Journal of Choice Modelling 56 (2025) 100562

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

Understanding the decision-making process of choice modellers

Gabriel Nova [a,*], Sander van Cranenburgh [a], Stephane Hess [a,b]

[a] Transportation and Logistics Group, Department of Engineering Systems and Services, TU Delft, Netherlands
[b] Choice Modelling Centre - Institute for Transport Studies, University of Leeds, England, United Kingdom

ARTICLE INFO

Keywords:
Choice modelling
Serious game
Choice modeller workflows

ABSTRACT

Choice Modelling is a widely used framework for understanding human choice behaviour across disciplines. Building a choice model is a complex, semi-structured process that involves a combination of prior assumptions, behavioural theories, and statistical methods. This complex set of decisions, coupled with diverse workflows, can lead to substantial variability in model outcomes. To investigate these modelling processes, we introduce the Discrete Choice Modelling Serious Game (DCM-SG), a novel tool that mimics the workflow of choice modellers and tracks the modelling decisions participants make. In our application, participants developed models to estimate willingness-to-pay values for reducing noise pollution. Their actions were tracked, enabling analysis of workflow patterns and modelling strategies. Forty participants, most with over five years of experience, completed the game. Our contributions are twofold. Methodologically, the DCM-SG captures sequential data on modellers' workflows, which we analyse using telemetry and sequential pattern mining techniques to uncover dynamic patterns of in-game tool usage, phase transitions, and model specification approaches. Substantively, there was a strong preference for data visualisation and frequent specification of simpler models (Multinomial Logit), alongside attempts to specify more complex models. These findings suggest that in time-constrained or resource-limited settings, modellers may underexplore important factors such as covariates, non-linearities, and complex specifications. Moreover, participants who engaged more thoroughly in data exploration and iterative model comparison consistently achieved superior model fit and parsimony. These results demonstrate how sequential data from the DCM-SG can uncover variations in modelling practices and provide a foundation for understanding the art of choice modelling.

1. Introduction

Discrete Choice Modelling (DCM) is a theoretical and applied framework used across various scientific disciplines to study human choice behaviour. These fields include, but are not limited to, transport, health, and environmental economics (Louviere, 2000; Hess and Daly, 2024; Mariel et al., 2021; Haghani et al., 2021). The aim of this work is to specify models for explaining current choices and predicting future choices (Ben-Akiva and Lerman, 1985). On the one hand, by calibrating models on empirical choice data, choice modellers can estimate and infer preferences over alternatives and their attributes, which correspond to features or qualities that define them. On the other hand, by using the estimated parameters to simulate choice scenarios, they can predict future behaviour and responses to changes in policy or market conditions. This enables analysts not only to study the decision-making process and the factors that influence individual decisions, but also to analyse choice behaviour in different contexts, forecast demand, and evaluate policy changes (Ben-Akiva and Bierlaire, 2003).

* Corresponding author.
E-mail address: G.Nova@tudelft.nl (G. Nova).

https://doi.org/10.1016/j.jocm.2025.100562
Received 30 September 2024; Received in revised form 16 July 2025; Accepted 24 July 2025
Available online 25 August 2025

Discrete choice modelling brings together individuals with diverse backgrounds and expertise to understand and forecast choice behaviour through a series of research steps. Regardless of the purpose, choice modellers typically engage in workflows that involve formulating a research question, designing experiments, collecting data, conducting exploratory and descriptive analyses, specifying models, interpreting model outcomes, and reporting results (Ben-Akiva and Lerman, 1985; Hensher et al., 2015; Mariel et al., 2021). Throughout each research phase, these professionals balance various formal behavioural theories with statistical methods, experimental applications, their own domain knowledge, and professional judgement to develop models that represent the choices in the data under study (Paz et al., 2019). Although these phases generally follow a chronological order, they are often conducted in a semi-structured manner. This flexibility enables modellers to determine how and when to incorporate subjective knowledge acquired during the process, ultimately guiding the selection of a model after evaluating multiple specifications (Rodrigues et al., 2020; Van Cranenburgh et al., 2022). This inherent flexibility, combined with diverse workflows and subjective decision-making, can lead to model specifications that do not truly capture the data generation process. Modellers may interpret data differently, emphasise different aspects of the model's functional form, or even select model families that do not align with observed preferences, which in turn leads to considerable variability in modelling results and conclusions. For this reason, choice modelling is often considered an art, involving a high degree of freedom in decision-making that allows modellers to use their expertise to make decisions within their research.

Hitherto, studies have focused primarily on the model specification phase (Daly et al., 2012; McFadden, 1974; McFadden and Train, 2000; Walker and Li, 2007). This phase involves a trial-and-error process in which choice modellers determine both the model structure and the parameters to be considered in the model (Paz et al., 2019). For instance, analysts must decide on the model family, the inclusion of linear or non-linear transformations of variables, the incorporation of observed and unobserved heterogeneity, the distribution of the random coefficients, and their potential correlations, among other considerations (Train, 2009; Beeramoole et al., 2023; Mariel et al., 2021). This iterative and time-consuming process continues until modellers have estimated each desired specification and obtained goodness-of-fit indicators and validation metrics to assess model performance, parameter consistency, and alignment with existing literature (Parady et al., 2021). As modellers may use different cost functions to balance these goodness-of-fit metrics, this process may yield multiple specifications that are deemed acceptable to address their research question.

The current choice modelling landscape reveals a knowledge gap in understanding modellers' decision-making processes. Although recent developments have introduced algorithms designed to assist in model specification (typically using goodness-of-fit indicators as objective functions) (Paz et al., 2019; Ortelli et al., 2021; Beeramoole et al., 2023), these algorithms only partially automate certain specification decisions. They fail to account for decisions made during the descriptive analysis phase, the trade-offs made during model specification to constrain the model space, and the model selection at the end of this process. This inherent flexibility of choice modellers' workflows can lead to diverse results, interpretations, and conclusions even when working with the same choice dataset. Similar concerns have been observed in psychology, where concepts such as *researcher degrees of freedom* (Simmons et al., 2011) and the *garden of forking paths* (Gelman and Loken, 2013) emphasise how flexibility in data collection and multiple potential tests based on the same dataset, along with the pursuit of meaningful parameters, can increase the risk of false positives. Furthermore, crowd-science experiments have demonstrated significant variability in research processes, highlighting a lack of consensus in decision-making and divergent outcomes when different researchers analyse the same data (Botvinik-Nezer et al., 2020; Wicherts et al., 2016; Silberzahn et al., 2018). While these degrees of freedom promote exploration and methodological innovation, they also carry a risk of poor decision-making and undesirable outcomes. A better understanding of these processes not only fosters debate about best practices but also paves the way for improved practices within the modelling community.

This study aims to shed light on the choice modelling research process by introducing a novel methodological approach that combines the design of a serious game, empirical data collection, and sequential pattern mining analysis. Specifically, it presents the first serious game designed for investigating the workflow of choice modellers. The game covers the entire choice modelling process, from data exploration and model specification to outcome interpretation and results reporting. The serious game was conducted by participants at two international conferences and an additional sample of online recruits, with their actions tracked throughout. Rather than evaluating participants against a predefined "best" modelling strategy, this study aims to reveal empirical insights into how modellers navigate the modelling process, respond to feedback (such as model fit statistics), and iterate on their modelling assumptions. These data then offer a unique opportunity to explore modelling workflows, reveal the degrees of freedom available to choice modellers, and analyse how their use of in-game tools and workflows influences reported results. In doing so, this study not only contributes a methodological instrument for analysing decision-making in choice modelling but also demonstrates how sequential data derived from the game can reveal differences in modelling practices and serve as a starting point for understanding these workflows.

The remainder of the paper is structured as follows. Section 2 discusses related work, focusing on the progression from data analysis to modelling outcomes. It also introduces a conceptual framework, which outlines the modelling phases that inform our serious game design. Section 3 describes the methodology, covering the serious game design, the in-game tools available throughout the choice modelling process, and the stated preference dataset used. Section 4 details the gameplay session, while Section 5 discusses the results and gives an overview of modellers' workflows. Finally, Section 6 presents the main conclusions.

2. Theoretical framework: Beyond data analysis to modelling

DCM enables analysts to uncover preferences and forecast choices across a wide range of disciplines. Unlike conventional data analysis approaches, DCM involves not only the application of statistical techniques, but also requires analysts to build, evaluate,

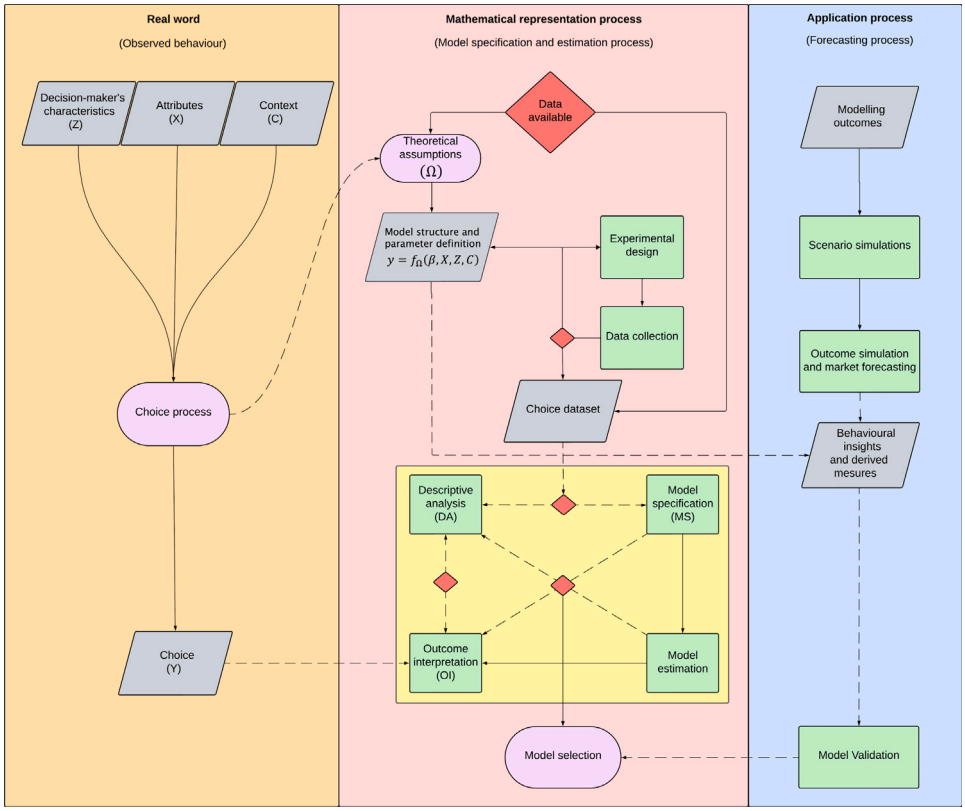


Fig. 1. Conceptual overview of the choice modelling research process.

and validate models that capture the underlying trade-offs individuals make when choosing among alternatives (Ben-Akiva and Lerman, 1985). While some modellers focus on developing and testing theoretical frameworks, others apply these models to real-world datasets to gain insights into specific choice contexts. Their tasks include collecting and analysing choice data, developing and validating models, interpreting results, and calculating econometric metrics to better understand decision-making and the factors that influence choice behaviour. Given the complexities involved in the modelling process, this section introduces a general framework that outlines modellers' degrees of freedom and the main research phases for developing choice models, grounded in well-established literature (e.g., Ben-Akiva and Lerman (1985), Train (2009), Louviere (2000), Mariel et al. (2021), Hess and Daly (2024), among others).

