able to account for the influence of the errors. However the extended model also brings additional problems, primarily the lack of convergence for several estimates decreases the reliability of the results. Furthermore, the general model outperforms the extended model in terms of hitrate and likelihood on several simulations. The extended model mainly seems to outperform the general model for high human error probabilities or more randomness, for smaller error probabilities the general model performs better. Finally a concern is the run time since the extended model takes an additional hour on 20.000 iterations, this run time could pose a serious problem on bigger data sets.

In conclusion, the extended model shows potential and is able to improve the choice predictions for respondents, however it is currently not completely reliable. Recognizing human errors could therefore be possible with this extended method but further research is needed to improve the performance. In Section 7 I discuss the implications, limitations and possible next steps to further the research and develop a model which is able to reliably account for human errors.

7 Discussion

This section first discusses the limitations which have occurred during the research. Then steps for further research are suggested, both to prevent these limitations as to expand the research. Finally the implications of the research are discussed.

In terms of limitations and future research I first discuss the limitations regarding the data set-up. A big factor is the fact that all data was simulated. This means the results can not with certainty be carried over to real life data. A next step would be to implement the used methods on real life data to envision the influence of the extension in this case. There currently is no certainty if the simulated errors are indeed how errors behave and occur in true data. For clean comparison purposes this research is usable but it does not reflect true data with certainty. Furthermore, the current implemented errors are not individual dependent. Whilst it could be expected to have more error prone individuals. This is a discrepancy with real life data and could be included in future simulation research. Also, the simulated repeat error has the same probability of occurring for 2 repeat answers as for 10 repeat answers. The probability of this error occurring can be expected to increase as the amount of repeat answers increases. In future research it would be advisable to implement this error with an increasing probability. Additionally, expanding on the comparisons made, the simulations only compare the influence of certain factors as opposed to the base case. To have a more clear view of the correspondence of certain factors, future research could include more simulations with for instance a random grid for the factor settings. Finally, the designs as used are random designs. Huber & Zwerina (1996) shows there are better design structures available when attempting to get as much information from the choices of respondents as possible. For future research, a different design set-up could be implemented.

Then, in terms of the definition of the model and the extended model, firstly the extended model is not able to account for the different types of human errors as introduced in Section 2.2. The model is only based on random errors. As seen in Section 5, the effect of random errors as opposed to the repeat errors is different on the performance of the estimation. So extending the model to make a difference between repeat and random errors could influence the estimations. Furthermore, the convergence problems in the extended estimation method decrease the reliability of the results. The exact cause of these problems should be found to increase the reliability and performance. As explained in Section 5, the convergence problems in the model may be (partly) caused by the fact that γ and β are sampled consecutively. Furtermore, the results show a large variability in the estimates of γ which could also influence the estimates of the part-worths in the extended model. This uncertainty about the estimates makes the results less reliable. Research in the direction of making the γ estimate more stable could improve the convergence of the estimates. A first step in this research could be the inclusion of additional parameters in γ and possibly making γ individual or question dependent. Making γ a function of for instance the effect of *how many questions have been answered already* and the effect of *how fast is the individual answering the question.* Another solution to the problems of convergence may be to sample γ and β at the same time. This could be done by for instance using a 10 random walk values for γ, sampling β for every γ and then deciding on the best next step. This however may also bring new down-sides. So for future research it would be beneficial to attempt different sampling and estimation structures than currently used.

In terms of implementation of the methods, the estimation was only implemented for 20.000 iterations due to time constraints. With more iterations, the methods might have converged better. However since the results seemed to not be moving towards a specific point but just move around too randomly to converge, I do not expect more iterations to make a big impact on the results. The implementations were also done from scratch, which is relatively error prone. In future research packages could be created and used to implement these methods. This could also improve the speed of the methods. Furthermore, I set the hyperparameters determining the acceptance ratios to the same value for all simulations. This in some cases led to a higher or lower than intended acceptance probability. A higher acceptance probability could also have caused lower convergence. So in future research, the hyperparameters should be set separately for every simulation to evaluate this effect and conclude more reliably on the convergence capabilities of the extended model. Finally, the Inverted Wishart sampling was done following Equation 11. However, for instance, Train (2009) states a slightly different variance for this sampling. In the implementation, a factor v may have been omitted from the equation. Further research into the correct implementation should be done. The variance of Σ may have been too small due to this difference, causing the samples of β to be more alike than intended. This may have for instance slowed down the convergence speed, however as this was also controlled by δ, this has not led to any direct problems.

To finally discuss the limitations of the results, firstly the performance evaluations are possibly not completely reliable. Both the estimation of the model and the out of sample performance are based on simulations. Meaning the MSE could not be evaluated on a out of sample data set. Also for the hitrate and likelihood, although new choices were simulated for the evaluation, the true respondent part-worths have been held constant. In real life data, there might be more variance or differences between the estimation and evaluation data. In future research, true data should be included and all evaluation set-ups should be expanded to include out of sample data.

This same risk of overfitting is increased by the fact that I attempt to catch a certain error, however the respondent choices also contain a certain randomness through the gumbell error. So it is possible that the extended model not only attempts to catch human errors, but also the randomness of respondent choices.

With the current simulations and results, I am not able to indicate how much of the effect which γ catches is the effect of human errors. γ can also catch different randomness. For future research, it would be interesting to implement the methods using different gumbell errors to see the effect of this error on the estimates. And again, it would be usefull to perform out of sample evaluation.

