ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis [Business Analytics and Quantitative Marketing]

Modelling Slipping and Carelessness in a Discrete Choice Model using a Four-Parameter Logistic Function

Student:

Jasper van der Vos

481019

Supervisor:

D. Fok

Second assessor:

K. Gruber

February 6, 2024

Abstract

This research introduces the implementation of a four-parameter logistic (4PL) function in a discrete choice model, in this case a multinomial logit, aiming to account for carelessness and slipping in the context of investigating the true preferences of individuals through choice tasks. A Gibbs sampler with Metropolis-Hastings step is used to estimate the parameters. The simulation experiment results indicate that the 4PL model does not perform worse in standard multinomial logit (2PL) scenarios, while it outperforms the 2PL model in datasets constructed with a 4PL scenario. The 4PL model is compared against the 2PL model on four real-life datasets, where the focus relied on model fit. For the datasets where the 2PL model was superior, the differences were relatively small, whereas the performance differences were larger when the 4PL model outperformed the 2PL model. Besides this, it is noted that in most datasets almost no careless responding or slipping was detected, which may raise the question whether the new model is useful in real-life datasets. This does not deter us from concluding that the 4PL model shows enhanced performance on datasets that are based on a 4PL scenario, where it does no harm for estimating parameters in a standard multinomial logit model.

Contents

1 Introduction

Consumers make choices between substitution goods and services all day long. For example, they need to choose which brand of pasta to buy for dinner or which hair dresser to go to. In these decision-making processes, consumers make a trade-off between the available alternatives and strive to choose the good or service that brings them the highest utility. Nevertheless, it is important to recognize that the consumer not always manages to choose the option that will bring them the highest utility. It is possible that they make a mistake by accident or that they are not fully informed about all the attributes of the product. Other factors that could play a role are that consumers behave careless or do not keep their attention when making the choice. Consequently, individuals may not consistently make choices that truly reflect their underlying preferences. Whether this is the case for a certain choice or not can not be observed. However, it is known that these mistakes occur. Therefore, one has to take the mistakes into account when analyzing the results of choice tasks with the goal to reveal the true preferences of the respondents. Otherwise, the presence of the errors could influence the results and conclusions from the decision-making model.

In previous studies, researchers tried to understand the behavior of careless respondents and identify such individuals to exclude them from their analysis or assign them lower weights in the model to decrease their impact on the results. These approaches seem to be good ways to deal with this issue. However, Ward and Pond (2015) stated that deleting careless respondents still influences the results of a research, because the sample size is reduced in a non-random way. An alternative way of dealing with mistakes is to incorporate them in the model. This minimizes their influence on the parameter estimation and does not lead to a loss of data.

In this paper, the idea is introduced to include the imperfect choices of consumers in a discrete choice model based on item response theory (IRT) models. The general field of IRT is psychometrics, but the models are used in multiple other fields. An example of the application of an IRT model is that it can be used in psychology to measure latent constructs through questionnaires. Next to this, it can also be used in personality assessments (Reise and Waller, 2009). Furthermore, IRT models are used to analyse items, i.e. questions, in a test and produce ability scores for students. In this example, a logistic function is used for all questions in a test to plot the abilities of students against the probability that they will answer an item correct. The logistic function is used to display the relationship between those two variables.

The IRT model has been researched extensively in the past and multiple variations of IRT models

have been created. The one-parameter logistic (1PL) IRT model (Rasch, 1993) is the simplest model and only includes a parameter for the difficulty of an item, whereas the two-parameter logistic (2PL) model (Birnbaum, 1968) also includes a factor for a different slope, which corresponds to the level of discrimination of an item. Next to this, complex models have been constructed in the recent years to model the answers of students more accurately. For example, guessing of correct answers plays a large role in the analysis of test items, particularly in the context of multiple choice questions where it is always possible for a student to guess the correct answer. To incorporate the influence of guessing, the three-parameter logistic (3PL) model was created (Birnbaum, 1968). The 3PL model does not fix the lower asymptote at 0; it can also be above 0. The lower asymptote represents the probability that the student with the lowest possible ability answers a question correct. By allowing the lower asymptote to be above 0, it is ensured that the student with the lowest ability has a non-zero chance of guessing the correct answer. On the other hand, it also occurs that the student with the highest ability answers a question incorrect, for example caused by carelessness or slipping. This is in contrast with this student's ability level and the corresponding chance of answering the question incorrectly. To cope with these situations, the four-parameter logistic (4PL) model was created, which includes the terms of slipping and carelessness (Barton and Lord, 1981). In the 4PL model, the upper asymptote is not fixed at 1. The upper asymptote represents the probability that the student with the highest possible ability answers a question correctly. By not fixing this parameter at 1, it is allowed for a student with the highest possible ability to make a mistake and have a probability less than a hundred percent to answer a question correctly.

The concept that individuals can slip or respond carelessly during a test in the 4PL IRT model can be integrated into the situation where consumers make a choice between products, because it is possible that a consumer chooses a product that is not of its true preference. Therefore, the objective of this paper is to incorporate the influences of mistakes into a discrete choice model for consumer choices by using the logistic function of the 4PL IRT model. In order to achieve this, certain modifications are required in the discrete choice model. In education, responses to items can be correct or incorrect, which is not the case when consumers are faced with multiple products. Nevertheless, the idea of introducing asymptotes that are not fixed at 0 and 1 can be implemented, thereby acknowledging the concept that consumer choices are never guaranteed to match their true preferences.

The implementation of non-fixed asymptotes is needed, because when trying to estimate parameters when the data of a choice task includes mistakes can lead to incorrect conclusions about

real-life preferences of people. An example of the influence of using a model without accounting for mistakes on the parameter estimates is given by Loken and Rulison (2010). They compared the results of a 2PL and 4PL IRT model, where it was assumed that the 4PL model was the correct model to use. They showed that the estimated slopes reduced in the 2PL model and also the estimated difficulty of the items increased. The underestimation of the slope parameter and overestimation of the difficulty of items was also shown by Fu et al. (2021). These examples show that using the incorrect model can indeed influence the results. Therefore, it would be good to investigate the use of non-fixed asymptotes in the context of discrete choice models. To test whether this extension of the model improves the performance of the model, the following research question is formulated:

Does implementing the four-parameter logistic function of IRT into the discrete choice model provide an opportunity to account for mistakes in these models?

The idea of implementing the logistic function of the IRT model into a discrete choice model seems to be an interesting and promising idea. However, it also increases the number of parameters, and it is expected that this will increase the runtime of the parameter estimation. Therefore, it is also aimed to set off the improvement of the model fit against the runtime of the model. Another disadvantage of the increasing number of parameters is that there will be more parameter uncertainty. To address the concerns with the 4PL model, the following question is posed:

To what extent does the potentially improved performance of the four-parameter logistic discrete choice model outweigh the increased complexity of the model?

By answering this question, this paper contributes to the current literature by exploring a new way to incorporate human mistakes into a discrete choice model. This will result in a better estimation of the true preferences of people.

In this research, the used discrete choice model is a multinomial logit model. The 4PL multinomial model is created to test whether this model can capture mistakes in discrete choice models. The initial testing of the model uses simulation data. Generated datasets are used to compare the estimation of the 4PL multinomial logit model to the estimation of a standard multinomial logit model (referred to in this study as a 2PL multinomial logit model). First, two datasets are simulated using 2PL and 4PL scenarios and serve as a basis for comparing the performance of the estimation models. Subsequently, multiple simulated datasets are employed to explore the model's performance across various 4PL datasets with different features, such as a large number of alter-

natives. Thereafter, a comparative analysis is conducted between the two estimation models using real-life datasets to evaluate the applicability of the model beyond simulated scenarios.

To estimate the parameter values for both models (2PL and 4PL), a Bayesian approach is used. Specifically, a Gibbs sampler with Metropolis-Hastings step is chosen. The Gibbs sampler facilitates the handling of high-dimensional data and is employed to find the posterior distribution of the parameters together with the Metropolis-Hastings step, including an acceptance-rejection algorithm.

The marginal likelihood and Deviance Information Criterion (DIC) serve as two measures for assessing the model fit during the evaluation of the estimation methods. Additionally, the mean percentage marginal effect (MPME) and mean absolute difference marginal effect (MADME) are computed to compare the relative and absolute sizes of the marginal effects of both models. Moreover, the runtime and parameter uncertainty are used to weigh the increasing complexity of the model against its performance.

The results indicate that there is no substantial difference in performance for the simulated 2PL dataset. However, various simulated 4PL datasets demonstrate superior performance for the 4PL model. In terms of marginal effects, the 2PL model underestimates the effects when trying to estimate them for a 4PL dataset. In other words, the actual marginal effects are significantly larger than the 2PL model estimated. These findings suggest that the 4PL model outperforms the 2PL model for 4PL datasets. Nonetheless, it is questionable whether the conclusions about decision behavior drawn from these results differ from those based on the 2PL model.

For the real-life datasets, no model consistently outperforms the other. This can possibly be caused by different origins of the data, a 2PL or 4PL scenario. It is observed that when the 2PL model slightly outperforms the 4PL model, the estimates of both models are nearly identical, whereas the reverse is not true. This suggests using the 4PL model is a safe choice. Despite the drawback of a longer runtime for the 4PL model, the difference in runtime is small enough be to outweighed by any inferior results of the 2PL model. Notably, in most dataset almost no slipping or carelessness is detected by the models. This raises the question whether the use of the model is useful in real-life.

All in all, incorporating non-fixed asymptotes into a discrete choice model, leading to the development of the 4PL multinomial logit model, demonstrates a high potential in the context of simulated datasets. However, this potential is not directly fulfilled when the model is applied to real-life datasets. On the other side, opting for a 4PL multinomial logit model does no harm to

the results, whereas choosing not to incorporate the non-fixed asymptotes, when preferred, could significantly impact the outcomes.

The remainder of the paper is structured as follows. In Section 2, discrete choice models, item response theory and human errors in discrete choice models are introduced. Section 3 demonstrates the used methods in this paper. Section 4 describes the simulated and real-life datasets that are used in this research. Section 5 demonstrates the results. In Section 6, a conclusion is drawn upon the results. Finally, Section 7 discusses limitations and further research.

2 Literature

In this section, previous work that is relevant to the current study will be discussed. First, item response theory is discussed. Subsequently, discrete choice models are introduced. Finally, the concepts of slipping and carelessness in the context of discrete choice models are discussed.

2.1 Item Response Theory

Item response theory (IRT) is most frequently used in education to analyse tests and measure abilities of students. The theory assumes that not all items in a test are equally difficult (Embretson and Reise, 2013). This is in contrast with another frequently used theory in education called classical test theory (Lord and Novick, 2008). IRT can be used to estimate a latent trait of students and can have multiple forms. In the context of education, this latent trait is for example a student's spelling or mathematics ability. IRT makes use of Item Characteristic Curves (ICC) for each individual item. The ICC sets the probability that a student will answer a question correct off against the ability of the student. Various forms of IRT models have been created in the past with their own level of complexity.

The simplest IRT model is the Rasch model (Rasch, 1993), which includes one parameter that reflects the difficulty of an item. As this model has one varying parameter in the logistic function, it is also called the one-parameter logistic (1PL) model. The probability p_j of choosing the correct answer is denoted by:

$$p_j = \frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}}\,, \tag{1}$$

where b_j denotes the difficulty of item j and θ denotes the ability of the individual. When the value of b_j becomes larger, the logistic function will shift to the right. This leads to the result that students with ability θ have a smaller chance to answer a question correct. To account for

situations that could occur in real-life, the 1PL model is extended by including a second parameter which allows for varying discriminating properties of the items. Incorporating the second parameter results in the the two-parameter logistic (2PL) IRT model (Birnbaum, 1968), which ensures that the slope of the ICC can differ per item. The corresponding formula is

$$p_j = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}\,, \tag{2}$$

where a_j denotes the parameter that influences the slope of the logistic function. A higher value for a_j results in a larger discrimination of the question, meaning that the item differentiates individuals abilities better.