Fig. 1 presents a theoretical framework for discrete choice modelling, illustrating the conceptual flow from observed choice behaviour to formal model representation and subsequent application cases. The framework is organised into three interconnected processes, each demarcated by vertical sections: the real-world choice process (left), the mathematical representation process (centre), and the application process (right). In the diagram, processes are represented by rectangles; data, inputs and outputs are depicted by parallelograms; internal procedures performed by decision-makers or choice modellers appear as ovals; and modeller decisions are indicated by diamonds. Solid arrows denote sequential progression, while dashed arrows indicate influences between components.

In the real world (left), the decision-makers' choice process is represented using the canonical representation, where individuals with specific characteristics (Z) face a choice situation defined by alternatives and their associated attribute levels (X) within a given hypothetical context (C). These inputs guide the decision-makers' choice process, which ultimately results in an observable choice outcome (Y) that modellers aim to understand and forecast.

The mathematical representation process captures the entire modelling workflow, including theoretical formulation, data handling and exploration, model estimation, and model selection. To model real-world choices, choice modellers must begin with a choice dataset that addresses their initial research question, which is shaped by both theoretical interests and data availability. If a suitable dataset already exists, choice modellers may proceed directly to data handling and exploration. However, in the absence of an existing dataset, several decisions about experimental design and data collection methodologies must be made. These decisions are again guided by the modellers' theoretical assumptions (Ben-Akiva and Lerman, 1985).

Experimental design and data collection require modellers to make several decisions, often influenced by their theoretical assumptions and subjective judgements, which ultimately shape the resulting dataset. Although discrete choice models can be estimated using revealed preference (RP), stated preference (SP), or both, this study exclusively focuses on SP data. RP involves observing actual choices through methods such as self-reported data, passive data collection, activity diaries, or even psychophysiological measures (Bierlaire and Frejinger, 2008; Arriagada et al., 2022; Hancock et al., 2022; Barría et al., 2023; Xu

et al., 2018). For instance, in transport modelling, RP data collection involves determining the basic unit of analysis, the spatial range, and essential information such as origin, destination, purpose, start time, end time, payment method, transport mode, and socio-demographic variables (Axhausen, 2024). While RP data reflects real behaviour, its design is inherently limited to existing and observed alternatives, making it difficult to evaluate preferences for hypothetical or future scenarios.

In contrast, SP experiments allow modellers to design hypothetical choice tasks that capture preferences and choice behaviour under controlled conditions. Designing such experiments involves a series of sequential decisions that not only affect the ability to identify primary attribute effects but also influence the statistical power to detect significant relationships between variables, particularly in studies with typical sample sizes (Rose and Bliemer, 2009). These decisions include defining the number of choice tasks, the alternatives, the attributes, and their levels (Hensher, 2004; Caussade et al., 2005; Rose and Bliemer, 2009; Meyerhoff et al., 2015; Meißner et al., 2016; Mariel et al., 2021). Moreover, RP data can be used to inform SP design by suggesting attribute range values or common trade-offs, thereby improving the plausibility of hypothetical scenarios. The survey instrument is then constructed and potentially refined through pilot studies or focus groups.

The resulting dataset is used across several interconnected modelling phases: descriptive analysis (DA), model specification (MS), model estimation (ME), and outcome interpretation (OI). These phases are interdependent, forming a feedback loop that reflects the trial-and-error nature of refining model assumptions and specifications until a suitable model — or multiple models — is/are selected. It is well established that most modellers begin this process with exploratory and descriptive analysis, which is essential for understanding the structure and composition of the data, as well as for preprocessing it for subsequent modelling tasks. This phase includes statistical analysis, graphical representation, handling missing values, and recoding or scaling to prepare the data for model estimation. Beyond these analyses, exploratory analysis plays an important role in revealing relationships and correlations between attributes, covariates, and outcomes. These insights guide the formulation of initial modelling hypotheses, offering a sense of which variables to include, how they may interact, and what behavioural patterns are being modelled. Although some modellers might skip this phase to save time, doing so risks overlooking valuable information that could significantly inform and improve model specification.

Model specification is a pivotal phase where modellers translate theoretical assumptions into a formal mathematical representation, determining both the functional form and the behavioural mechanisms to be captured. Modellers must first choose a model family based on assumptions about the decision process (e.g. compensatory *vs* non-compensatory), followed by within-family decisions, for example the distribution of error terms and heterogeneity across decision-makers (e.g., Multinomial Logit (MNL), Generalised Extreme Value (GEV), Latent Class (LC), Mixed Logit (MMNL)). Then, modellers must determine the model specification itself, including which variables to include, how to include them (e.g., linear, logarithmic, piecewise), and whether to incorporate interactions with covariates (e.g., gender, age, household composition) (Ortelli et al., 2021; Rodrigues et al., 2020). These degrees of freedom are not only essential for capturing the complexity of real-world choices (Beeramoole et al., 2023), but also directly affect both model interpretability and performance (Parady et al., 2021; Van Cranenburgh et al., 2022). This is an iterative process involving the estimation of multiple model specifications, validation of their internal and external performance, and the interpretation of model outcomes through comparisons of model fit, parameter consistency, and alignment with existing literature (for more details, see Parady et al. (2021)).

In the final phase, the selected model(s) are applied to generate behavioural insights, forecast demand, and evaluate policy scenarios across different contexts. This involves using the estimated parameters to derive measures such as willingness-to-pay (WTP), elasticities, and market shares, and predict responses to changes in policy or market conditions. The application process enables choice modellers to derive practical implications from their models, providing valuable insights and guidance for policymakers and practitioners.

To summarise, the choice modelling framework provides a structured approach to understanding and forecasting individual choice behaviour, by capturing the choice process through several model specifications. Beyond data collection and analysis, the multiple workflows and decisions involved in model specification, estimation, outcome interpretation, and model comparison can create an opaque environment for both understanding the actual data generation process and the factors that influence decision-makers in the choice situation studied by choice modellers. This has significant implications not only for the field itself, but also for the reliability of the modelling outcomes on which policymakers rely for policy formulation and analysis. Enhancing transparency and addressing these degrees of freedom can strengthen the faith that policymakers place in the results provided by choice modellers.

3. Method

3.1. Serious games for choice modelling

Serious Games (SGs) are game-based tools designed with a primary purpose other than entertainment, particularly for addressing real-world problems by providing training, learning, or encouraging behavioural change for participants (Michael and Chen, 2005). Characterised by explicit rules and defined goals, SGs are intentionally designed to be applied to relevant, often complex, issues where players can experiment in a safe environment with different in-game tools and see the consequences of their decisions (Corti, 2006; Dörner et al., 2016; Squire, 2006). Although initial applications aimed to improve decision-making skills in diverse simulated environments — such as educational (Rawitsch, 1978) and military combat contexts (Krulak, 1997) — later efforts have concentrated on tracing and analysing players' actions and behaviours within game environments (Medler and Magerko, 2011).

SGs have gained popularity in behaviour evaluation due to their potential to capture real-time data and support decision-making analysis in different contexts, such as policy formulation, resource allocation, or scenario analysis (Donaldson and Grant-Vallone,

2002; Olejniczak et al., 2020; Van Dijk and De Dreu, 2021). In transport research, SGs have been used to explore diverse applications, including road safety education (Plass et al., 2015), the assessment of perceived transport justice (Vecchio, 2024), and the promotion of public participation in urban planning processes (Poplin, 2012). Moreover, they have been employed for training transport planners through role-playing and resource allocation exercises in urban mobility contexts (Ortúzar and Willumsen, 1978, 1982; Willumsen and Ortúzar, 1985). These games often simulate realistic mobility scenarios and allow for observing responses in controlled settings.

Although data collection can be conducted during different phases — such as before, during, and after gameplay — which offers diverse methods and data types (Mayer et al., 2014), most early SGs did not focus on capturing in-situ information on player behaviour (Smith et al., 2015). Surveys, questionnaires, and self-reports were the most commonly used methods due to their ease of implementation and data collection. However, these fail to provide detailed insights into decision-making processes, as participants may adjust their behaviour during the game to align with the researchers' objectives (Podsakoff et al., 2003), and do not capture participants' actual thinking processes (Lazar et al., 2017). Therefore, with technological progress, efforts have been made to collect behavioural data during gameplay in simulated environments with controlled variables, which can be used to demonstrate relationships between players' knowledge, degrees of game experience, or backgrounds and their performance in different tasks (Chung, 2014; Snow et al., 2014).

Integrating SG into choice modelling offers a novel method for understanding the decision-making process, conceptualised as a sequence of interconnected and complex research phases. The Discrete Choice Modelling Serious Game (DCM-SG) was therefore developed based on the conceptual framework outlined in Section 2, assuming the availability of a previously collected SP choice dataset. Based on this, the game allows users to move through the phases of descriptive analysis, model specification, estimation, and outcome interpretation, with the objective of deriving measures such as willingness-to-pay estimates. Capturing modellers' decisions as real-time data facilitates an analysis of actual behaviour throughout the performed workflows. Furthermore, the collected data enables an analysis of the relationships between participants' experience, field of expertise, and knowledge, and how these factors influence their decision-making processes.