Furthermore, the results for the evaluation metrics are based on the average of the estimates over the last 10.000 iterations. Some of the parameters had not converged so these estimates may not be entirely reliable. Although I do not expect my conclusions to differ greatly due to this limitation, improving the convergence could again improve the reliability of the results obtained. Finally, although the MSE of the extended model outperforms the general model for almost all settings, the same does not hold for the likelihood and hitrate. This of course raises the question of which is most important. But eventhough in my research the MSE is most important, the difference in performance still leads to doubt about the performance of the extended model.

To conclude the limitations and future research, I discuss the implications my research has. As stated in this section, there are many limitations present. Combined with the fact that the extended model as proposed does not have completely reliable estimates, the model as proposed should not be used in real life settings as is. However, what this research does prove is the fact that the extended model is able to recognize and account for the human errors in some sense. By recognizing the potential of this extension and performing future research to increase the reliability and sophistication, a method could be developed which is able to provide more reliable results in the case of human errors than the currently used method of Hierarchical Bayes.

References

Adams, M. J. (2017). Failures to comprehend and levels of processing in reading. In *Theoretical issues in reading comprehension* (pp. 11–32). Routledge.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics*, *89*(1-2), 57–78.

Bhatnagar, A., & Ghose, S. (2004). A latent class segmentation analysis of e-shoppers. *Journal of business research*, *57*(7), 758–767.

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to bayesian statistics.* John Wiley & Sons.

Bradley, M., & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, *21*(2), 167–184.

Calderhead, B. (2014). A general construction for parallelizing metropolis- hastings algorithms. *Proceedings of the National Academy of Sciences*, *111*(49), 17408–17413.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1–58.

Chrzan, K., & Orme, B. (2000). An overview and comparison of design strategies for choice-based conjoint analysis. *Sawtooth software research paper series*, *98382*, 161–178.

DeSarbo, W. S., Ramaswamy, V., & Cohen, S. H. (1995). Market segmentation with choice-based conjoint analysis. *Marketing Letters*, *6*, 137–147.

Edmondson, A. C. (2004). Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *The journal of applied behavioral science*, *40*(1), 66–90.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis.* CRC press.

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, *31*(3_supplement), S56–S73.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgemental data. *Journal of Marketing Research*, *8*(3), 355–363.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research*, *5*(2), 103–123.

Greenberg, E. (2012). *Introduction to bayesian econometrics.* Cambridge University Press.

Hess, S., Hensher, D. A., & Daly, A. (2012). Not bored yet–revisiting respondent fatigue in stated choice experiments. *Transportation research part A: policy and practice*, *46*(3), 626–644.

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association*, *60*(311), 806–825.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Huber, J., & Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing research*, *33*(3), 307–317.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, *39*(1), 103–129.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the neo-pi-r. *Journal of Personality Assessment*, *76*(2), 315–332.

Louviere, J. J. (1988). Conjoint analysis modelling of stated preferences: a review of theory, methods, recent developments and external validity. *Journal of transport economics and policy*, 93–119.

Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *Journal of marketing research*, *20*(4), 350–367.

Magidson, J., & Vermunt, J. K. (2007). Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference. In *Sawtooth software conference* (Vol. 139).

Moore, W. L. (2004). A cross-validity comparison of rating-based and choice-based conjoint analysis models. *International Journal of Research in Marketing*, *21*(3), 299–312.

Natter, M., & Feurstein, M. (2002). Real world performance of choice-based conjoint models. *European Journal of Operational Research*, *137*(2), 448–458.

Norman, D. A. (1981). Categorization of action slips. *Psychological review*, *88*(1), 1.

Otter, T., Tüchler, R., & Frühwirth-Schnatter, S. (2004). Capturing consumer heterogeneity in metric conjoint analysis using bayesian mixture models. *International Journal of Research in Marketing*, *21*(3), 285–297.

Paap, R. (2021a, September). *lecture3.* Erasmus Universiteit Rotterdam.

Paap, R. (2021b, October). *lecture7.* Erasmus Universiteit Rotterdam.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from `https://journal.r-project.org/archive/`

Pokropek, A., Żółtak, T., & Muszyński, M. (2022). Mouse chase: Detecting careless and unmotivated responders using cursor movements in web-based surveys.

Regier, D. A., Ryan, M., Phimister, E., & Marra, C. A. (2009). Bayesian and classical estimation of mixed logit: an application to genetic testing. *Journal of health economics*, *28*(3), 598–610.

Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, *22*(3), 304–328.

Roy, V. (2020). Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, *7*, 387–412.

RStudio Team. (2021). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA. Retrieved from http://www.rstudio.com/

Rumelhart, D. E. (2017). Schemata: The building blocks of cognition. In *Theoretical issues in reading comprehension* (pp. 33–58). Routledge.

Savage, S. J., & Waldman, D. M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics*, *23*(3), 351–371.

Sawtooth Software, I. (2021, March). *The cbc/hb system, technical paper v5.6.* Retrieved from https://sawtoothsoftware.com/resources/technical-papers/cbc-hb-technical-paper

Spiro, R. J., Bruce, B. C., & Brewer, W. F. (2017). *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education* (Vol. 11). Routledge.

Statisticat, & LLC. (2021). Laplacesdemon: Complete environment for bayesian inference [Computer software manual]. Bayesian-Inference.com. Retrieved from https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software (R package version 16.1.6)

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

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge university press.

Vriens, M., Wedel, M., & Wilms, T. (1996). Metric conjoint segmentation methods: A monte carlo comparison. *Journal of Marketing Research*, *33*(1), 73–85.