In the 1PL and 2PL model, the probability of answering a question correct falls within the range between 0 and 1. However, when students take a test with multiple-choice questions, there is always a chance that they guess the correct answer. Therefore, the probability of answering a question correct will always be above 0. To address this, Birnbaum (1968) proposed a model that could deal with guessing by creating a lower asymptote of the logistic function that was above 0. The formula for the three-parameter logistic (3PL) IRT model that shows the probability p_j that a student with ability θ answers item j correct is

$$p_j = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}\,, \tag{3}$$

where c_j denotes the parameter for the lower asymptote. The value of c_j corresponds to the probability that a student with the lowest possible ability answers the question correct. In other words, the probability that students guess a certain question correct.

After the 3PL was introduced, Barton and Lord (1981) proposed the idea that even the student with the highest possible ability would never reach a probability of 1 for getting the correct answer due to the influence of slipping and carelessness. To take this into account in the ICC, they introduced a fourth parameter that allowed the upper asymptote to vary between 0 and 1. This resulted in a four-parameter logistic (4PL) IRT model with the formula

$$p_j = c_j + (d_j - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}\,, \tag{4}$$

where d_j denotes the parameter for the upper asymptote of item j. $1 - d_j$ corresponds to the proportion of students who slip or behave careless for item j. The other variables have the same interpretation as for the 3PL model.

Including the parameters for guessing and carelessness, the 4PL model is a model that can deal with guessing in a test and imperfections of the human kind when making a test. Having the upper

asymptote lower than 1 corresponds to the fact that students might slip during a test or behave careless. In other words, even the best students will not always answer a question correctly. This extension of the IRT models does not only create a model that is more accurate, but it also improves measurement efficiency and estimation of the ability of high-achieving students (Yen et al., 2012).

Shortly after the 4PL model was created, there was not much attention for these models with two main reasons (Loken and Rulison, 2010). The first reason was that there was only a small benefit of this model over other models. Secondly, the fitting of the model seemed very difficult with maximum-likelihood methods. However, as time went on, more researchers began to recognize that an upper asymptote not equal to 1 could get a more accurate representation of the data (Osgood et al., 2002; Reise and Waller, 2003). On top of this, new methods appeared to estimate the parameters in an efficient way in the 4PL model. Multiple extensions of the Expectation-maximization algorithm were created to estimate the parameters in the 4PL model (Monte Carlo EM, Quasi Monte Carlo EM, Metropolis-Hastings Robbins-Monro) (Kalkan, 2022). Next to that, Bayesian methods were used to estimate the parameters (Markov Chain Monte Carlo with Gibbs Sampling or Metropolis-Hastings). The new ways of estimation and new interest in the possibilities of non-fixed asymptotes renewed the interest in the 4PL IRT model.

2.2 Discrete Choice Models

Discrete choice models describe the choices that are made by consumers between multiple discrete alternatives. In these models, consumers choose a specific product rather than a continuous quantity, which is the case for continuous choice models. An example of a discrete choice scenario could involve an individual deciding which transport to take to work: bike, bus or car. Currently, several kinds of models are available to model these decision-making processes. These include models created for choices between two alternatives (binomial models) or multiple alternatives (multinomial models), and distinctions are also made between probit and logit models. The basis of discrete choice models lies in the utility theory, which states that consumers will choose the product or service that provides them the highest utility corresponding to the pleasure and happiness that a consumer experiences when buying a product.

In this research, a multinomial logit model will be used as basis. A multinomial logit model tries to explain a choice task with a logistic formula which represents the probability that a consumer will buy a certain product. The number of alternatives to choose from in the choice task is larger than two. The assumption is made that there is no correlation among the error terms of the alternatives

in the task. Not only individual-specific variables are incorporated in the model, but also product-specific variables. The formula for the probability that person i chooses alternative j out of the set of alternatives K is equal to

$$p_{ij} = \frac{e^{\beta_j x_i + \gamma_j z_{ij}}}{\sum_{k=1}^{K} e^{\beta_k x_i + \gamma_k z_{ik}}}, \tag{5}$$

where β_j corresponds to the specific parameters of alternative j, x_i to the intercept and personal characteristics of consumer i (e.g. income or gender), z_{ij} to the product-specific characteristics, and γ_j to the product-specific parameters (e.g. price or brand name). To ensure identification in the estimation procedure, the β values of the first alternative are set to zero.

2.3 Slipping and Carelessness in Discrete Choice Models

In research questionnaires and other experiments, careless behavior is a problem (Maniaci and Rogge, 2014). Careless behavior occurs when individuals respond to questions in a different way than what would reflect their real preferences (Ward and Pond, 2015). This behavior can influence the results of analyses because the data contains impurities. To deal with carelessness in questionnaires, multiple methods have been proposed. For example, cleaning the data before using it in an analysis. This was done by identifying participants that showed careless behavior in their answers and remove them from the dataset (Meade and Craig, 2012; Ward and Meade, 2023). Other studies tried to weigh the data of respondents with careless behavior differently than the data of other participants (Hasselhorn et al., 2023; Ulitzsch et al., 2023). The problem is that careless responses of participants could still cause psychometric problems even after removing these respondents correctly or using weights to decrease their influence (Ward and Pond, 2015). The reason for this is that the sample size is reduced in a non-random way, which could shape the distribution of the sample size. Besides that, it is also possible that an individual responds carelessly at a small number of tasks. Therefore, it would be a loss of data when all the information of this individual would be removed.

Careless behavior is a problem when trying to retrieve information about people's true preferences from experiments. However, this is only one part of the problem. Careless behavior is conscious behavior of respondents. Next to this, there is also a possibility that a participant does not answer based on their true preferences when they did not intend to do this. The term that is used for this phenomenon is slipping. Slipping includes all mistakes that are not made on purpose. For example, action slips could cause the fact that consumers do not choose a product of their real preference. Another example of slipping that could occur during a task where consumers need to choose between different products is rational inattention. Rational inattention defines the concept

of a participant not choosing something of their real preference, because they are not able to make a trade-off between all possible attributes of the products or the costs of the trade-off are too high (Matějka and McKay, 2015). In other words, there is an overload of information available and therefore they overlook information that could be interesting for their choice between the alternatives. Rational inattention refers to the fact that consumers make a choice on imperfect information instead of all the information that is available, simply because they do not have the capacity to process all the information (Sims, 2003, 2010).

In this research, rational inattention will not be taken into account directly, as the data will not consist of this many variables that rational inattention could play a role. On the other hand, it is possible that the careless behavior and slips could be present in the choice data, influencing the parameter estimates undesirably. The 4PL IRT model can account for guessing, slipping and carelessness. This study extends the application of these terms to a discrete choice model, particularly in situations where a consumer must choose between multiple products. Unlike the educational context, involving correct or incorrect answers, guessing is irrelevant in this context as there is no correct or incorrect in choosing between products. Consequently, the extra parameters in the 4PL IRT model are in this context not about answering a question correctly by guessing or unintentionally answering incorrectly. However, individuals may accidentally choose a product that does not align with their true preferences, which can be seen as a human mistake. When modelling people's choices, these accidental mistakes can be identified as errors. This can be done by allowing the lower asymptote to be above 0 (which means that the chance that a certain alternative is chosen will always be above zero percent), and adding the possibility that the upper asymptote can be below 1 (allowing the maximum probability to choose an alternative to be below one hundred percent).

Without adding terms to the model that could deal with the influences of carelessness and slipping, there is no possibility that these mistakes are taken into account. Therefore, this study contributes to the current literature by investigating a new way to cope with these mistakes in discrete choice models. This is done by incorporating terms in a discrete choice model that could deal with the errors caused by careless responding and slips.

3 Method

This section starts with an introduction of the four-parameter multinomial logit model that is created. Subsequently, the estimation procedure, incorporating the Gibbs sampler with Metropolis-Hastings step, is demonstrated. Finally, the measurements for model fit and parameter estimates to analyse the results of the estimations are discussed.

3.1 Four-Parameter Multinomial Logit

In this research, the four-parameter logistic model of the item response theory is generalized to a context where consumers make choices between products. One problem implementing this is that a question in a test can be answered correctly or incorrectly. Consequently, there are only two possible outcomes. This corresponds with a binomial model. However, the model in this research reflects the choice of an individual between more than two alternatives and therefore a multinomial model is used.

The idea of asymptotes in the 4PL function is implemented in the context of a multinomial logit model, where the probability that alternative j is chosen by individual i is equal to

$$p_{ij} = c_j + (d_j - c_j)\frac{e^{\beta_j x_i + \gamma_j z_{ij}}}{\sum_{k=1}^{K} e^{\beta_k x_i + \gamma_k z_{ik}}} \,. \tag{6}$$

In this model, the IRT component in the exponent is replaced by the multinomial logit component with the individual-specific variables x_i and the product-specific variables z_{ij} with their corresponding parameters β_j and γ_j. The β's and γ's are allowed to change per alternative. c_j and d_j denote the lower and upper asymptotes for alternative j respectively, which implies that $c_j \leq p_{ij} \leq d_j$. c_j corresponds to the chance that individual i accidentally chooses alternative j. $1 - d_j$ denotes the probability that individual i slips or behaves carelessly, and not chooses alternative j. The part that is left, $d_j - c_j$, is explained by the individual and product-specific variables.

In the context of the binomial IRT models, the lower asymptote of the correct answer and upper asymptote of the incorrect answer automatically add up to 1. This can be explained by the fact that a binomial model has complementary probabilities. In the context of multiple choices, the model needs to be changed from two alternatives to three or more. To ensure that the upper asymptote of one alternative and the other lower asymptotes of all other alternatives add up to 1, a new constraint is defined. The formula corresponding to this constraint is

$$d_j + \sum_{k \neq j}^{K} c_k = 1, \forall j \in K \,, \tag{7}$$

where d_j is the upper asymptote of alternative j and together with the sum of all the lower asymptotes of the other alternatives, it needs to add up to 1. This constraint is also sufficient to ensure that the probabilities of different alternatives will always add up to 1. The proof for this statement is shown in Appendix A.

The proof in Appendix A also shows that the term $(d_j - c_j)$ has to be constant for all alternatives j. In this study, the value of $(d_j - c_j)$ is equalized to κ. Some rewriting of Equation 7 (see Appendix A) leads to the following formulation of κ:

$$\kappa = 1 - \sum_{j=1}^{K} c_j \, . \tag{8}$$

Incorporating κ into the 4PL multinomial logit model leads to the following formula:

$$p_{ij} = c_j + \kappa \, \frac{e^{\beta_j x_i + \gamma_j z_{ij}}}{\sum_{k=1}^{K} e^{\beta_k x_i + \gamma_k z_{ik}}} \, . \tag{9}$$

All in all, the multinomial logit model is extended by adding parameters for the lower asymptotes for all alternatives ($j \in K$) and the parameter κ, which represents the proportion of the probabilities that is explained by product and individual-specific variables.

The log-likelihood for N individuals and K alternatives, which can be used to measure how well a model explains the observed data, is expressed as:

$$\log(\, p(y|\beta, \gamma, c, \kappa) \,) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \cdot \log(p_{ij}) \, , \tag{10}$$

where y_{ij} is equal to 1 if individual i chooses alternative j, and 0 otherwise.