3.2. Discrete choice modelling serious game design

To reveal how choice modellers make decisions throughout the modelling process, we designed and developed the DCM-SG based on the framework proposed by Meijer (2009). The game is intended for students, researchers, practitioners, and analysts with at least a basic understanding of choice modelling. No programming skills are required. All modelling actions are performed through a user-friendly interface, where predefined triggers guide participants through the research phases, such as data exploration, model specification, and outcome interpretation.[1]

During the game session, participants are assigned the role of choice modellers, and the game is played within a reference system. This reference system is confined to a real-world problem that involves developing choice models using a stated preference choice dataset, where participants with different backgrounds may apply different approaches to solve the problem. Participants are presented with a well-defined context: *"Imagine a colleague has asked for your help in analysing a stated choice dataset on residential location preferences. This dataset has been designed to determine the willingness-to-pay for reducing noise pollution. Respondents were faced with three unlabelled neighbourhoods (A, B, C) and asked to select the one they preferred to live in. The data were collected across four different cities and are representative of the target population"*. Thus, they are required to apply their knowledge, technical skills, modelling expertise[2] to generate modelling outcomes to inform policymakers, while their actions and responses are tracked. Specifically, the game objective is stated as: *"Develop a choice model to estimate the Willingness-to-Pay (WTP) for noise pollution reduction. Your WTP estimate will be used by policymakers to make informed urban planning decisions"*, as shown on the instruction page (Fig. 8). Therefore, this problem-solving context situates our method within the testing-and-retention quadrant, as described by Olejniczak et al. (2020).

To ensure that the game's design was appealing and as realistic as possible, we specified clear rules and constraints throughout the game experience. These were explained at the beginning of the game sessions and also appeared within the game instructions. On the one hand, the rules stated that participants could perform any allowable actions within the game, while avoiding the sharing of information or results with other participants during the 45 min game session. On the other hand, certain constraints were introduced to limit the degrees of freedom available in the model specification phase. These restrictions, detailed in Section 3.3, include limits on the number of random parameters in MMNL models, a maximum of three latent classes in LC models, and the inclusion of all attributes in certain model types, along with restrictions on interaction terms, to constrain the modelling space.

3.3. DCM-SG tools

To facilitate participants' workflows in solving this problem, the game simulates four main phases of choice modelling research, derived from the conceptual framework: descriptive analysis (DA), model specification (MS), outcome interpretation (OI), and reporting (R). Participants can iterate through each phase as many times as they deem necessary before moving on to the next phase,

[1] For source code and installation instructions, please refer to the project's GitHub repository: https://github.com/TUD-CityAI-Lab/DCM-SG.

[2] Practical judgement, modelling heuristics, and preferences that modellers develop through experience but are rarely formalised in guidelines or training courses.

or even returning to previously completed phases, thus mimicking the intrinsic trial-and-error nature of the modelling process. This iterative approach, which aligns with the conceptual framework depicted in Fig. 1, enables participants to continually refine their specifications and analyses until they report their findings.

Descriptive analysis: During this phase, we incorporated in-game tools based on general practices in choice modelling and recommendations from the literature (Tukey, 1977; Páez and Boisjoly, 2022). This allows modellers to perform a range of exploratory data analysis tasks on the raw choice dataset, enabling them to better understand the data distribution prior to the model specification phase. Specifically, participants can view a data dictionary to understand the context and meaning of each variable; inspect the first five rows of the dataset; and consult descriptive statistics (e.g., mean, median, max, min, standard deviation) for each variable. Participants can also examine the frequency of choices (i.e., the distribution of choices made by decision-makers) and see an example of a choice task. Additionally, they can sort the dataset by any variable. For handling missing data, participants can display missing values, delete missing values, or replace them using the mean, mode, or median. Finally, the visualisation tools available include box plots, histograms, correlation matrices, scatter plots, pie charts, and bar charts, which allow users to plot any variable and explore the structure of the dataset. While we acknowledge that the list of actions is not exhaustive, these tools were selected to reflect techniques commonly used in the field and to maintain consistency among participants. Thus, our serious game ensures comparability across workflows and highlights how participants engage with the data under similar conditions. An overview of this phase is shown in Fig. 9.

Model specification: During this phase, researchers are given the flexibility to specify models from diverse families that are foundational and widely used in discrete choice modelling, as shown on the respective page (Fig. 10). The base model is the Multinomial Logit (MNL) model (McFadden, 1974), which is the workhorse and benchmark of discrete choice models. To specify an MNL model, the modeller must make various decisions, such as incorporating alternative-specific constants, including attributes, selecting between generic or alternative-specific coefficients, and considering interactions with socio-demographic variables. They can also apply non-linear transformations, such as logarithmic and Box–Cox functions, to account for nonlinearities.

To capture unobserved heterogeneity, the DCM-SG also allows the specification of the Mixed Multinomial Logit model (McFadden and Train, 2000), which is commonly used to address random taste variation across individuals and correlation in unobserved factors, making it a powerful tool for modelling a range of behavioural specifications (Train, 2009). Within this model family, modellers can decide which parameters follow a random distribution (normal or lognormal), while maintaining the remaining ones as population-level parameters. Additionally, they can still account for observed heterogeneity by interacting attributes with socio-demographic variables. Due to the exponential number of specifications a modeller might consider, some limitations are introduced for MMNL models:

1. Alternative-specific constants are included for all models.
2. All attributes are included in the utility function and treated as generic across alternatives.
3. The number of draws is fixed and constant across all models; this aspect is not analysed.
4. A maximum of two random parameters can be included in the utility specification.
5. Interactions between a random parameter and a socio-demographic variable are not permitted.

The serious game also considers Latent Class models (Walker and Li, 2007), which enable the probabilistic allocation of decision-makers into discrete classes by assuming that each class has distinct preferences. This model is included as it is among the most extensively used in the choice modelling literature (Hensher et al., 2015). In these model specifications, modellers need to not only define the number of latent classes, but also have to determine which covariates to include in the class membership function to calculate the probability of belonging to each class. Similar to the MMNL case, some limitations are introduced:

1. Alternative-specific constants are included.
2. All attributes are included in the utility function and treated as generic across alternatives.
3. Models may include up to three latent classes.
4. Covariates are dummy coded and may be included in the class allocation model.

We decided not to include machine learning (ML) models. While we acknowledge the growing interest in models such as Random Forests, Support Vector Machines, Gradient Boosting Decision Trees, and Artificial Neural Networks for analysing choice preferences, their adoption is still relatively limited (Hagenauer and Helbich, 2017; Wang et al., 2017, 2020; Martín-Baos et al., 2023). We believe that including ML models would have introduced unnecessary complexity rather than enhance the realism of the serious game.

Finally, regarding the estimation of the models, MNL models are estimated on the fly upon user request. To enable this, the DCM-SG integrates Biogeme methods and classes, allowing users to dynamically request and estimate any specification within this model family according to their specific needs. Due to the longer estimation times required for MMNL and LC models, we pre-estimated them using Delft Blue (Delft High Performance Computing Centre (DHPC), 2024) and local machines, employing both Apollo (Hess and Palma, 2019) and Biogeme (Bierlaire, 2003). The results generated by these software packages were stored in a database to serve as a repository for the outcome interpretation phase. In total, we pre-estimated 78,604 MMNL models and 8832 LC models. Lastly, we calculated the standard errors of the willingness-to-pay estimates using the Delta method (Daly et al., 2012). Both the WTP values and their standard errors are made available to participants in the outcome interpretation phase.

Outcome interpretation: During this phase, we display the common modelling results table for the estimated model, which contains parameter names along with their estimated values, standard errors, t-tests, and p-values, as shown on the respective page (Fig. 11). These results represent the most elementary outputs following model estimation and provide the information needed to

Table 1
Data dictionary.

Variable	Description	Type/Levels
ID	ID number of the respondent	Integer
Task ID	Number of the choice task	Integer
Stores	Walking time to grocery store	2, 5, 10, 15 min.
Transport	Walking time to public transport stop	2, 5, 10, 15 min.
City	Distance to city centre in kms	<1, 1 to 2, 3 to 4, > 4 km
Noise	Street traffic noise	None, Little, Medium, High
Green	Green areas in residential neighbourhood	None, Few, Some, Many
Cost	Monthly change in housing cost *vs* current	-€150, -€50, €50, €150
Choice	Indicates the choice	1 = A, 2 = B, 3 = C
Age	Decision-maker age in years	< 30, 30 to 50, ≥ 50
Woman	Indicates if respondent is a woman (1) or not (0)	Binary
Homeowner	Indicates if respondent is a homeowner (1) or not (0)	Binary
Carowner	Indicates if respondent is a car owner (1) or not (0)	Binary
Respcity	Indicates corresponding city	Categorical
Job	Indicates if respondent is working (1) or not (0)	Binary

assess the significance of parameters and their alignment with theoretical expectations. Modellers typically consider these results before examining additional goodness-of-fit metrics. Participants can then select from a range of metrics for the estimated model and choose to analyse them. These include: the number of parameters estimated in the model; the sample size used in the estimation process; the null log-likelihood; the initial log-likelihood; the final log-likelihood; the likelihood ratio test against the null model; the ρ^2 and adjusted ρ^2 against the null model; the Akaike Information Criterion (AIC); the Bayesian Information Criterion (BIC); the final gradient norm; the time taken for model estimation; and the willingness-to-pay estimates for attributes. Finally, participants can compare models in two ways: by making direct comparisons between models in terms of parameters and goodness-of-fit indicators, or by displaying elbow graphs to visualise latent class model metrics.

Reporting phase: At the end of the simulated research process, choice modellers are asked to report their findings to policymakers, as shown on the respective page (Fig. 12). To facilitate this, they can review the estimated models along with summaries of their results in order to select the most appropriate ones. In addition, participants are required to submit a short written report (several sentences), in which they detail the main findings and interpret the modelling results to address the objective set at the beginning of the game.

3.4. Stated preference choice dataset

The above discussion presents a general overview of the DCM-SG, independently of the dataset used. For our specific application, and in order to represent a research scenario that reflects what practitioners encounter in their real-world work, we use a modified raw stated preference dataset collected by Liebe et al. (2023), which aimed to analyse residential location choice. The dataset consists of three unlabelled alternatives (A, B, and C), each defined by six attributes: distance to the grocery store, distance to transportation, distance to the city centre, street traffic noise, green areas in the residential area, and monthly housing cost variation. This dataset contains 2430 individuals, each facing four choice tasks, resulting in 9720 observations. It also includes socio-demographic variables such as age, gender, home ownership, car ownership, place of residence, and employment status. A summary of the data is shown in Table 1.

4. Serious game data

4.1. Gameplay data

During gameplay in the DCM-SG, two types of data are recorded to enable behavioural analysis of decision-making in choice modellers' workflows. Firstly, the game stores each participant's identifier, timestamp of actions, and any task performed over the course of the research phases, collected in situ and in real time (telemetry). For example, there are 15 descriptive analysis in-game tools, including statistical analysis, missing value handling, graph creation, and database sorting. In the model specification phase, there are 34 interactive game devices (such as buttons, drop-down menus, and checklists) for specifying MNL, MMNL, and LC models. Any use of these is stored in the database. During the outcome interpretation phase, there are 18 options for comparing modelling results. Finally, the model(s) selected and their main findings are stored. Table 2 displays the variables and their descriptions, which allow us to track participants' interactions with the game.