With the parameter estimates of a multinomial logit model, marginal effects can be calculated. Marginal effects show the change in probability with a one-unit change, which improves the interpretability of a model. For the 4PL model, the same formula can be used as for a standard multinomial logit model. The corresponding formula for the marginal effect obtained from Heij et al. (2004) for an individual-specific variable x_i on alternative j is

$$\frac{\mathrm{d}p_{ij}}{\mathrm{d}x_i} = p_{ij} \left(\beta_j - \sum_{k \neq j}^{K} p_{ik} \beta_k \right) \tag{11}$$

To compute the marginal effects for a model, the average for all individuals is taken. To calculate the marginal effects for a product-specific variable, the same formula is used where β is replaced by γ.

3.2 Bayesian Approach

To estimate the parameters of the model, a Bayesian approach is used. The ideas in this section on Bayesian statistics are primarily based on information in Greenberg (2012) and Train (2009). The Bayesian estimation of parameters uses prior knowledge in estimating the parameters. The basis lies in Bayes' theorem to update probability distributions, which is equal to

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)}\,. \tag{12}$$

The conditional posterior distribution of θ given y, is equal to the conditional distribution of y on θ times the prior distribution of θ divided by the distribution of y. As $p(y)$ is often difficult to calculate and does not influence the distribution because it is a constant value with respect to θ, the following formula is often used

$$p(\theta|y) \propto p(y|\theta)\, p(\theta)\,. \tag{13}$$

The earlier discussed prior information is represented in the term $p(\theta)$ and $p(y|\theta)$ can be easily computed. Using these two terms, the density of the parameters (θ) given certain data (y) can be calculated. To sample new values for the parameters from this posterior distribution, an algorithm is needed.

3.2.1 Gibbs Sampler with Metropolis-Hastings Step

A Markov Chain Monte Carlo (MCMC) method is used in this research to simulate from the posterior distribution. In the MCMC method Gibbs sampling (Gelfand and Smith, 1990; Geman and Geman, 1984) is used. This is a recursive method to construct a Markov chain that has the posterior distribution as its limiting distribution. Gibbs sampling splits up the dimensionalities when sampling a new parameter value. Most of the times this method is used when the high dimensionality of a model makes it difficult to draw a next point. The Gibbs sampler iteratively draws new values for all the different parameters from the distribution conditional on the other parameters. Within the Gibbs sampler, a Metropolis-Hastings algorithm is used, because of its high flexibility which makes it of great use for difficult distributions. The Metropolis-Hastings algorithm is incorporated in a Gibbs sampler to update the samples.

In this study, not all variables are updated at the same time, but they are divided into multiple blocks as in a Gibbs sampler for efficiency. When drawing parameters for a certain block, the Metropolis-Hastings step is executed. The first blocks consist of the β's and the γ's. These variables

are separated per variable, meaning that the blocks will consist of the values for all alternatives for a specific β or γ. The last block includes the parameters of c_j and κ.

The Metropolis-Hastings algorithm is used in this research because there is no readily available distribution conditional on the data and on other parameters. Therefore, this algorithm is used to choose whether to jump to the next point or not. It calculates whether to move on to the new point based on the likelihoods of the two points. In this study, a random walk is used as proposal function, which implies that there is no term for the proposal function present in the acceptance-rejection formula. According to Greenberg (2012), the formula corresponding to the situation with a random-walk proposal function is

$$\alpha(\theta^o, \theta^n) = \begin{cases} \min\left\{\frac{p(y|\theta^n)}{p(y|\theta^o)}, 1\right\}, & p(y|\theta^o) \neq 0, \\ 0\,, & \text{otherwise.} \end{cases} \tag{14}$$

Here, θ^o represents the current set of the parameter values and θ^n the proposed/new set. In this research, the likelihood from Equation 10 is incorporated into this expression as $p(y|\theta)$. Equation 14 shows that if the density is higher for θ^n compared to θ^o, the move to that point will certainly be made. When it is the other way around, there is still a chance that the move will be made, but with a lower probability. In the formula the possibility of $p(y|\theta^o) = 0$ is included. However, it is very unlikely that this situation occurs, because it means that the current value does not exist.

The full Gibbs sampler with Metropolis-Hastings step in the MCMC simulation will be described here. First, a new sample of values is drawn for the variables in the current block. Then $\alpha(\theta^o, \theta^n)$ in Equation 14 is calculated using this new point. After this, U is drawn from $U(0, 1)$. When $U \leq \alpha$ the new point θ^n is accepted. Otherwise it continues with the old values and proceeds to the next block. When the loop over all blocks is done, one iteration has passed. The first half of iterations is used as burn-in period. Also in this period, the standard deviations of the random walk are tuned. After every 1000 iterations the acceptance rates are calculated to set the standard deviations of the proposal distributions. The goal is to keep the acceptance rates between 0.2 and 0.4. Therefore, the standard deviations are adjusted if the acceptance rate falls outside this range. This is done by multiplying the standard deviation by 1.5 when the acceptance rate is too high, or by 0.67 when the rate is below 0.2.

3.2.2 Priors

For the Bayesian estimation of the parameters in Equation 9, it is necessary to specify priors, which needs to be done by intuition, logic, or earlier research (Train, 2009). To introduce the priors, it is essential first to introduce the formulation of κ and the c_j's in this research. One approach could involve a Dirichlet distribution with values for both κ and the c_j's together, as this distribution ensures that values of set of variables add up to 1. However, to keep the interpretation simple and the parameters independent from each other, in this research a different approach is chosen. Specifically, the value of κ is separated from the c_j's, meaning that also the priors are separated. To ensure that the value of κ is within the range of 0 and 1, a beta distribution is used as prior for κ:

$$\kappa \sim \text{Beta}(7, 2) \, . \tag{15}$$

The parameters of the beta distribution are chosen such that the distribution is left-skewed. This choice is driven by the idea that the characteristics should account for the largest part of the choice probabilities.

With the formulation in this study, the value of κ defines the part of the probabilities that is composed of the characteristics of the alternative and the individual. The complementary part, $(1 - \kappa)$, represents the sum of all lower asymptotes of all the alternatives (as in Equation 8), which can be divided between the lower asymptotes. The parameters associated with the asymptotes (c_j's) typically use a beta distribution as a prior (Mislevy, 1986). However, since this research incorporates the concept of asymptotes in a multinomial context, the Dirichlet distribution, also known as a multinomial beta distribution, is used. The exact distribution of the probabilities over the various alternatives is defined by

$$c_j \sim (1 - \kappa) \, \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_J) \, , \tag{16}$$

where α represents the parameter values, set as positive real numbers with a value of 1.5 in this study. Using a Dirichlet distribution ensures that the sum of values, drawn from this distribution, always equals 1. By multiplying these values with the term $(1 - \kappa)$, the c_j's are computed and the constraint in Equation 7 is satisfied.

For both κ and the c_j's, the random walk is used as proposal function. However, allowing this without restrictions could lead to values below 0 and above 1. Consequently, the algorithm for κ is restricted between 0 and 1, resulting in a bounded random walk. If the drawn value exceeds 1, it is set to 1, and if it is below 0, it is set to 0. To make sure that the constraint in Equation 16

is satisfied with the new c_j's, the new values of the c_j's and κ need to add up to 1. Achieving this involves drawing new values through a random walk and scaling them. The random walk for the c_j's is bounded at 0, ensuring that only positive numbers are drawn.

Lastly, priors for the individual-specific and product-specific parameters are specified. For the β's and γ's in the model, a standard normal distribution is chosen as the prior distribution and the proposal distribution is also a random walk.

3.2.3 Implementation Check

To evaluate the implementation of the Bayesian algorithm, Simulation-Based Calibration (SBC) is used (Talts et al., 2018). This method can identify inconsistencies in the implementation of the model by constructing histograms of rank statistics. Using the rank statistics a uniform distribution appears when the implementation of the model is correct. The idea behind SBC is based on the self-consistency of a joint posterior distribution. Talts et al. (2018) stated that 'for any model the average of any exact posterior expectation with respect to data generated from the Bayesian joint distribution reduces to the corresponding prior expectation'. This corresponds with the following formula

$$p(\theta) = \int p(\theta|\tilde{y})\, p(\tilde{y}|\tilde{\theta})\, p(\tilde{\theta})\, \mathrm{d}\tilde{y}\, \mathrm{d}\tilde{\theta}\,, \tag{17}$$

where $p(\tilde{\theta})$ corresponds to the density from the prior and $p(\tilde{y}|\tilde{\theta})$ to a posterior density that is drawn from $p(y|\tilde{\theta})$. The rank statistic counts the number of times that the value of the random variable, in this research the likelihood, from posterior densities is smaller than the value of the same random variable evaluated at the prior sample and is calculated by

$$r(\{f(\theta_1), \ldots, f(\theta_L)\}, f(\tilde{\theta})) = \sum_{l=1}^{L} \mathbb{I}[f(\theta_l) < f(\tilde{\theta})] \in [0, L]\,. \tag{18}$$

For any one-dimensional variable, $f : \Theta \to \mathbb{R}$, this rank statistic will be uniformly distributed. When this is not shown in the histogram, it means that something went wrong with the implementation or derivation. Since a multidimensional model is investigated in this study, the procedure with the one-dimensional rank statistic is executed multiple times, i.e. for every variable.

Furthermore, before interpreting the parameter estimates, the convergence of the parameter estimates is assessed. Convergence deems to occur when the estimate values do not alternate anymore. The evaluation of convergence is done with the Potential Scale Reduction Factor (PSRF), a measure derived from the Gelman-Rubin method (Gelman and Rubin, 1992). Brooks and Gelman (1998) generalized the PSRF in the context of monitoring convergence in iterative simulations and

relies on multiple chains with overdispersed starting points. The PSRF value is computed per parameter to determine whether convergence has been achieved, indicated by a value below 1.1.

3.3 Measurement

To evaluate which model represents the data better, the 4PL multinomial logit model is compared to the standard (2PL) multinomial logit model. This comparison involves evaluating the marginal likelihoods of both models, calculated using the Chib-Jeliazkov method (Chib and Jeliazkov, 2001). This is an importance sampling method which computes the marginal likelihood by sampling from the posterior distribution. The marginal likelihood can be used as the 2PL model can also be interpreted as a 4PL model with the values of c restricted to 0 and κ to 1. Additionally, the Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002) is used to compare the model fit. This criterion is based on the AIC and is calculated using the posterior distribution.

Next to the parameter estimates, for the individual and product-specific variables the marginal effects are computed with the values of the β and γ estimates. With this information for each model, the mean percentage marginal effect (MPME) and the mean absolute difference of the marginal effect (MADME) between the 2PL and 4PL model are computed. The mean percentage of the marginal effect represents the size of the parameter estimates of the 2PL model relative to the 4PL model. This metric is calculated based on the expectation that the size of the marginal effects for the 2PL model will be smaller than for the 4PL model (Loken and Rulison, 2010; Fu et al., 2021). Additionally, the MADME between the two estimation models is computed to examine the absolute differences in marginal effects.

For a more in-depth comparison of the models, for some datasets the parameter estimates (including the 95% highest density intervals, representing the smallest interval containing 95% of the points) and the marginal effects are evaluated. Additionally, fitted probability curves are made to give a graphical representation of the difference between the estimation models.

3.4 Estimation Procedure

For each dataset, five different starting points are used, and the Markov chain consists of 20.000 iterations per model, with the first 10.000 iterations serving as the burn-in period. During the burn-in period, the stepsize of the Metropolis-Hastings random walk is tuned to ensure an acceptable acceptance rate of approximately 30%. The five starting points for the different Monte-Carlo Markov Chains are overdispersed, facilitating the computation of the PSRF.

After the burn-in period, all data points from the five chains are pooled, resulting in a total of 50.000 posterior sample points. These samples are used to calculate the marginal likelihood and DIC, enabling a comparison of the two estimation models. Additionally, the posterior mean and the posterior highest density interval are derived from these samples.

4 Data

In this section, the datasets that are used in this research are discussed. First, simulated datasets are discussed. These are used to investigate whether the estimation models showed better performance on datasets that used the same model for data generation. On top of that, datasets incorporating interesting features (e.g. a large number of alternatives) that are used for a more elaborate exploration of the performance of the estimation models are discussed. Lastly, datasets from real life are discussed, which are used to test whether the new model also performs well for non-simulated datasets. Combining the results of the different datasets will provide an insight into whether and how well the 4PL model works, and whether the mistakes can be discovered in real-life datasets.

4.1 Simulated Datasets

All simulated datasets that are used for this research were generated in the context of a hypothetical discrete choice experiment involving more than two alternatives. Each dataset consists of 1000 individuals with their unique 'characteristics', drawn from a standard normal distribution. The parameter values for the β's and γ's are drawn from the same distribution. For the β's, the value for the first alternative is set at 0, which is done to ensure identification of the parameters during the estimation. In cases where a dummy variable is incorporated in a simulated model, its values are drawn from a binomial distribution with a 0.4 success rate.

To start with, two basis simulation datasets are created to compare the performance of the 2PL and 4PL estimation models. Both datasets include 3 alternatives, 1 intercept, 2 individual-specific variables (x_i), and 2 product-specific variables (z_{ij}). The first dataset is constructed using a 2PL scenario (with $\kappa = 1$ and $c_j = 0$, $\forall j \in \{1, 2, 3\}$), while the second dataset was created with a 4PL scenario (with $\kappa = 0.7$ and $c_j = 0.1$, $\forall j \in \{1, 2, 3\}$).

Datasets with special cases of features are simulated, in addition to these basis datasets. All datasets are used to test the performance of the 2PL and 4PL models in a more elaborate way. Dataset A is characterized by the same number of variables as the basis models, but includes 8

alternatives. To avoid a small number for κ of 0.2, the values of c_j are set to 0.05, which corresponds with a κ value of 0.6. This adjustment ensures a more meaningful and realistic dataset. Dataset B represents another special case with no individual-specific variables, meaning that (besides the intercept) only product-specific variables are present in this dataset. The model has 3 alternatives, so the values for κ and c are the same as in the basis model, and 2 product-specific variables are included in the model. Dataset C introduces dummy variables to test whether the model also can deal with categorical data. The dummy variables replace the continuous individual-specific variables of the basic model in this dataset. Additionally, the model retains 3 alternatives, 1 intercept, and 2 product-specific variables.

4.2 Real-Life Datasets

Real-life datasets are also considered. These datasets are used to explore whether the 4PL estimation method performs better than the 2PL estimation method beyond simulated datasets.

4.2.1 Fishing

The first dataset, referred to as 'Fishing' in this research, was obtained from the Ecdat package in R (Croissant and Graves, 2006). This package contains various open datasets in the economic field that can be used for research purposes. The Fishing-dataset was originally collected for the research of Herriges and Kling (1999), and was from a discrete choice experiment. It includes both types of variables, individual and product-specific. Within the Fishing-dataset, 1182 individuals in the United States were asked to choose their preferred recreational mode of fishing. The available alternatives were beach, pier, boat, and charter. These alternatives where chosen 134, 178, 418, and 452 times respectively, and each option had its own price and catch rate that could vary for each individual and alternative. Additionally, individual income served as an individual-specific variable. Prior to the parameter estimation, both individual and product-specific variables were normalized, ensuring that the same priors and proposal functions could be used as for the simulated models.

4.2.2 Mode of Travel

The second dataset, also sourced from the Ecdat package (Croissant and Graves, 2006), was originally retrieved from Train (2022). This dataset focuses on the travel mode of 453 individuals. The available modes are by car, carpool, bus, or trail, which were chosen 218, 32, 81, and 122 times respectively. Notably, this dataset only consists of product-specific variables (cost and time).

Furthermore, similar to the Fishing-dataset, the variables were normalized before estimating the parameters, ensuring that the same priors and proposal functions could be used as for the simulated models.

4.2.3 Heating

Another real-life dataset, sourced from the Ecdat package (Croissant and Graves, 2006) and originally from Train (2022), is about choices for heating systems for houses within the United States. This dataset considers various individual-specific variables, including income, age of the head of the house, and the number of rooms. Next to this, product-specific variables that are included are the installation costs of a heating system and the annual operational costs. The alternatives were between gas central (573 times chosen), gas room (129 times), electric central (64 times), electric room (84 times), and heat pump (50 times). Before conducting parameter estimation, individual and product-specific variables were normalized. This step was taken to ensure that the same priors and proposal functions could be used as for the simulated models.

4.2.4 Travel Mode Students

Finally, a dataset incorporating dummy variables is used. This dataset is sourced online and previously collected for the research of Müller et al. (2008). The dataset focuses on the travel mode choices of German students to their school and substantially larger than the other datasets used in this research. Within this dataset, 8556 individuals made a choice between walk (1858), bike (1484), car (539) or public transport (4675). The dataset includes a mix of continuous variables, such as distance and energy spent when cycling, and dummy variables, including season (winter or summer), car available, gender and whether the school is located on the other side of the river or not. No product-specific variables are included. Prior to the parameter estimation, the non-dummy individual variables were normalized, ensuring that the same priors and proposal functions could be used as for the simulated models.

5 Results

In this section, the results of the parameter estimation with the different models are discussed. First, the implementation and convergence are checked. Subsequently, the outcomes of the estimation for the simulated datasets are presented, followed by the results obtained from the real-life datasets.

All analyses were executed using R (RStudio Team, 2020). Since no existing package contained the employed method and models, the entire program was developed from scratch.

5.1 Implementation and Convergence

The correctness of the implementation was verified using the SBC algorithm. The algorithm demonstrated that the histograms for all variables (β, γ, c, and κ) showed an approximately uniform distribution, indicating that the implementation was correct. The check with the SBC algorithm was conducted using values of $L = 100$ and $N = 1000$, and the plots are available in Appendix B. Next to this, convergence is needed before the estimates of the parameters will be discussed. The assessment of convergence involved calculating the PSRF values for all models and datasets. It is noteworthy that only 1 out of all 18 models (2PL model for the Travel Mode Students dataset) displayed a maximum PSRF value above 1.1. This indicates that, for this specific model, one or more parameter estimates have not yet converged, likely due to an insufficient number of iterations. The Travel Mode Students dataset was clearly the largest dataset with more than 8000 individuals and 7 individual-specific variables, which could possibly need more iterations to converge. The mean value of the PSRF for this model was equal to 1.049, which is far below 1.1, leading to the conclusion that most of the parameters of this model have converged. Moreover, the violation of the parameter estimates that have a PSRF value above 1.1 is potentially due to the large number of parameters that are tested and small enough to cause no problems for interpreting the results. Descriptive statistics of all the PSRF values for the models are provided in Appendix C. The PSRF values provide proof that the estimates have stabilized and reached a consistent value for each model.

5.2 Simulated Data

The simulated datasets can be split in two groups. First, the results of the basis models are extensively discussed. Subsequently, the outcomes of more specific datasets are demonstrated.

5.2.1 Basis Data

In Table 1, the performance measures of the estimation models are presented for the basis datasets, with 3 alternatives, 1 intercept, 2 individual-specific variables and 2 product-specific variables. As expected, for the 2PL dataset, the performance of the 2PL estimation model is slightly better than the performance of the 4PL estimation model when considering the marginal likelihood (2PL: -570.3

Table 1: Model Comparison for the Basis 2PL and 4PL Dataset

Dataset	**Model**	**Marg.lik.**	**DIC**	**Runtime**	**MPME**	**MADME**
2PL	2PL	-570.3	1135.9	1.9	0.96	0.004
	4PL	-573.6	1140.5	2.2		
4PL	2PL	-869.3	1745.5	1.9	0.53	0.058
	4PL	-843.8	1692.6	2.3		

Note: The runtime is given in minutes.

vs. 4PL: -573.6) and DIC (2PL: 1135.9 vs. 4PL: 1140.5). Additionally, the runtime for the 4PL model is 1.2 times the runtime for the 2PL model. The MPME has a value of 0.96, indicating that, on average, the marginal effects of the 2PL model are 96% of the size of the 4PL model. This implies no substantial difference in parameter estimates, which is supported by the MADME with a value of 0.004. Conversely, when evaluating the performance on the 4PL data, it is shown that the 4PL estimation model outperforms the 2PL model based on the marginal likelihood (2PL: -869.3 vs. 4PL: -843.8) and DIC (2PL: 1745.5 vs. 4PL: 1692.6). The performance difference is larger than for the 2PL data, suggesting that the 4PL model is more suitable in this scenario. This comes at the cost of a larger runtime, with ratio being almost the same as for the 2PL data. Furthermore, the MPME of 0.53 indicates that the parameters values are smaller for the 2PL than for the 4PL, and the MADME of 0.058 shows that there is a larger difference in absolute parameter estimates than for the 2PL dataset. Based on these results, it is concluded that the 4PL model shows almost the same results for the 2PL dataset as the 4PL model, but the 4PL model does outperform the 2PL model in the context of the 4PL dataset.

In this part, a more in-depth analysis of the parameter estimates of the models for the basis datasets will be conducted. Table 2 displays the parameter estimates and the corresponding 95% highest density intervals for both the 2PL and 4PL models applied to the basis 2PL dataset. The real values are presented in *italics*. Overall, the parameter estimates for both models do not significantly deviate from the real values. However, it is noteworthy that the absolute values of the 4PL model are slightly larger than those of the 2PL model for all estimates. Particularly interesting are the estimates for the c values and κ for the 4PL dataset. The estimated c values are very close or equal to 0, and the value of 0.00 is also within the highest density interval, indicating that the estimated value is close to the real value. The same holds for κ, where the posterior mean is 0.98, and the real

Table 2: Parameter Estimates of the Basis 2PL Dataset for the 2PL and 4PL Estimation Model

	Variables													
	intercept		β_1		β_2		γ_1		γ_2		c		κ (SE)	
2PL model														
Alt.1	0.00		0.00		0.00		0.64	[0.45,0.82]	-0.91	[-1.11,-0.72]	0.00		1.00	
	0.00		*0.00*		*0.00*		*0.57*		*-0.92*		*0.00*		*1.00*	
Alt.2	-1.33	[-1.63,-1.04]	-0.75	[-1.00,-0.50]	-0.25	[-0.50,0.00]	1.29	[1.05,1.56]	0.88	[0.65,1.12]	0.00			
	-1.39		*-0.71*		*-0.32*		*1.31*		*0.99*		*0.00*			
Alt.3	0.65	[0.44,0.86]	0.21	[0.02,0.41]	1.09	[0.85,1.32]	1.70	[1.46,1.93]	-1.54	[-1.77,-1.30]	0.00			
	0.69		*0.25*		*1.01*		*1.65*		*-1.44*		*0.00*			
4PL model														
Alt.1	0.00		0.00		0.00		0.66	[0.47,0.85]	-0.94	[-1.14,-0.73]	0.00	[0.00,0.02]	0.98	[0.96,1.00]
	0.00		*0.00*		*0.00*		*0.57*		*-0.92*		*0.00*		*1.00*	
Alt.2	-1.39	[-1.75,-1.07]	-0.78	[-1.04,-0.51]	-0.26	[-0.54,0.01]	1.35	[1.07,1.62]	0.92	[0.67,1.18]	0.00	[0.00,0.01]		
	-1.39		*-0.71*		*-0.32*		*1.31*		*0.99*		*0.00*			
Alt.3	0.66	[0.42,0.90]	0.22	[0.02,0.43]	1.14	[0.89,1.41]	1.77	[1.51,2.03]	-1.60	[-1.85,-1.34]	0.01	[0.00,0.03]		
	0.69		*0.25*		*1.01*		*1.65*		*-1.44*		*0.00*			

Note: The posterior means are given with the 95% highest density intervals between brackets and true values in *italics*

Table 3: Marginal Effects of 2PL and 4PL Model for Basis 2PL Dataset

Model		**Variables**			
		β_1	β_2	γ_1	γ_2
2PL	Alt.1	0.019	-0.074	-0.116	-0.036
		0.013	*-0.068*	*-0.123*	*-0.046*
	Alt.2	-0.072	-0.061	0.017	0.175
		-0.069	*-0.064*	*0.022*	*0.179*
	Alt.3	0.053	0.135	0.099	-0.139
		0.056	*0.132*	*0.102*	*-0.133*
4PL	Alt.1	0.019	-0.078	-0.120	-0.038
		0.013	*-0.068*	*-0.123*	*-0.046*
	Alt.2	-0.075	-0.065	0.018	0.183
		-0.069	*-0.064*	*0.022*	*0.179*
	Alt.3	0.055	0.145	0.106	-0.148
		0.056	*0.132*	*0.102*	*-0.133*

value of 1 is within the highest density interval. Investigating the highest density intervals of the parameters between the two models demonstrates that the widths of the intervals for the 2PL model are the same as for the 4PL model, indicating that the parameter uncertainty for both models is approximately the same for this dataset. When comparing the marginal effects of both models on the 2PL dataset in Table 3, the conclusion aligns with the previous analysis. Both models produce nearly the same output, indicating that the marginal effects of both models are approximately the same. Additionally, the 4PL marginal effects are slightly stronger than the 2PL marginal effects, consistent with the value of the MPME of 0.96 in Table 1.