Table 2
Variable descriptions and types.

Variable	Description
timestamp	Time at which participants performed any task
user_id	Participant ID
task_id	Task ID performed in DA or OI
model_id	Model ID (to identify task performed in model specification)
model	1 for MNL, 2 for MMNL, 3 for LC
ASC	1 to include alternative specific constants (0 otherwise)
att_i	1 to include attribute i (0 otherwise)
s_i	1 to indicate that attribute i is alternative-specific (0 otherwise)
t_i	1 for applying linear transformation to attribute i, 2 for box-log, 3 for logarithmic
int_i	indicates whether attribute i interacts with a single socio-demographic variable (only once per attribute is allowed at a time): none (= 0), woman (= 1), age (= 2), location of residence (= 3), home owner (= 4), car owner (= 5)
$dist_i$	0 to indicate that attribute i is fixed at the population level, 1 follows a normal distribution, 2 follows a lognormal distribution
n_class	indicates the number of classes
$covariates_j$	1 to indicate that covariate j is included in the membership class function for latent class (0 otherwise)
r_models	identify reported models by the participants
reporting	discussion and findings of selected models

Secondly, both qualitative and quantitative data are collected at the end of the game to characterise participants in terms of their background and expertise. Thus, we characterised each participant using a unique `user_id`, and collected personal information such as `gender` (participant's gender), `age_dcm` (years involved in choice modelling), `main_field` (primary field the participant is working in), and expertise (self-assessment of their knowledge in choice modelling). We also asked whether the participant had been a teacher or a student in a choice modelling course; their most commonly used programming language or software; their predominant modelling approach used in applications; and their `h_index` (as a proxy for scholarly impact). Finally, we recorded the initial and final timestamps of when participants began playing the game (`init_time`) and submitted their report (`end_time`).

4.2. Participants

The DCM-SG was administered in person to attendees at two conferences — the International Choice Modelling Conference and the International Conference on Travel Behaviour Research — and was also distributed online to researchers and practitioners known to work with or on choice modelling. Participation was voluntary. Before taking part, all individuals were informed that the game was part of a study on decision-making in discrete choice modelling. Additionally, all participants read and accepted the informed consent form as part of the initial step in the game. Data were collected and analysed anonymously.

A total of 40 participants were involved; 38 of them reported their models and completed their analyses. For the remaining participants, we inferred their reported models as those that were specified towards the end of their modelling processes and demonstrated high performance. Although the distribution of experience in choice modelling was varied, most participants had more than five years of experience (10% had less than one year, 32% had one to five years, 45% had five to ten years, and 13% had more than ten years). Moreover, 85% of participants had taken a choice modelling course, while 58% had experience as a teacher, teaching assistant, or lecturer. These values indicate that most participants were familiar with more than just the basics of choice modelling. Many could be considered experts, combining formal training, teaching experience, and several years of involvement in the modelling process. Most participants focused on transportation (70%), while others worked in economics (10%), urban planning (8%), environmental valuation (5%), and other fields (7%). In terms of self-assessed expertise, 40% rated themselves at a medium level, and 25% at medium–high, with smaller percentages at lower levels. Scholarly impact scores ranged from 0.0 to 20.0, with 8.0 being the most common score (20%). The gender distribution was predominantly male (68%), with 32% female or undisclosed. In addition, participants reported their primary modelling approaches: the MNL and MMNL models were the most commonly used, followed by LC models. Other approaches — such as Integrated Choice and Latent Variable, Multiple Discrete–Continuous Extreme Value, and Ordered Logit models — were reported less frequently. Finally, although we did not directly measure task engagement, we found that participants specified multiple models, interacted extensively with in-game tools, and occasionally requested extended time. Their responses to post-game questions about realism and completeness also indicated a high level of engagement, though some noted that certain modelling options limited their ability to explore more complex models.

5. Data analysis

In this section, we focus on analysing choice modellers' decision-making behaviour as captured through the DCM-SG, utilising a behavioural observation framework (Bakeman and Gottman, 1997). This approach allows us not only to discuss the common

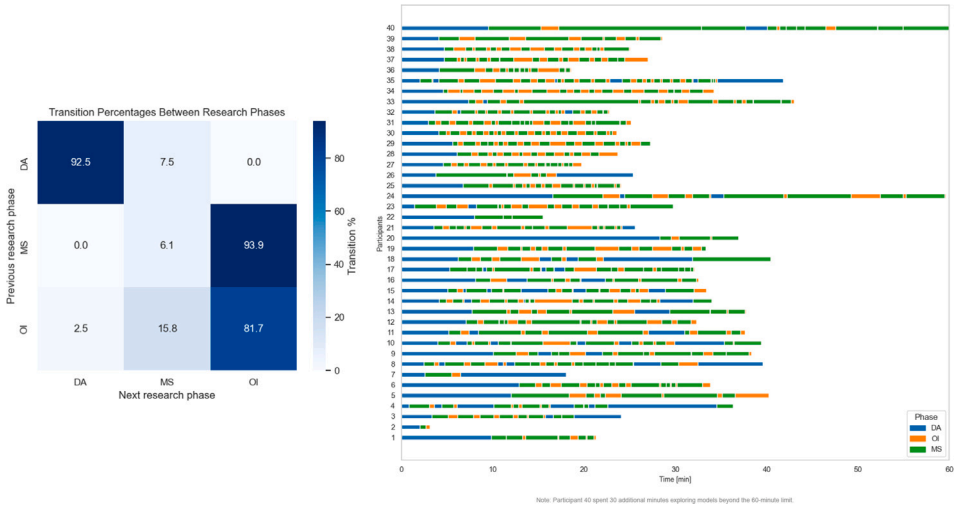


Fig. 2. Workflow transitions and time allocation in choice modelling phases.

methods and analyses employed by participants but also to identify sequential patterns in their model specification and refinement processes. We then clustered participants to explore how different factors, such as in-game usage, transitions between research phases, and modelling patterns, affect their reported modelling results.

5.1. *Workflows and time spent*

We begin our analysis by examining choice modellers' workflows, focusing on the transitions between research phases and the time allocated to each, as shown in Fig. 2. The high intra-phase retention rates, such as 93% in DA and 82% in OI, indicate that modellers typically use all available in-game tools within a phase before gradually moving on to model specification. Also, intra-phase transitions within the MS phase (6%) suggest that modellers often choose to specify a new model without considering other modelling results beyond the estimated parameters. This behaviour may reflect a perceived misspecification or unexpected estimation results, leading to a revision of the specification. Notably, few modellers (only 2.5% of transitions) return to reanalyse the data after proceeding to model specification and outcome interpretation, which suggests that iteration is concentrated mainly within these two phases.

On the right-hand side, we display the timeline of each choice modeller, representing their progress through the research phases. The graph shows that DA generally dominates the early workflows, with modellers typically exploring, visualising, and cleaning data before moving on to more complex phases. A significant amount of time is then spent on MS, reflecting efforts to refine models and incorporate assumptions based on analyses conducted during DA and OI. Interestingly, there are instances where modellers briefly return to earlier phases, as discussed previously. This highlights the iterative nature of the process, where modellers may revisit the DA phase to revise model hypotheses or conduct further descriptive analysis based on new insights gained from later phases.

5.2. *In-game tools analysis*

Table 3 provides a summary of the in-game tools used by participants across all research phases. The mean usage shown in the table is calculated only for those participants who used the respective tool at least once. The interaction percentage represents the relative frequency of use within each specific research phase, rather than across the entire modelling process. This approach allows for a comparison of tool engagement in the context of each individual phase.

In the descriptive analysis phase, participants demonstrated a strong preference for data visualisation tools and statistical descriptions, which can facilitate a deeper understanding of the data prior to model specification. In particular, histograms accounted for 19.17% of interactions in this phase, followed by data dictionary visualisation (13.27%), box plots and pie charts (both 7.11%), and correlation matrix plot (6.76%). Although most participants viewed the main data statistics and the percentage of choice between alternatives to evaluate the data distribution among choice-makers, only approximately 73% of the modellers (29 out of 40) deleted missing values, while 33% (13 out of 40) replaced them before moving to model specification. Since both actions could be performed, the 13 participants who replaced missing values may overlap with those who deleted them. Thus, while the use of visualisation tools appears to support the formulation of initial modelling hypotheses, the exact methodology adopted for handling missing values and preprocessing the dataset, even in cases where there is a small amount of missing values, remains unclear.

In the model specification phase, participants attempt to specify the three model families available in the SG: MNL, MMNL, and LC models. While many of these models were successfully estimated, some specifications failed to converge due to issues such as misspecification (e.g., incorrect functional forms or the inclusion of socio-demographic variables that do not vary across alternatives) or because the raw choice dataset lacked sufficient variability to identify certain parameters. As shown in Table 3, we subsequently analysed each model family to identify the most common specifications.

Table 3
Summary of in-game tools used across research phases.

Tools	Users	Mean (SD)	Interaction [%]
View summary statistics	38	1.84 (0.97)	6.07
View data dictionary	40	3.92 (3.12)	13.27
Check missing data	39	1.56 (0.75)	5.29
View first 5 rows of data	37	2.00 (2.39)	6.42
View percentage of choices	36	1.33 (0.89)	4.16
View choice task example	36	1.47 (1.00)	4.60
View histogram	34	6.50 (4.98)	19.17
Delete missing values	29	1.21 (0.41)	3.04
View boxplot	26	3.15 (2.17)	7.11
Sort dataset by variable	22	2.14 (1.46)	4.08
View correlation	25	3.12 (1.64)	6.76
View two-variables scatter plot	17	3.24 (1.92)	4.77
Replace missing values	13	1.23 (0.44)	1.39
View pie chart	13	6.31 (7.90)	7.11
View bar chart	19	3.68 (3.00)	6.07
Multinomial logit model	40	5.92 (3.04)	51.68
Latent class model (2 classes)	31	2.45 (1.48)	17.00
Latent class model (3 classes)	26	1.73 (1.08)	10.07
Mixed logit model	27	2.26 (1.51)	13.64
Model misspecification	16	2.12 (1.85)	7.61
View final log-likelihood	33	6.58 (4.24)	15.16
View initial log-likelihood	27	2.44 (1.72)	4.61
Calculate Willingness-to-Pay	33	6.58 (4.24)	15.16
Model comparison	33	6.21 (4.54)	14.33
View number of parameters	24	3.12 (3.05)	5.24
View number of individuals	17	2.00 (1.41)	2.38
View log-likelihood at equal shares	21	2.43 (1.83)	3.56
View ρ^2	23	5.22 (4.48)	8.39
View adj. ρ^2	25	5.67 (4.86)	10.06
View number of CPU cores	17	2.18 (1.55)	2.59
View number of data rows	17	2.00 (1.41)	2.38
View number of outputs	15	2.27 (1.53)	2.38
View Akaike Information Criterion	20	6.25 (5.09)	8.74
View Bayesian Information Criterion	20	6.20 (5.07)	8.67
View time taken for estimation	10	1.20 (0.42)	0.84

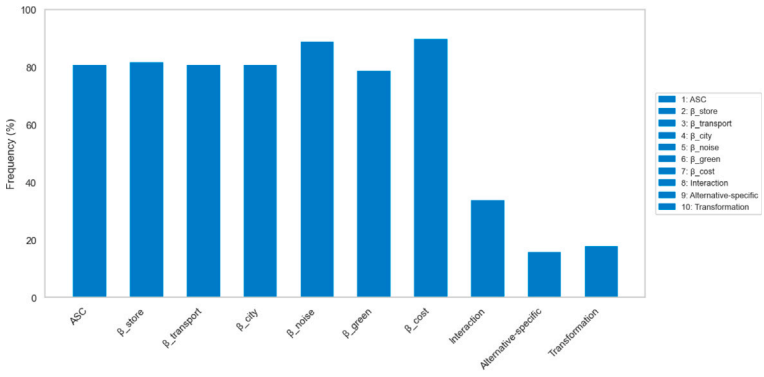


Fig. 3. Multinomial logit model specifications among participants.