Table 4 presents the parameter estimates and 95% highest density intervals for the 4PL datasets obtained using both estimation methods. The real values are given in *italics*. Notably, in the 2PL model, the values for c and κ are fixed at the wrong value. While the sign of the estimated parameters seems correct, their magnitudes are significantly lower than the real values. In some instances, the real value falls outside the range of the highest density interval. In contrast, for the 4PL model, where the values of c and κ are not fixed, the estimated values are close to the real values, and the real values are also within the highest density interval. For the other parameters, the estimates are

Table 4: Parameter Estimates of the Basis 4PL Dataset for the 2PL and 4PL Estimation Model

	Variables											
	intercept		β_1		β_2		γ_1		γ_2		c		κ (SE)	
2PL model														
Alt.1	0.00		0.00		0.00		0.28	[0.13,0.42]	-0.45	[-0.60,-0.30]	0.00		1.00	
	0.00		*0.00*		*0.00*		*0.57*		*-0.92*		*0.10*		*0.70*	
Alt.2	-0.51	[-0.70,-0.30]	-0.45	[-0.64,-0.26]	-0.09	[-0.28,0.10]	0.59	[0.41,0.76]	0.42	[0.25,0.59]	0.00			
	-1.39		*-0.71*		*-0.32*		*1.31*		*0.99*		*0.10*			
Alt.3	0.31	[0.14,0.48]	-0.05	[-0.21,0.12]	0.41	[0.24,0.58]	0.87	[0.71,1.03]	-0.73	[-0.89,-0.57]	0.00			
	0.69		*0.25*		*1.01*		*1.65*		*-1.44*		*0.10*			
4PL model														
Alt.1	0.00		0.00		0.00		0.49	[0.25,0.74]	-0.69	[-0.97,-0.44]	0.07	[0.03,0.11]	0.78	[0.71,0.86]
	0.00		*0.00*		*0.00*		*0.57*		*-0.92*		*0.10*		*0.70*	
Alt.2	-1.33	[-1.89,-0.84]	-0.97	[-1.36,-0.60]	-0.37	[-0.72,-0.02]	1.38	[0.98,1.81]	0.85	[0.53,1.20]	0.09	[0.06,0.12]		
	-1.39		*-0.71*		*-0.32*		*1.31*		*0.99*		*0.10*			
Alt.3	0.62	[0.26,1.02]	-0.03	[-0.29,0.21]	0.72	[0.41,1.04]	1.59	[1.19,2.01]	-1.38	[-1.78,-0.98]	0.06	[0.02,0.11]		
	0.69		*0.25*		*1.01*		*1.65*		*-1.44*		*0.10*			

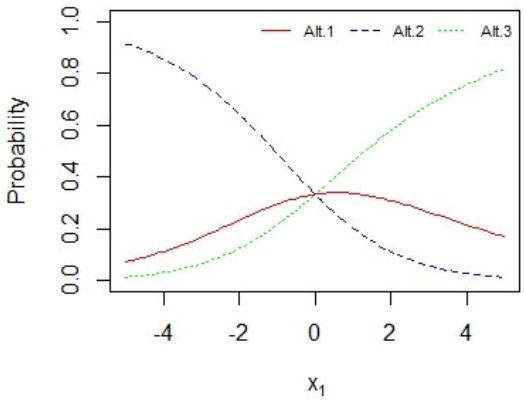
Note: The posterior means are given with the 95% highest density intervals between brackets and true values in *italics*

Table 5: Marginal Effects of the 2PL and 4PL Models for the Basis 4PL Dataset

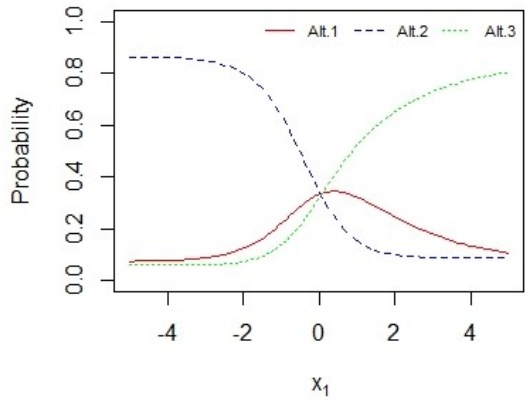
Model		**Variables**			
		β_1	β_2	γ_1	γ_2
2PL	Alt.1	0.039	-0.038	-0.088	-0.033
		0.020	*-0.090*	*-0.169*	*-0.069*
	Alt.2	-0.062	-0.043	0.003	0.146
		-0.121	*-0.123*	*0.023*	*0.313*
	Alt.3	0.023	0.081	0.085	-0.113
		0.101	*0.213*	*0.146*	*-0.244*
4PL	Alt.1	0.068	-0.048	-0.171	-0.034
		0.020	*-0.090*	*-0.169*	*-0.069*
	Alt.2	-0.133	-0.104	0.046	0.264
		-0.121	*-0.123*	*0.023*	*0.313*
	Alt.3	0.064	0.152	0.128	-0.232
		0.101	*0.213*	*0.146*	*-0.244*

closer to the real values than those obtained with the 2PL model, which can also be observed in Table 5. This table reveals that the marginal effects of the 4PL model are also stronger than those of the 2PL model, and the 4PL values are closer to the real values. Comparing the width of the highest density intervals of the models in Table 4 shows that the ranges of the 4PL intervals are larger than for the 2PL model. This indicates that the parameter uncertainty is larger for the 4PL model for this dataset. The MPME in Table 1 indicates that the sizes of the marginal effects of the 2PL model are only 53% of the size of the 4PL model marginal effects. This implies that the effect of changing one unit in the case of the 4PL is stronger in the case of the 4PL model compared to the 2PL model.

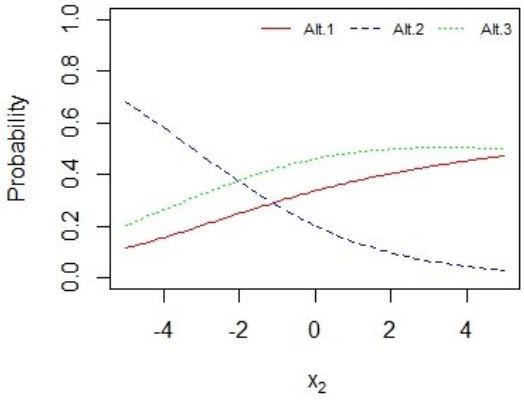
To visually compare the 2PL and 4PL models, fitted probability curves are created. The plots for the individual-specific variables are given in Figure 1. While these plots are used as examples, the probability curves of the other variables show the same pattern, and are included in Appendix D. Figure 1a and 1b show the plots for the first individual-specific variable. The overall shape of the probability curves is consistent across the models. However, it can be observed that the curves for the 4PL model reach their asymptote faster. Moreover, the slopes of the curves are steeper. The



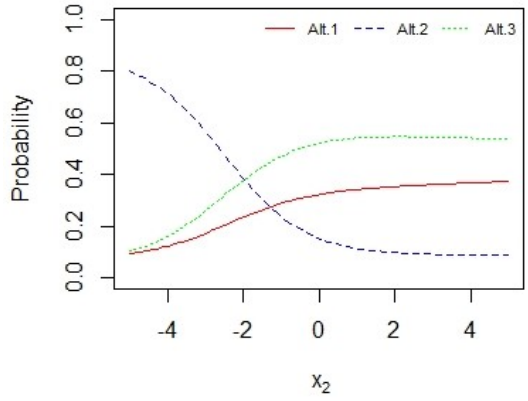
(a) Fitted probability curve of x_1 for the 2PL model



(b) Fitted probability curve of x_1 for the 4PL model



(c) Fitted probability curve of x_2 for the 2PL model



(d) Fitted probability curve of x_2 for the 4PL model

Figure 1: Fitted Probability Curves of the Individual-Specific Variables in the 4PL Estimation Model for the Basis 4PL Dataset

same outcomes are observed for fitted probability curves for x_2 based on Figure 1c and 1d.

When examining the parameter estimates, marginal effects and the fitted probability curves for the basis 4PL dataset, it is concluded that the model estimates vary: the estimates of the 4PL model are larger, the marginal effects are stronger, and the probability curves have steeper slopes that reach their asymptotes earlier. However, the question is whether these differences in results lead to different conclusions regarding people's true preferences when interpreting the results. This uncertainty arises from the fact that the sign of the marginal effects and parameter estimates is always the same for both estimation models. Furthermore, if a parameter estimate for the 4PL model is significant different from 0 (indicated by the absence of this value in the highest density

Table 6: Model Comparisons for the Datasets with Specific Features

Data	Model	Marg.lik.	DIC	Runtime	MPME	MADME
A: 8 alternatives	2PL	-1718.0	3484.3	3.7	0.23	0.037
	4PL	-1678.1	3417.1	4.4		
B: Product-spec. variables	2PL	-996.3	1997.7	1.3	0.52	0.081
	4PL	-990.5	1991.17	1.7		
C: Dummy variables	2PL	-876.3	1759.9	2.1	0.31	0.062
	4PL	-857.8	1723.7	2.6		

Note: The runtime is given in minutes.

interval), this is consistently observed for the 2PL model. Similar consistency is observed in the fitted probability curves. The 4PL model differs in details as the speed to reach the asymptote and the slope, but the overall shape of the probability curves seems to be the same. Consequently, while the 4PL model may lead to more precise estimates of the true preferences for this dataset, it is questionable whether this would lead to different conclusions compared to those drawn from the 2PL model.

5.2.2 Specific Data

Table 6 provides an overview of the marginal likelihood, DIC and other performance measures for datasets with different feature sets than the basis 4PL dataset. Across all datasets, the marginal likelihood and DIC consistently show a better performance of the 4PL model compared to the 2PL model. The runtime for all 4PL models is longer than those for the 2PL model, with a factor around 1.2. The MPME values highlight that the marginal effects of the 4PL models are stronger than for the 2PL models. Notably, the difference in model fit, indicated by the marginal likelihood and DIC, of dataset A and C is larger than for dataset B. This distinction is also reflected by the MPME, where the difference in marginal effects for both models is smaller for dataset B (0.52) compared to datasets A (0.23) and C (0.31).

5.3 Real-Life Data

The results from the simulated datasets suggest that implementing non-fixed asymptotes could be promising for enhancing the performance of discrete choice models and providing new insights into

Table 7: Model Comparisons for the Real-Life Datasets

Data	**Model**	**Marg.lik.**	**DIC**	**Runtime**	**MPME**	**MADME**
Fishing	2PL	-1186.7	2359.7	2.0	0.99	0.005
	4PL	**-1184.9**	**2358.3**	2.6		
Mode of Travel	2PL	**-360.0**	**721.0**	0.9	0.95	0.004
	4PL	-362.9	723.1	1.3		
Heating	2PL	-1017.9	**2047.7**	2.9	0.23	0.043
	4PL	**-1010.6**	2049.3	3.5		
Travel Mode Students	2PL	**-4725.9**	**9272.2**	15.1	0.98	0.002
	4PL	-4737.9	9292.3	17.1		

Note: The runtime is given in minutes.

the data. To assess whether the implementation also yields benefits for real data, four datasets from real-life applications were used. Table 7 shows the performance measures for the two estimation models applied to the real-life datasets. The best values for each model fit measure are highlighted in **bold** font. Next to this, Table 8 shows the estimated values for the lower asymptotes (c_j's) and the κ for the real-life datasets.

The differences in model fit for the real-life datasets between the models, given in Table 7, are generally smaller compared to the differences for simulated datasets. For the Fishing dataset, both model fit measures show a slightly better performance for the 4PL model than for the 2PL model (Marg.lik. 2PL: -1186.7 vs. 4PL: -1184.9, and DIC 2PL: 2359.7 vs. 4PL: 2358.3). However, the MPME of 0.99 indicates small differences between the models, a conclusion supported by the MADME of 0.005. The values in Table 8 show that slipping and carelessness could be present in this dataset, as the highest density interval of κ does not include the value of 1.

In the case of the Mode of Travel dataset, the marginal likelihood exhibits slightly better performance for the 2PL model (2PL: -360.0 vs. 4PL: -362.9), and the DIC supports the use of the 2PL as well (2PL: 721.0 vs. 4PL 723.1). Small differences are also demonstrated for the MPME (0.95) and MADME (0.004). Furthermore, the highest density interval of κ (see Table 8) includes the value of 1 which also could indicate that the dataset is obtained from a 2PL scenario.

Contradictory results are observed for the Heating dataset, where the marginal likelihood favors the 4PL model (2PL: -1017.9 vs. 4PL: 1010.6), the DIC slightly favors the 2PL model (2PL: 2047.7

Table 8: Parameter Estimates for the c_j's and κ for the Real-Life Datasets

Dataset	**Alternative**	**Variables**			
		c		κ	
Fishing	beach	0.004	[0.000,0.008]	0.964	[0.938,0.997]
	pier	0.003	[0.000,0.008]		
	boat	0.014	[0.000,0.037]		
	charter	0.007	[0.000,0.021]		
Mode of Travel	car	0.007	[0.000,0.023]	0.974	[0.942,1.000]
	carpool	0.007	[0.000,0.022]		
	bus	0.004	[0.000,0.013]		
	trail	0.005	[0.000,0.013]		
Heating	gas central	0.597	[0.549,0.640]	0.145	[0.083,0.220]
	gas room	0.108	[0.076,0.142]		
	electric central	0.051	[0.025,0.077]		
	electric room	0.055	[0.024,0.089]		
	heat pump	0.044	[0.026,0.062]		
Travel Mode Students	walk	0.000	[0.000,0.001]	0.999	[0.997,1.000]
	bike	0.000	[0.000,0.001]		
	car	0.000	[0.000,0.001]		
	public transport	0.000	[0.000,0.001]		

Note: The posterior means are given with the 95% highest density intervals between brackets.

vs. 4PL: 2049.3). The MPME of 0.23 and MADME of 0.043 highlight large differences in marginal effects. This could be attributed to the 2PL model showing large values for the intercept, while the 4PL model has relatively lower values for the intercept and higher values for the c's, which can be seen in Table 8.

Lastly, for the Travel Mode Students dataset, the results support the use of the 2PL model according to both performance measures (Marg.lik. 2PL: -4725.9 vs. 4PL: -4737.9, and DIC 2PL: 9272.2 vs. 4PL: 9292.3). However, the values of MPME (0.98) and MADME (0.002) indicate that the differences in marginal effects between both methods are minimal. The idea that this data is obtained from a 2PL scenario is also supported by the values in Table 8, where the value of κ is

almost equal to 1 and all the lower asymptotes to 0.

6 Conclusion

Careless responding or slipping could occur when someone makes a choice. When trying to examine the true preferences of individuals, these unintentional actions could influence the results of analyses. This research explores the inclusion of non-fixed asymptotes, which are normally used in IRT, in discrete choice models. The research question addressed is: *Does implementing the four-parameter logistic function of IRT into a discrete choice model provide an opportunity to account for mistakes in these models?* To answer this question, a multinomial logit model was extended using a four-parameter logistic function (4PL model) and was compared to a standard multinomial logit model (2PL model). This comparison was initially conducted on simulated datasets and subsequently on real-life datasets to assess the model's efficacy beyond simulated scenarios.

The results of the basis 2PL model indicate that, for data generated using a 2PL scenario, there is no significant difference between the two estimation models. This holds true for both model fit and parameter estimates. The small difference in marginal likelihood could be attributed to the correction for using extra parameters. An explanation for these nearly identical results could be that the 4PL model can adapt to a 2PL scenario by adjusting the values of c towards 0 and the value of κ towards 1. On the other hand, the outcomes of the estimation for the basis 4PL dataset reveal that, in this case, fixing parameters at 0 and 1 leads to an inaccurate representation of the parameters. The effects of variables were underestimated when the incorrect model (the 2PL model for a 4PL dataset) was used.

In summary, the results of the basis models suggest that using a 4PL model for 2PL data does not harm the parameter estimates. Conversely, using a 2PL model for 4PL data leads to underestimated and, consequently, incorrect values. However, it is questionable to what extent using the 2PL model actually leads to incorrect conclusions about behavior drawn from this model.

When evaluating the outcomes of the datasets with specific features, the similar conclusion is drawn that the 2PL model performs worse on 4PL data. These three datasets were simulated in a 4PL context, and the 4PL model showed a clearly better fit than the 2PL logit model. Once again, the marginal effects for the 2PL model were obviously underestimated compared to the 4PL model. This finding aligns with previous research by Loken and Rulison (2010) and Zhang et al. (2007), who observed a similar pattern when comparing 4PL and 2PL in the context of Item Response

Theory.

These results present a promising start for the 4PL multinomial logit model. However, when examining the outcomes of the real-life datasets, these results are not unambiguous. Some models show a better model fit for the 2PL estimation model, while others favor the 4PL model. In one case, the preferred model differs even between the two model fit measures for the same dataset. This difference could possibly be caused by the origin of the dataset, a 2PL or 4PL scenario. Besides this, the values of the MPME consistently indicate that, for all models, the marginal effects of the 4PL model are larger than those of the 2PL model. Moreover, it is worth noting that for the datasets where the marginal likelihood was higher for the 2PL models, the marginal effects were very close to each other (as reflected in the MPME and MADME values). Conversely, in the comparison of one dataset where the 4PL model performed better, the parameter estimates varied a lot between the two estimation models. For another dataset where the 4PL outperformed the 2PL model, the differences in parameter estimates were small. Overall, this supports the overarching idea that could answer the research question: if the 2PL model is the true model, the use of the 4PL model would not adversely impact parameter estimations. However, if the 4PL model is the correct one, it is possible that the parameter estimates and marginal effects derived from the 2PL model may not accurately represent reality.

At this point, the 4PL multinomial logit model appears to be a promising model for future use to detect carelessness and slipping in choice tasks. However, the model is more complex, leading to increased parameter uncertainty and runtime for the estimation model. This brings us to the second research question: *To what extent does the potentially improved performance of the four-parameter logistic discrete choice model outweigh the increased complexity of the model?* The increased complexity in this model is represented in both the runtime and parameter uncertainty of the model. In a 2PL scenario, the parameter uncertainty was the same for both models. However, in a 4PL scenario, the parameter uncertainty was larger for the 4PL model compared to the 2PL model. Notably, the intervals for the 2PL model did not consistently contain the true values, whereas this was the case for the 4PL model. Considering the parameter uncertainty, there is no reason opting for the 2PL over the 4PL. Analyzing the runtime of the estimation of the two different models reveals that the runtime of the 4PL model is consistently longer compared to the runtime of the 2PL model, as expected due to an increased number of parameters. The runtime ratio between the 2PL and 4PL models is, in almost all cases, around 1.2, with varying magnitudes of runtime. The maximum ratio observed is 1.4, which occurred for a model with a very short runtime. Even for the most

complex model (Travel Mode Students dataset with more than 8000 individuals and 7 variables), the runtime increased by a factor of 1.13. These results indicate a consistent and not exponential increase in runtime. Therefore, it appears that the potential improvement in performance outweighs the complexity of the model.

In conclusion, the extended multinomial logit model incorporating a four-parameter logistic function, originally from IRT, demonstrates its potential across various datasets. While substantial differences are not yet observed in real-life datasets, it is noteworthy that the use of this model does not adversely affect parameter estimates when the true model turns out to align with a standard multinomial logit model. In this case, the asymptotes can be simply set to 0 and 1.

7 Discussion

This section of the research discusses the limitations of the study and several ideas for future research are presented.

The first limitation of this research lies in the uncertainty regarding the practical use of the newly created 4PL multinomial logit model in real-life. While the model demonstrated superior performance for the simulated datasets constructed with a 4PL scenario, the outcomes for real-life datasets were mixed. Only one model showed a better model fit for the 4PL model, while other demonstrated a lower fit, possibly due to the correction for using additional parameters in the 4PL model, when the 2PL model was the correct model. The use of more parameters could be seen as overestimation of the model. The key question arises regarding whether the identified problem of mistakes in discrete choice models is present in real-life situations, or that the observed improvement was an incident that occurred for one real-life dataset. Further research could explore additional real-life datasets to determine whether other datasets also show enhanced performance for the 4PL model, or if the idea of non-fixed asymptotes is only applicable to simulated data. Exploring datasets where the impact of careless responding or slipping has been previously demonstrated, such as online surveys known for containing a significant number of careless respondents, could be an interesting start for further investigation.

An additional consideration regarding the possible absence of significant improvement in real-life datasets is that the possibility that the influence of the slipping and carelessness may be minimal in these datasets, making it challenging to detect. On top of this, in the IRT field, the probability fraction to choose an alternative can easily approach 0 (as shown in Formula 2). In such cases, the

probability of the corresponding alternative would also be close to 0. In contrast, in a multinomial logit model (as described in Formula 5), the probability of a particular alternative reaching 0 is less likely, given the presence of a large number of variables that can influence the probability. This suggests that small amounts of slipping and carelessness may already be effectively addressed by the standard multinomial logit model. Furthermore, it is important to note that in the simulated datasets of this research, a relatively high value of 0.1 for the lower asymptotes (c) was used. One could say that it is ridiculous to assume that individuals would choose based on their preference 70% of the time and make a mistake or behave careless 30% of the time. Accordingly, it could be valuable for further research to explore smaller values for the lower asymptotes in the simulation dataset to examine whether these can also be accurately determined by the estimation method and if the new method also shows enhanced performance in this case.