Fig. 3 highlights trends within the specification of MNL models. Despite working with a stated preference database with unlabelled alternatives, approximately 80% of the models included alternative-specific constants (ASCs) and most did not incorporate all attributes. This suggests that choice modellers were attentive to possible differences in baseline utilities. As their estimated model results show, ASCs were often statistically significant and improved model performance, which possibly reflects efforts to account for design effects or lexicographic behaviours in the decision-makers' choice process. In general, attribute effects were included as linear-additive, the Store attribute in 82% of the models, Transport in 82%, City in 82%, Noise in 88%, Green in 79%, and Cost in 89%. Transformations of attributes were limited, with only 18% of the specifications including logarithmic or Box–Cox transformations. Furthermore, while the majority of attributes were treated as generic, which is consistent with the context of unlabelled alternatives, approximately 16% of the cases were tested with alternative-specific taste parameters. Regarding model interactions, 34% of the specified models include at least one attribute-covariate interaction. The most frequently observed were between Transport, City, Noise, and Green with the respondent's Age, as well as interactions between Store and gender (Woman),

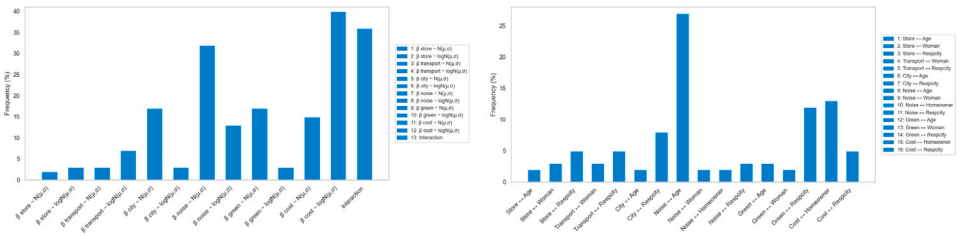


Fig. 4. Mixed multinomial logit model specifications among participants.

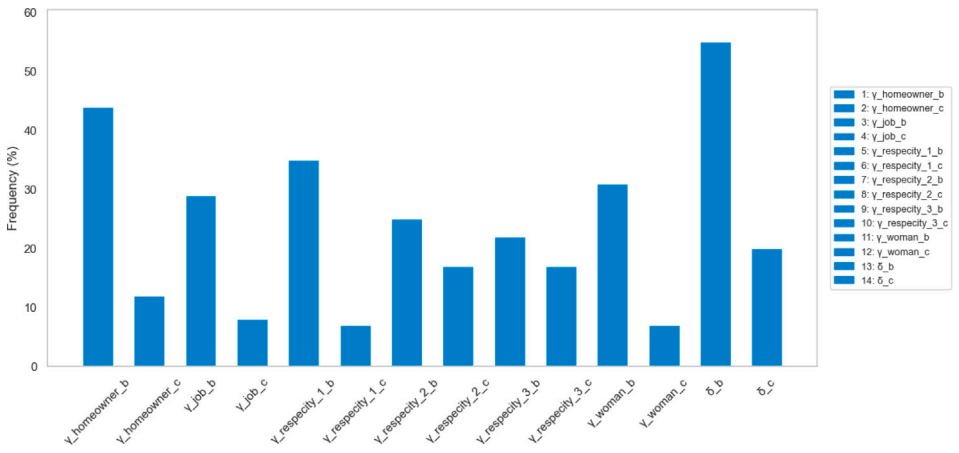


Fig. 5. Latent class model specifications among participants.

and between Cost and home ownership status (Homeowner). However, there was a limited exploration of interactions related to the residence location (respcity), even though the data were collected in four different cities. This evidence reveals a clear preference for linear parameter specifications, with a tendency to test for attribute exclusion and, less frequently, the integration of socio-demographic characteristics. This pattern may reflect limitations in the model specification phase of the game, which did not allow for modelling categorical attributes as dummy variables. This limitation may have prevented modellers from exploring other ways they could have considered. Furthermore, the limited exploration of interactions may be due to time constraints, which prevented modellers from fully investigating hypotheses related to preference variations of decision-makers across cities.

Fig. 4 shows the MMNL model specifications, in which all attributes were considered linear and generic by default. This restriction imposed by the DCM-SG does not appear to have affected participants' modelling decisions, as it is consistent with the specifications observed in the MNL models. In terms of parameter distribution, 44.3% of specifications considered a single distributed parameter, with 55.5% of them following a normal distribution and the remainder being log-normally distributed. When two random parameters were included, 44.9% were modelled as normally distributed, 35.6% as log-normally distributed, and 19.5% as a combination of both distributions. Notably, Noise and Cost were the most frequently specified attributes with normal and log-normal distributions, respectively. Moreover, 63.9% of the models did not include interactions with socio-demographic variables, with city being the least frequently considered in such interactions. When interactions were included, the most common were with age, followed by residence location and gender, though their frequency remained low. These findings evidence an approach focused on capturing taste variations for noise and cost, along with an exploration of the observed heterogeneity. However, considering a normal distribution for the cost attribute has important implications, since it always results in undefined moments for the willingness-to-pay values for noise, a problem that can be addressed by using log-normal distributions, which restrict the values to a non-negative space and provide greater consistency in the economic interpretation of the results.

Fig. 5 shows the latent class model specifications. As in the MMNL case, all attributes were specified as linear by default, with the same functional form applied across all classes. In the modellers' decisions, 55.3% of the LC specifications involved latent class models with two classes, while the remaining models utilised three classes. This suggests that many modellers preferred to work with less complex models that may be easier to interpret and estimate. Moreover, only six modellers successfully specified three-class latent models despite several trials. This suggests that due to the early issues of parameter non-identification, modellers tended to focus on models with fewer classes or other model families. In terms of the membership function, the inclusion of covariates such as `Homeowner`, `Woman`, `Job`, and `Respcity` ranged from 8% to 44%, which demonstrates an attempt to capture socio-demographic heterogeneity in the membership class function. However, the relatively low usage rates of `Respcity` in three-class models indicate that modellers may not have fully explored the potential influence of location-based heterogeneity. This aligns with observations in the MNL and MMNL models, where interactions with residential locations were also limited.

In the outcome interpretation phase, when modellers estimated a model, they were immediately provided with a standard results table generated by the DCM-SG, which included the parameter names, estimated values, robust standard errors, t-tests against zero,

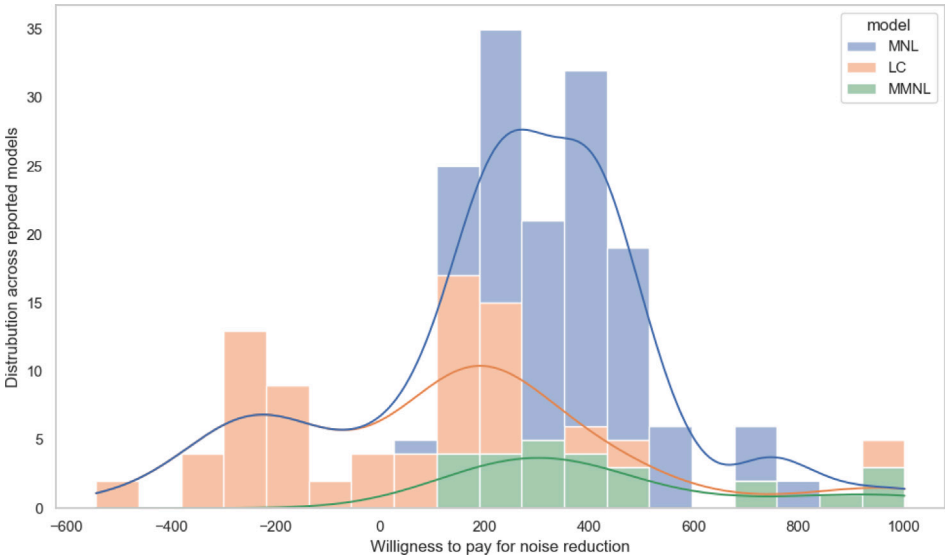


Fig. 6. Reported Willingness-to-Pay values for noise reduction across final model families.

and corresponding *p*-values. Analysis of the tools used for interpreting outcomes revealed that modellers focused mainly on the calculation of willingness-to-pay (15.16% of the interactions) and on comparison between models (14.33% of the total interactions). Similarly, the observation of log-likelihood (13.07%) was also highly reviewed, which indicates the importance modellers attach to the internal consistency of the model and its predictive capability. However, modellers paid limited attention to other metrics such as the (adjusted) ρ^2, the Akaike Information Criterion, and the Bayesian Information Criterion, each used in approximately 9% of interactions. These preferences suggest that model refinement was guided primarily by goodness-of-fit and economic interpretation of the parameters, using the previous model as a benchmark for comparison, but also involved neglecting metrics that assess the trade-off between model fit and complexity—an essential consideration for selecting well-performing and parsimonious models.

Finally, in the reporting phase, Fig. 6 shows the willingness-to-pay reported by modellers towards the conclusion of the simulated research, which reveals important implications for informing policymakers. On the one hand, MNL and MMNL models consistently produced positive WTP estimates for noise reduction, indicating a generalised preference among case study decision-makers. Although LC models (LC2 and LC3) uncover significant preference heterogeneity, with some classes showing negative WTP values, MNL was the most commonly reported model type (55% of final reports), followed by LC models (32.5%) and finally MMNL (12.5%). This heterogeneity in results not only highlights the range of acceptable outcomes found by modellers but also reflects the complexity of their decision-making processes in finalising model reports. Furthermore, this clear distinction in reported WTP values between models serves as a reminder to choice modellers of the importance of considering multiple modelling approaches to fully capture both observed and unobserved heterogeneity in preferences.