Additionally, the model's design could present a challenge. In psychometrics, where the four-parameter logistic function is commonly employed, the model is used to measure a latent trait (such as a student's mathematics ability) through multiple questions. Typically, the number of latent traits measured is lower than the number of questions each individual answers. Allowing for careless responding and slipping helps mitigate the impact of these actions on the measured latent trait. However, in the multinomial logit context where the individuals make choices identifying these mistakes becomes more complex, because each individual provides just one response. Parameter estimates of multiple variables are calculated based on this single choice with varying characteristics per individual. Accordingly, for further research, it could be interesting to explore the performance of the 4PL multinomial logit model on data which aligns more with an IRT scenario, where consumers need to make multiple decisions consecutively. Examples could include a Choice-Based Conjoint analysis format or scenarios where individuals make choices on different days regarding their preferred mode of transportation to work.

Another limitation not related to the model, but to a measure to compare the models. To answer the second research question, the comparison of increasing complexity and its consequences was partly based on the runtime of models estimated with the 2PL and the 4PL model. However, using the runtime for this purpose in this study comes with certain drawbacks. Firstly, the algorithm for calculating the estimates was programmed from scratch by the researcher. Therefore, it is possible that the program was not fully optimized, potentially leading to variations between the observed and real differences in runtime. However, it is important to note that even if the code was perfectly optimized, any impact on runtime would likely not be significant enough to result in exponential

increases. Despite this, the validity of runtime as a measure can be questioned. The underlying assumption is that the 4PL multinomial logit model is more complex and, consequently, requires a longer runtime. This runtime could be attributed not only to the additional multiplications, but also to the potentially longer convergence time of the model. In this research, a fixed number of iteration is used. In a real-world scenario, it is plausible that the estimation procedure is stopped when convergence is reached. When looking at the results of convergence in this paper (given in Appendix C), it can also be concluded that the mean, minimum and maximum PSRF is smaller on average for the 2PL models than for the 4PL models. This suggests that the 2PL model converges faster than the 4PL model. Consequently, this method would need less iterations if the estimation was stopped at convergence. This would which potentially increases the differences in runtime significantly.

To conclude the discussion, the four-parameter multinomial logit model demonstrates potential for future applications of non-fixed asymptotes in discrete choice models, ensured by its ability to deal with mistakes made by individuals, coupled with its accurate performance when these peculiarities are not in the data. However, further research is needed to discover to what extent these slipping and carelessness impact the parameter estimates in real-life scenarios. The four-parameter multinomial logit model could play a crucial role in examining this and provide valuable insights into the practical applications of incorporating non-fixed asymptotes in a discrete choice model.

References

Barton, M. A. and Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

Croissant, Y. and Graves, S. (2006). Ecdat: Data sets for econometrics. *R package version 0.1-5, URL http://CRAN. R-project. org*.

Embretson, S. E. and Reise, S. P. (2013). *Item response theory*. Psychology Press.

Fu, Z., Zhang, S., Su, Y.-H., Shi, N., and Tao, J. (2021). A gibbs sampler for the multidimensional four-parameter logistic item response model via a data augmentation scheme. *British Journal of Mathematical and Statistical Psychology*, 74(3):427–464.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

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

Hasselhorn, K., Ottenstein, C., and Lischetzke, T. (2023). Modeling careless responding in ambulatory assessment studies using multilevel latent class analysis: Factors influencing careless responding. *Psychological Methods*.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., van Dijk, H. K., et al. (2004). *Econometric Methods with Applications in Business and Economics*. OUP Oxford.

Herriges, J. A. and Kling, C. L. (1999). Nonlinear income effects in random utility models. *Review of Economics and Statistics*, 81(1):62–72.

Kalkan, Ö. K. (2022). The comparison of estimation methods for the four-parameter logistic item response theory model. *Measurement: Interdisciplinary Research and Perspectives*, 20(2):73–90.

Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63(3):509–525.

Lord, F. M. and Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Maniaci, M. R. and Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48:61–83.

Matějka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298.

Meade, A. W. and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3):437.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51:177–195.

Müller, S., Tscharaktschiew, S., and Haase, K. (2008). Travel-to-school mode choice modelling and patterns of school choice in urban areas. *Journal of Transport Geography*, 16(5):342–357.

Osgood, D. W., McMorris, B. J., and Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance i: Item response theory scaling. *Journal of Quantitative Criminology*, 18:267–296.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Reise, S. P. and Waller, N. G. (2003). How many irt parameters does it take to model psychopathology items? *Psychological Methods*, 8(2):164.

Reise, S. P. and Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5:27–48.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Sims, C. A. (2010). Rational inattention and monetary economics. In *Handbook of Monetary Economics*, volume 3, pages 155–181. Elsevier.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

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

Train, K. E. (2022). Kenneth train's home page. `https://eml.berkeley.edu/~train/`. Accessed on January 15, 2024.

Ulitzsch, E., Shin, H. J., and Lüdtke, O. (2023). Accounting for careless and insufficient effort responding in large-scale survey data—development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*, pages 1–22.

Ward, M. and Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, 74:577–596.

Ward, M. K. and Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on internet-based surveys. *Computers in Human Behavior*, 48:554–568.

Yen, Y.-C., Ho, R.-G., Laio, W.-W., Chen, L.-J., and Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, 36(2):75–87.

Zhang, Z., Hamagami, F., Lijuan Wang, L., Nesselroade, J. R., and Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, 31(4):374–383.

A Proof Sum of Probabilities

This section shows the proof that the sum of the probabilities of individual i always add up to 1 when the upper asymptote of alternative j and the lower asymptotes of all other alternatives add up to 1.

First, the restriction that the d_j and the c_k's of the other alternatives is rewritten to show that the term $(d_j - c_j)$ is constant for all the alternatives j.

$$d_j + \sum_{k \neq j}^{K} c_k = 1 \tag{19a}$$

$$d_j - c_j + \sum_{k \neq j}^{K} c_k + c_j = 1 \tag{19b}$$

$$(d_j - c_j) + \sum_{k=1}^{K} c_k = 1 \tag{19c}$$

$$(d_j - c_j) = 1 - \sum_{k=1}^{K} c_k \tag{19d}$$

$$(d_j - c_j) = (d_l - c_l), \quad j \neq l, \forall j, l \in K \tag{19e}$$

$$\kappa = (d_j - c_j), \forall j \in K \tag{19f}$$

In words, this derivation shows us that the value $(d_j - c_j)$ is constant for all alternatives, because the term $\sum_{k=1}^{K} c_k$ is constant across all alternatives.

This information will be used to show that all probabilities of individual i add up to 1, which is proven by rewriting the information from the derivation in Equation 19 in combination with the 4PL multinomial logit formula.

$$p_{ij} = c_j + \kappa \frac{e^{\beta_j x_i + \gamma_j z_{ij}}}{\sum_{k=1}^{K} e^{\beta_k x_i + \gamma_k z_{ik}}} \tag{20a}$$

$$p_{ij} = c_j + \kappa \frac{e^{\beta_j x_i + \gamma_j z_{ij}}}{\sum_{k=1}^{K} e^{\beta_k x_i + \gamma_k z_{ik}}} \tag{20b}$$

$$\sum_{j=1}^{K} p_{ij} = \sum_{j=1}^{K} \left(c_j + \kappa \frac{e^{\beta_j x_i + \gamma_j z_{ij}}}{\sum_{k=1}^{K} e^{\beta_k x_i + \gamma_k z_{ik}}} \right) \tag{20c}$$

$$\sum_{j=1}^{K} p_{ij} = \sum_{j=1}^{K} c_j + \kappa \sum_{j=1}^{K} \left(\frac{e^{\beta_j x_i + \gamma_j z_{ij}}}{\sum_{k=1}^{K} e^{\beta_k x_i + \gamma_k z_{ik}}} \right) \tag{20d}$$

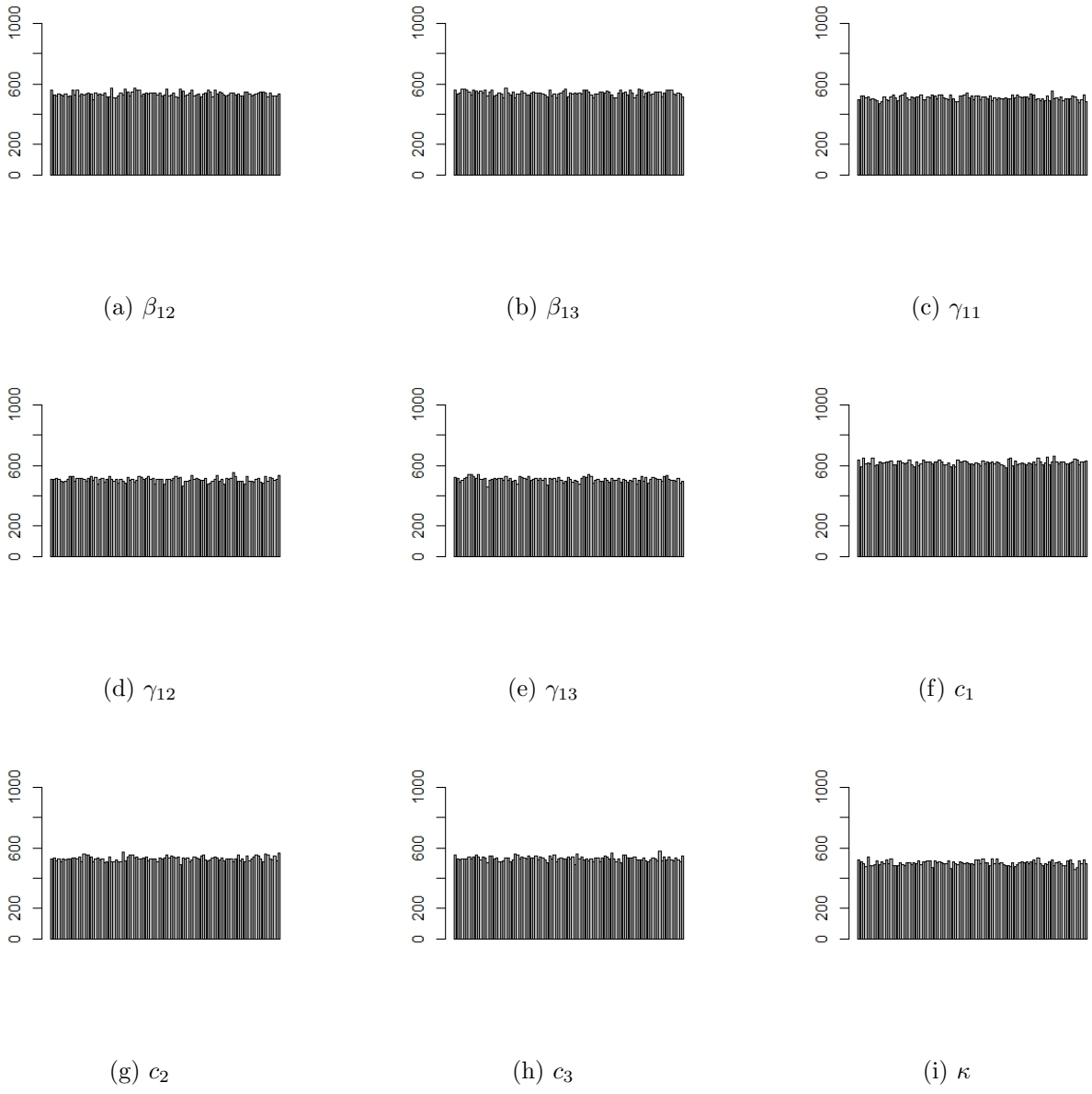
$$\sum_{j=1}^{K} p_{ij} = \sum_{j=1}^{K} c_j + \kappa \cdot 1 \tag{20e}$$

$$\sum_{j=1}^{K} p_{ij} = \sum_{j=1}^{K} c_j + (d_j - c_j) \tag{20f}$$

$$\sum_{j=1}^{K} p_{ij} = \sum_{k \neq j}^{K} c_k + d_j = 1 \tag{20g}$$

B Simulation-Based Calibration Results

The results of the Simulation-Based Calibration are shown in Figure 2. The approximately uniform distributions of the rank statistics show that the implementation was good. For β there is no value given for alternative 1, as this value is restricted to 0 for identification. Next to this, the SBC results are shown for one individual-specific variable and one product-specific variable as adding more results would not lead to new insights as the procedure for the different variables is the same.