5.3. Analysis of model specification workflows

To gain deeper insights into the modelling workflows employed by participants during the model specification phase, we display the temporal progression of models in Fig. 7. This graph provides an aggregated view of how modellers transitioned across different model families, such as MNL, LC, and MMNL, as well as cases of Misspecification (Miss), and the point at which they reported their final model (R). The diagram shows the evolution of the estimated models and their log-likelihood values, starting from their initial specification to the reported one.

As can be seen, during the initial stage of the modelling process, most participants (38 out of 40) began by specifying MNL models, which varied in terms of the attributes included, the transformations applied, and the interactions specified with socio-demographic variables. Notably, only four participants started with a fully linear specification that included all attributes. This may seem counter-intuitive, as one might expect modellers to begin by including all available variables from the stated choice tasks to capture primary and linear effects on choices. This behaviour could reflect the influence of prior beliefs about data complexity or previous modelling experience, which may have shaped their initial hypotheses. As participants progressed, they tested more complex functional forms within the MNL family. Some of these specifications resulted in misspecification errors, potentially due to the inclusion of overly intricate interactions with socio-demographic variables or non-linear transformations of attributes. These attempts to account for observed heterogeneity may have led to parameter non-identification, demonstrating the trade-offs between model complexity and statistical feasibility.

In later iterations, there was a shift towards more complex models, with participants gradually specifying LC and MMNL models. This evolution suggests a learning trajectory, where initial explorations with simpler MNL models laid the groundwork for the adoption of more flexible models capable of accounting for deterministic and random taste heterogeneity. This pattern aligns with standard modelling practice, in which simpler models are frequently estimated as a benchmark before progressing to more advanced approaches.

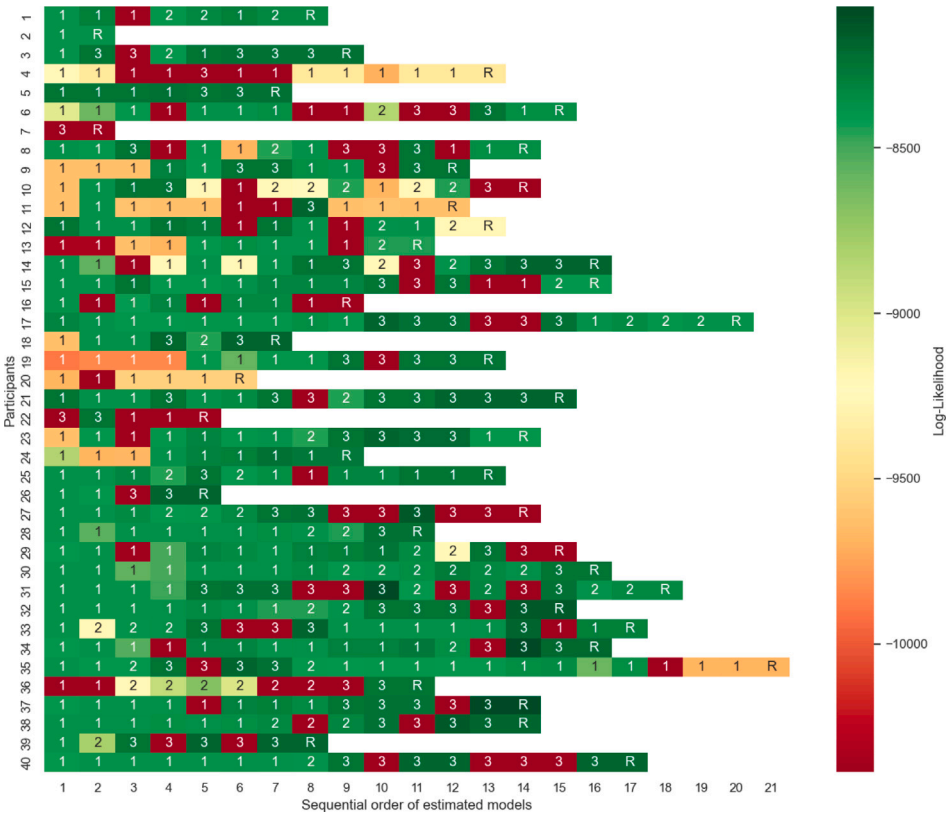


Fig. 7. Workflow transitions and time allocation in choice modelling phases.

5.4. Analysis of workflow differences

To evaluate the impact of participants' workflows on the improvement of choice modelling outcomes, we conducted an analysis of modelling patterns. Using the cSPADE algorithm, a data mining method designed to identify frequent patterns within temporal sequences (Zaki, 2001; Maimon and Rokach, 2005), we examined the most common patterns related to in-game tool usage, transitions between research phases, shifts between different model families, and specification strategies. This approach allows us to move beyond static summaries and explore the temporal dynamics of how participants interacted with the DCM-SG environment.

Firstly, participants were classified into two groups based on changes in goodness-of-fit metrics, such as log-likelihood, AIC, BIC, ρ^2, and adjusted ρ^2, from the initial model to the final reported model. The first group comprised 30 participants who demonstrated improvements across these metrics, suggesting a goal-oriented learning process aimed at progressively capturing variability in the choices through more complex model specifications. The second group consisted of 10 participants whose final models did not outperform their initial specifications, which may indicate challenges in refining their modelling approaches or limitations in their ability to adapt their modelling approach to better fit the data. Secondly, the most frequent patterns of in-game tool usage, transitions between research phases, and shifts between model types were found considering the cSPADE algorithm with minimum support thresholds of 70%. Thus, only sequential patterns that appeared in at least 70% of the observed workflows were included in the analysis. This threshold was chosen to reduce noisy patterns and extract the most common behavioural workflows within our small-to-medium sample, as suggested by Zhang and Paquette (2023), Kang et al. (2017). Finally, we calculated the frequency of each pattern for every choice modeller and compared the two groups using an independent samples t-test.

Table 4 shows statistical differences between the two groups of participants in terms of their use of in-game tools, transitions between research phases, and model specification workflow patterns. Participants who used tools such as 'View summary statistics', 'View bar plot', and 'View correlations', or who spent more time visualising histograms or specifying MNL models, were more likely to report improved modelling metrics compared to their initial estimations. Moreover, specific workflow sequences, such as moving from 'Handling missing values' to 'Viewing summary statistics' and then to 'Viewing first five rows', as well as revisiting the descriptive analysis phase after interpreting their modelling outcomes (OI → DA), were more common among modellers who refined their specifications over time. These patterns of behaviour suggest a higher involvement in data analysis and development of initial hypotheses, which contributed to improved models.

Another important distinction emerged in how participants analysed modelling outcomes. Those who adopted a comprehensive approach, considering not only parameter estimate values and their standard errors, but also model fit metrics such as ρ^2, $\bar{\rho}^2$, AIC, and BIC, were more likely to improve their modelling outcomes. In addition, the workflow 'Calculate WTP → Model comparison

Table 4
Difference across frequent patterns (p-value < 0.05).

In-game tools and workflows	t-statistic	p-value
Total uses: view summary statistics	−2.82	0.055
Total uses: view bar plot	−2.20	0.042
Total uses: view correlations	−2.69	0.013
Total uses: model comparison with previous one	−2.20	0.046
Total uses: view Akaike Information Criterion	−5.38	0.001
Total uses: handling missing data	−3.21	0.002
Total uses: replacing missing values	2.94	0.011
Time spent on visualising histograms	−2.82	0.027
Time spent on specifying MNL	−2.64	0.013
Calculate WTP → MNL	−2.75	0.009
$\bar{\rho}^2$ → Calculate WTP	−2.73	0.011
MNL → Log-Likelihood	−3.14	0.003
Calculate WTP → Model comparison with previous one	−3.03	0.005
BIC → calculate WTP	−2.81	0.009
Miss → Miss	−3.10	0.004
MMNL → LC	−2.69	0.013
DA →DA → MS → OI → OI → MS	−3.53	0.001
MS → OI → OI → OI → OI	−2.96	0.006
OI → OI → OI → OI → MS	−3.18	0.003
OI → OI → DA → DA → MS	−2.26	0.031

with previous model' reflects a deliberate focus on both economic interpretation and comparative model evaluation, often leading to better model refinement. Similarly, participants who engaged in sequences such as 'AIC → Calculate WTP' or 'BIC → Calculate WTP' demonstrated significantly better model improvements than those who did not. These patterns suggest that integrating statistical diagnostics with economic reasoning supports more informed and effective model development.

In terms of transitions between modelling phases, significant differences in workflows appeared when several DA and OI phase tools were used before specifying a model. For instance, sequences such as 'DA → DA → MS → OI → OI → MS' and 'DA → MS → OI → OI → MS' suggest that revisiting data analysis both before and after initial model specification further improved model refinement. Similarly, extended use of OI tools prior to re-specification, such as 'MS → OI → OI → OI → OI' and 'OI → OI → OI → OI → MS', was associated with improved model fit, whereas those who relied solely on estimation tables showed minimal or no improvement. Moreover, the pattern 'OI → OI → DA → DA → MS' further supports the idea that returning to data exploration after interpreting results can inform initial modelling hypotheses and ultimately lead to improved model performance. These results suggest that multiple specifications without adequate reflection and analysis of previous modelling results may lead to inefficient search processes, producing models that ultimately fail to outperform the initial one.

Finally, we found that participants who achieved improvements in model fit also tended to exhibit a higher frequency of model misspecification transitions (Miss → Miss) as well as transitions from MMNL to LC models. This behaviour suggests a continued effort to capture more complex functional forms and reflects an iterative learning process throughout the modelling. As shown in Fig. 7, the workflows of participants 4, 17, and 36, among others, demonstrate that repeated misspecifications were often followed by model specifications achieving better performance within the same model family.

Overall, the results indicate that participants who engaged in thorough data analysis, revisited earlier research phases to refine their modelling assumptions, and evaluated goodness-of-fit metrics were more successful in reporting models with an improved balance between fit and parsimony. These findings tangibly reflect the nature of discrete choice modelling as an intrinsically iterative, hypothesis-driven, and feedback-guided process, which relies not only on model fit metrics but also on alignment with expected behavioural realism.