(a) β_{12} (b) β_{13} (c) γ_{11}

(d) γ_{12} (e) γ_{13} (f) c_1

(g) c_2 (h) c_3 (i) κ

Figure 2: Rank Statistics

C Potential Scale Reduction Factor Results

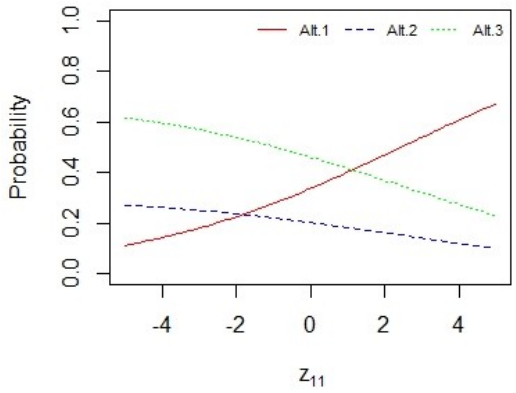
The PSRF results of for all datasets are given. For one dataset the maximum value exceeds the threshold of 1.1. It is also interesting to note that on average the PSRF values of the 4PL model are larger than for the 2PL model, indicating that the 2PL model converges faster than the 4PL model.

Table 9: Descriptive Statistics of PSRF Results for All Datasets and Models

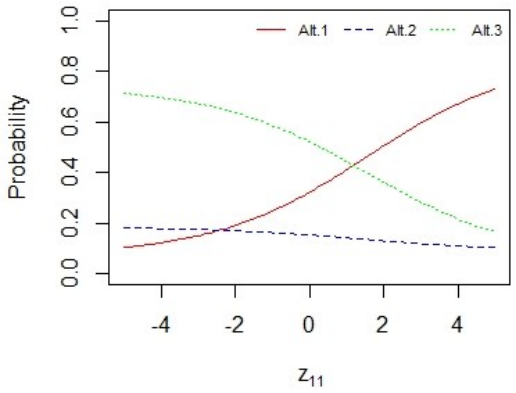
Data	**Model**	**Mean**	**Min.**	**Max.**
Basis 2PL	2PL	1.002	1.001	1.005
	4PL	1.002	1.000	1.005
Basis 4PL	2PL	1.002	1.000	1.004
	4PL	1.002	1.001	1.006
8 alternatives	2PL	1.007	1.000	1.018
	4PL	1.020	1.000	1.073
Product-specific variables	2PL	1.001	1.000	1.002
	4PL	1.004	1.000	1.007
Dummy variables	2PL	1.002	1.000	1.007
	4PL	1.003	1.000	1.007
Fishing	2PL	1.004	1.002	1.008
	4PL	1.011	1.003	1.045
Mode of Travel	2PL	1.003	1.001	1.008
	4PL	1.003	1.000	1.007
Heating	2PL	1.003	1.000	1.014
	4PL	1.003	1.000	1.008
Travel Mode Students	2PL	1.049	1.003	1.118
	4PL	1.026	1.002	1.095

D Fitted Probability Curves

This section shows the fitted probability curves of the product-specific variables of the 2PL and 4PL model for the basis 4PL dataset. The same conclusions can be drawn as for the fitted probability curves of the individual-specific variables: the curves for the 4PL model reach their asymptotes faster and have steeper slopes.

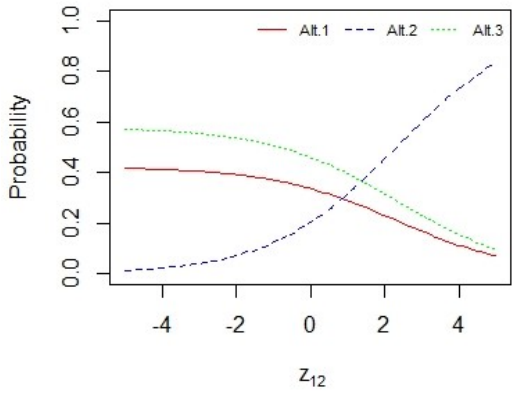


(a) Fitted probability curve of z_{11} for the 2PL model

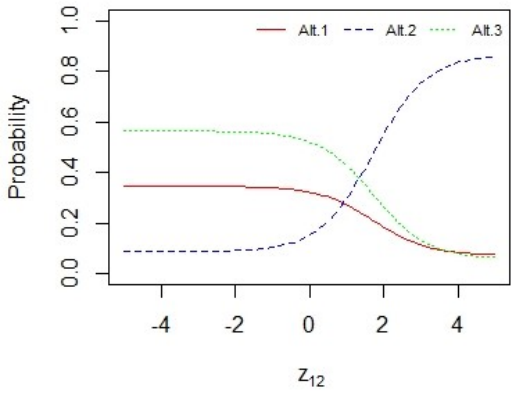


(b) Fitted probability curve of z_{11} for the 4PL model

Figure 3: Fitted probability curves of z_{11} for the basis 4PL dataset

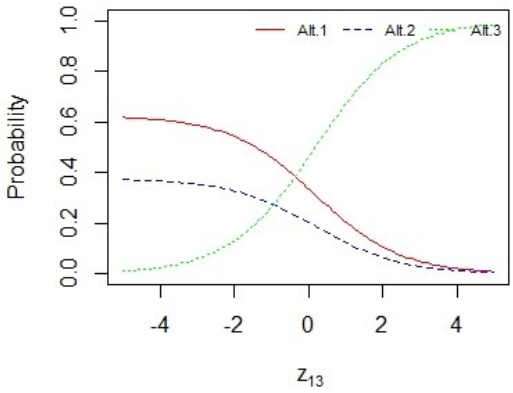


(a) Fitted probability curve of z_{12} for the 2PL model

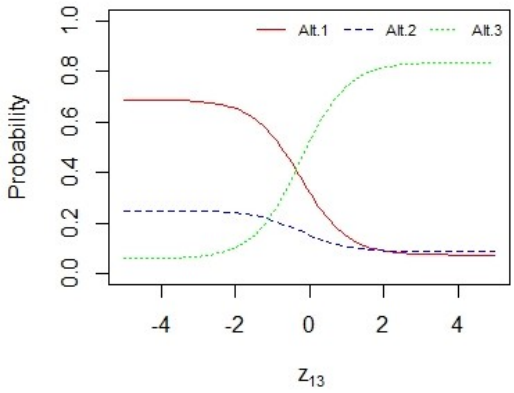


(b) Fitted probability curve of z_{12} for the 4PL model

Figure 4: Fitted probability curves of z_{12} for the basis 4PL dataset

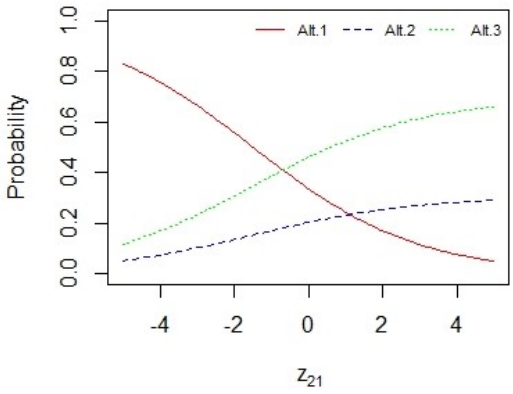


(a) Fitted probability curve of z_{13} for the 2PL model

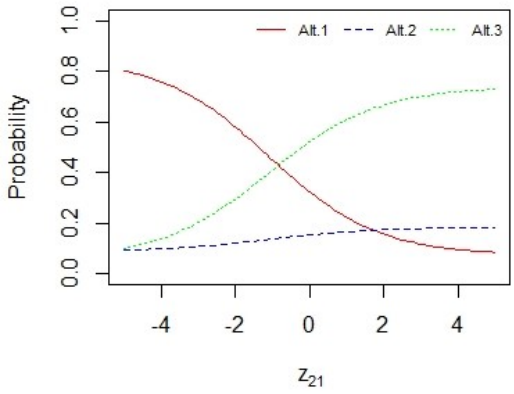


(b) Fitted probability curve of z_{13} for the 4PL model

Figure 5: Fitted probability curves of z_{13} for the basis 4PL dataset

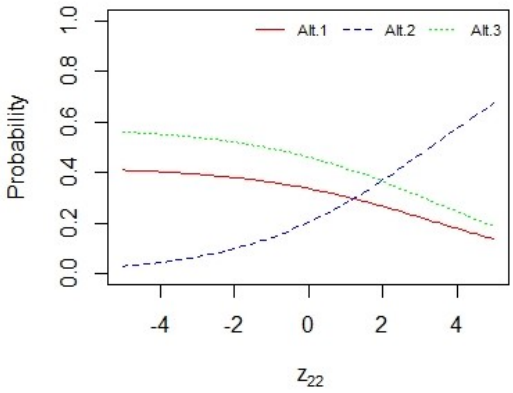


(a) Fitted probability curve of z_{21} for the 2PL model

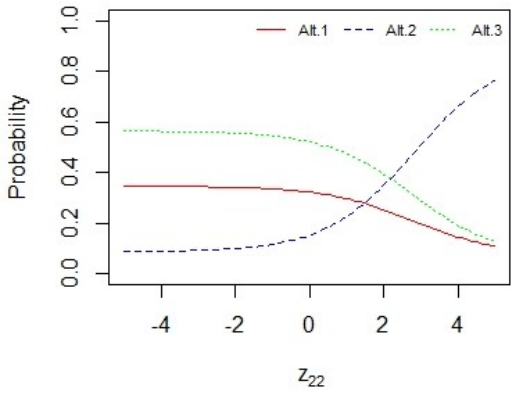


(b) Fitted probability curve of z_{21} for the 4PL model

Figure 6: Fitted probability curves of z_{21} for the basis 4PL dataset

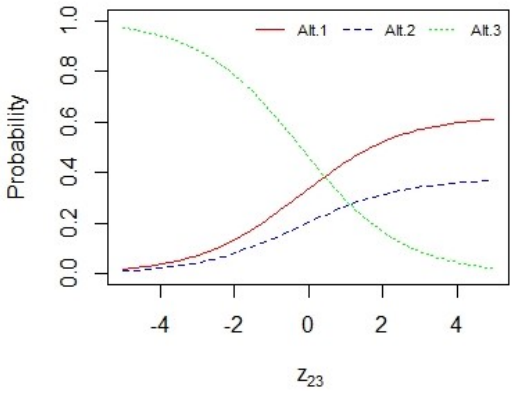


(a) Fitted probability curve of z_{22} for the 2PL model

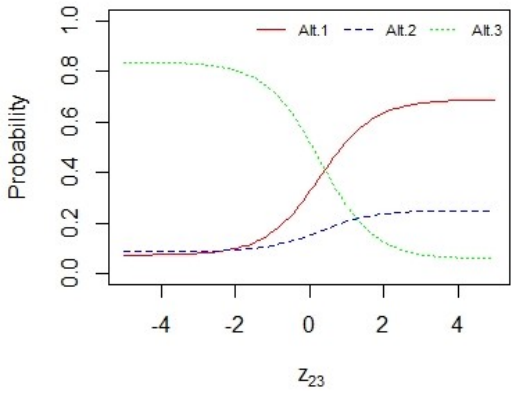


(b) Fitted probability curve of z_{22} for the 4PL model

Figure 7: Fitted probability curves of z_{22} for the basis 4PL dataset



(a) Fitted probability curve of z_{23} for the 2PL model



(b) Fitted probability curve of z_{23} for the 4PL model

Figure 8: Fitted probability curves of z_{23} for the basis 4PL dataset