6. Conclusions

Our study provides a twofold contribution to the choice modelling field. First, it introduces serious games as a methodological innovation to capture workflows of choice modellers and demonstrates how serious game data can effectively be analysed to better understand the choice modelling process. The Discrete Choice Modelling Serious Game (DCM-SG) provides an online environment that mimics the real-world research phases, enabling modellers to apply their knowledge while we track their actions. The code used to implement the DCM-SG will be publicly accessible via a GitHub repository, thereby facilitating further research using this tool within the community.[3]

Second, our study provides new substantive insights into the practices of choice modellers. Firstly, we find strong evidence of the iterative nature of the choice modelling process, with choice modellers moving back and forth between modelling phases, such as descriptive analysis (DA), model specification (MS), outcome interpretation (OI), and reporting. For instance, we have observed that — after the first descriptive analysis — most participants start with specifying Multinomial Logit (MNL) models, after which they either return to the descriptive analysis phase or move forward to more advanced models, such as Mixed Multinomial

[3] GitHub repository: https://github.com/TUD-CityAI-Lab/DCM-SG.

Logit (MMNL) and Latent Class (LC) models. Secondly, we find extensive support for the notion that modelling practices are heterogeneous. Specifically, we see that while data visualisation and statistical descriptions are commonly used, there is no clear approach to handling missing values. Thirdly, we observe that participants favoured simpler models despite having complex families available in the game. Finally, our results reveal that workflows, in-game tools usage, and model specification strategies significantly impact choice modelling outcomes. Participants who engaged in comprehensive data exploration, iterative comparisons, and made systematic use of econometric tools tended to improve goodness-of-fit and parsimony. Conversely, those who relied more heavily on limited metrics without exploring broader aspects of the model struggled to improve upon their initial specifications.

Our findings also invite a broader reflection on hypothesis formulation in choice modelling. As illustrated in our theoretical framework (Section Section 2), the process of specifying initial model assumptions is typically guided by exploratory analyses and theoretical reasoning, yet rarely documented as formal hypotheses. While this flexible, iterative approach allows choice modellers to adapt specifications based on insights from the data, it contrasts with emerging practices in the social sciences, such as pre-registration, aimed at improving transparency and reproducibility. Although pre-registration could be a problem given the exploratory nature of the model specification, the DCM-SG provides a structured environment in which initial hypotheses, modelling workflows, and decision points could, in future applications, be documented and studied systematically. This opens up opportunities for further methodological innovation around transparency, accountability, and reproducibility in choice modelling.

Our study has several limitations that provide avenues for future research. The current implementation of the DCM-SG, specifically in the model specification phase, imposes constraints that do not totally replicate real-world modelling scenarios. While the game is designed to support iterative exploration, the set of tools and model options remains limited compared to what modellers may use in practice. More importantly, the available in-game tools, interface structure, and modelling pathways were informed by the proposed framework grounded in the literature, as well as by the designers' own understanding and experience of the choice modelling process. This may influence participants' behaviour, strategies, and decision-making processes they might otherwise pursue. Additionally, the time constraints assigned to the game sessions might have pushed participants to use simpler models that they would have used in real life (63% of participants selected the MNL model as their reported model). After all, in the real world, modellers typically have more time and flexibility to explore complex model families and refine specifications. Moreover, contextual factor during data collection, such as the limited time available to participants, or the stated goal of the game, may have encouraged participants to prioritise completing the task over engaging in extended model exploration. While we did not include an explicit self-report measure of engagement, participants' behaviour (e.g., prolonged interaction with the platform, iterative specification of meaningful model, and unsolicited comments on the game's realism) indicates sustained and active engagement. Finally, participants' awareness that they were being monitored may also have introduced a bias, leading them to display desirable behaviours rather than their own modelling practices. For instance, while we observed some modellers revisiting earlier phases (e.g., returning to DA), it remains unclear whether such behaviour mirrors real-world practice or was influenced by the experimental setup. To mitigate this, all responses were collected anonymously, and participants were encouraged to engage freely with the platform.

To overcome these limitations, future research should involve larger samples with more diverse backgrounds, extend the duration of game sessions to allow deeper engagement with advanced models, and improve in-game tools to enable more complex approaches in the model specification phase.

CRediT authorship contribution statement

Gabriel Nova: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sander van Cranenburgh:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Stephane Hess:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that there are no conflicts of interest related to this study.

Acknowledgements

We would like to express our sincerest gratitude to all participants in our serious games. Their willingness, dedication, and active involvement in the game are greatly appreciated. In particular, we acknowledge the following individuals for their significant contributions and valuable insights, which greatly enhanced the research process: Bartosz Bursa, Victor Cantillo, Santiago Cardona, Laurent Cazor, Roel Faber, Sayed Faruque, Francisco Garrido-Valenzuela, Thomas Hancock, Bastian Henriquez-Jara, Chengxi Liu, Gregory Macfarlane, Evripidis Magkos, Petr Mariel, David Palma, Pablo Reyes, Jaime Soza-Parra, Melvin Wong, Amir Ghorbani, Baiba Pudāne, Peter King, Edward JD Webb, Nejc Geržinič, Rodrigo Tapia, Marco Kouwenhoven, and Sander Boxebeld. We gratefully acknowledge Ulf Liebe for making the dataset available and for providing comments on the design of the serious game. We also thank Michiel Bliemer for his early feedback on the game framework. Stephane Hess acknowledges support from the European Research Council through the Advanced Grant 101020940-SYNERGY.

Appendix

See Figs. 8–12.

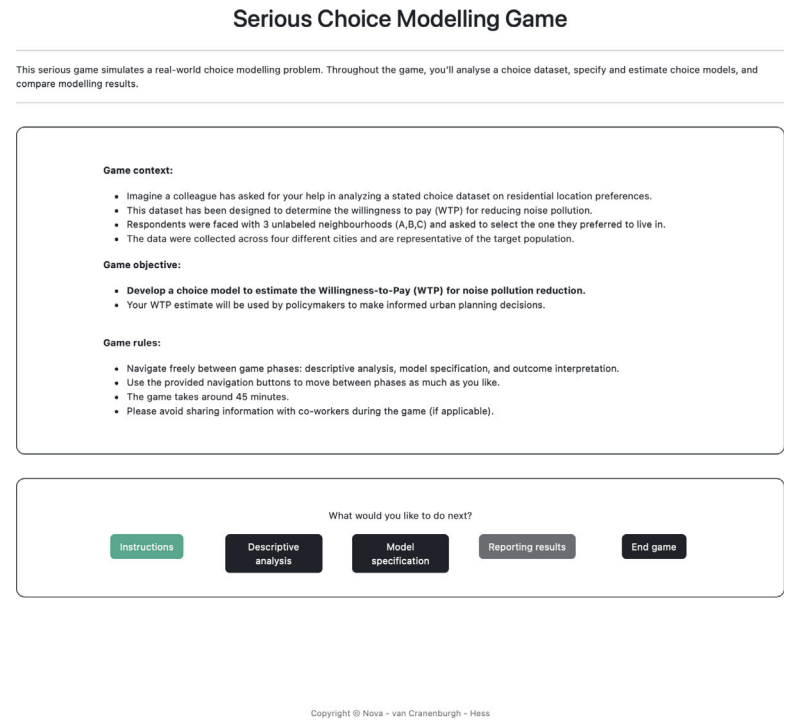


Fig. 8. Screenshot of the DCM-SG instruction page.

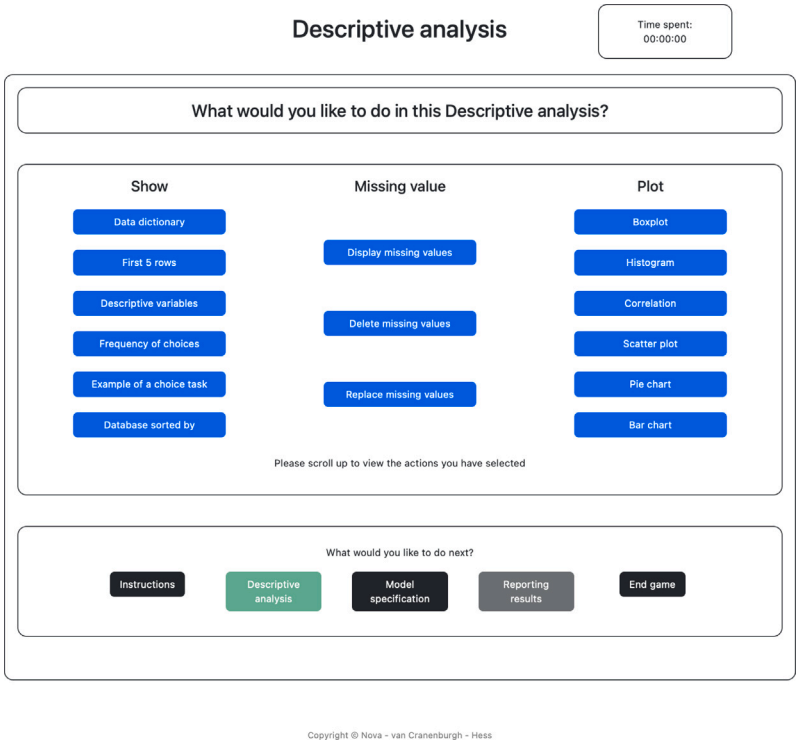


Fig. 9. Screenshot of the DCM-SG descriptive analysis page.

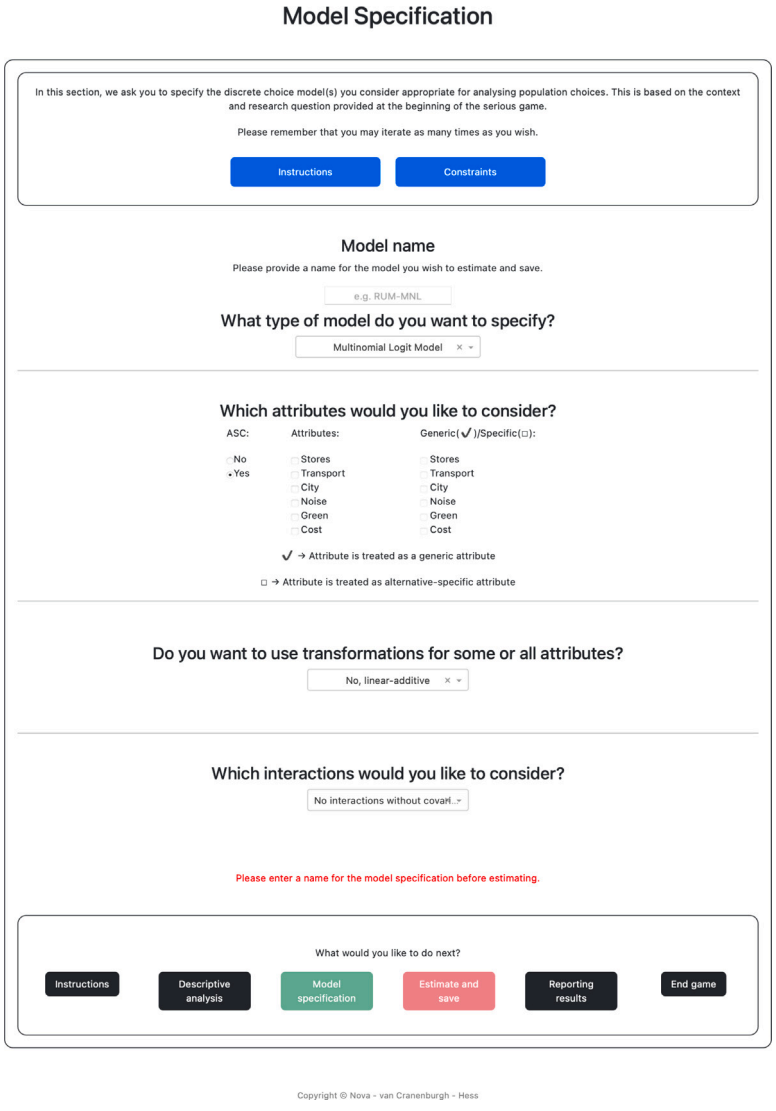
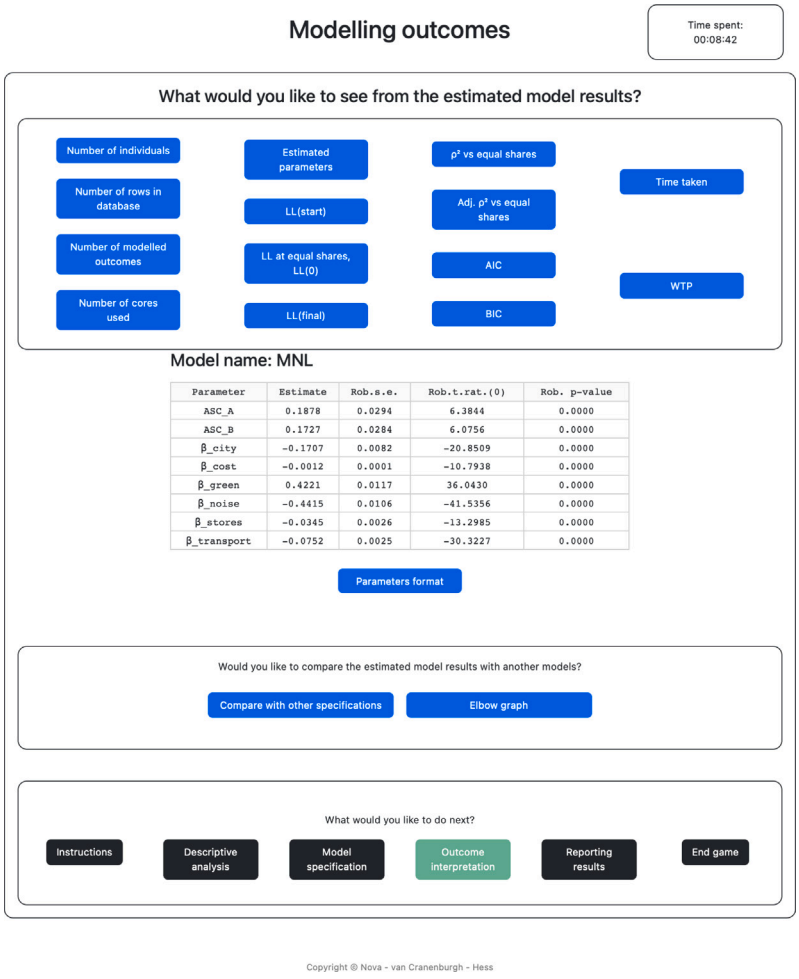


Fig. 10. Screenshot of the DCM-SG model specification page.



Parameter	Estimate	Rob.s.e.	Rob.t.rat.(0)	Rob. p-value
ASC_A	0.1878	0.0294	6.3844	0.0000
ASC_B	0.1727	0.0284	6.0756	0.0000
β_city	-0.1707	0.0082	-20.8509	0.0000
β_cost	-0.0012	0.0001	-10.7938	0.0000
β_green	0.4221	0.0117	36.0430	0.0000
β_noise	-0.4415	0.0106	-41.5356	0.0000
β_stores	-0.0345	0.0026	-13.2985	0.0000
β_transport	-0.0752	0.0025	-30.3227	0.0000

Fig. 11. Screenshot of the DCM-SG report page.

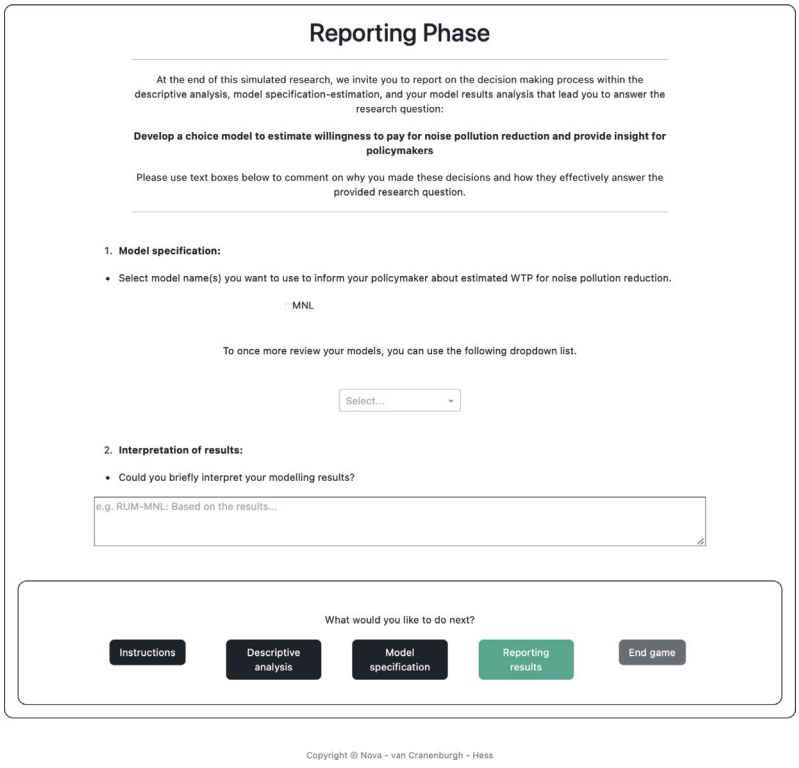


Fig. 12. Screenshot of the DCM-SG report page.

Data availability

The code used to implement the DCM-SG will be publicly accessible via a GitHub repository.

References

Arriagada, J., Munizaga, M.A., Guevara, C.A., Prato, C., 2022. Unveiling route choice strategy heterogeneity from smart card data in a large-scale public transport network. Transp. Res. Part C: Emerg. Technol. 134, 103467.

Axhausen, K.W., 2024. Self-tracing and reporting: State-of-the-art in the capture of revealed behaviour. In: Handbook of Choice Modelling. Edward Elgar Publishing, pp. 147–171.

Bakeman, R., Gottman, J.M., 1997. Observing Interaction: An Introduction to Sequential Analysis. Cambridge University Press.

Barría, C., Guevara, C.A., Jimenez-Molina, A., Seriani, S., 2023. Relating emotions, psychophysiological indicators and context in public transport trips: Case study and a joint framework for data collection and analysis. Transp. Res. Part F: Traffic Psychol. Behav. 95, 418–431.

Beeramoole, P.B., Arteaga, C., Pinz, A., Haque, M.M., Paz, A., 2023. Extensive hypothesis testing for estimation of mixed-Logit models. J. Choice Model. 47, 100409.

Ben-Akiva, M., Bierlaire, M., 2003. Discrete choice models with applications to departure time and route choice. In: Handbook of Transportation Science. Springer, pp. 7–37.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand, vol. 9, MIT Press.

Bierlaire, M., 2003. BIOGEME: A free package for the estimation of discrete choice models. In: Swiss Transport Research Conference.

Bierlaire, M., Frejinger, E., 2008. Route choice modeling with network-free data. Transp. Res. Part C: Emerg. Technol. 16 (2), 187–198.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., et al., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582 (7810), 84–88.

Caussade, S., de Dios Ortúzar, J., Rizzi, L.I., Hensher, D.A., 2005. Assessing the influence of design dimensions on stated choice experiment estimates. Transp. Res. Part B: Methodol. 39 (7), 621–640.

Chung, G., 2014. Toward the relational management of educational measurement data. Teach. Coll. Rec. 116 (11), 1–16.

Corti, K., 2006. Games-based learning; a serious business application. Inf. de PixelLearning 34 (6), 1–20.

Daly, A., Hess, S., de Jong, G., 2012. Calculating errors for measures derived from choice modelling estimates. Transp. Res. Part B: Methodol. 46 (2), 333–341.

Delft High Performance Computing Centre (DHPC), 2024. DelftBlue Supercomputer (Phase 2). URL: https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2.

Donaldson, S.I., Grant-Vallone, E.J., 2002. Understanding self-report bias in organizational behavior research. J. Bus. Psychol. 17, 245–260.

Dörner, R., Göbel, S., Effelsberg, W., Wiemeyer, J., 2016. Introduction to serious games. In: Serious Games Foundations, Concepts and Practice. Springer, pp. 1–34.

Gelman, A., Loken, E., 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Dep. Stat. Columbia Univ. 348 (1–17), 3.

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Syst. Appl. 78, 273–282.

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

Hancock, T.O., Choudhury, C.F., Hess, S., 2022. Secret in their eyes’: Incorporating eye-tracking and stress indicator data into travel behaviour models. Available at SSRN 4129033.

Hensher, D.A., 2004. Identifying the influence of stated choice design dimensionality on willingness to pay for travel time savings. J. Transp. Econ. Policy (JTEP) 38 (3), 425–446.

Hensher, D.A., Rose, J.M., Greene, W.H., 2015. Applied Choice Analysis. Cambridge University Press.

Hess, S., Daly, A., 2024. Handbook of Choice Modelling. Edward Elgar Publishing.

Hess, S., Palma, D., 2019. Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. J. Choice Model. 32, 100170.

Kang, J., Liu, M., Qu, W., 2017. Using gameplay data to examine learning behavior patterns in a serious game. Comput. Hum. Behav. 72, 757–770.

Krulak, C., 1997. Military Thinking and Decision Making Exercises. Report 1500.55, United States Marine Corps.

Lazar, J., Feng, J.H., Hochheiser, H., 2017. Research Methods in Human-Computer Interaction. Morgan Kaufmann.

Liebe, U., van Cranenburgh, S., Chorus, C., 2023. Maximizing utility or avoiding losses? Uncovering decision rule-heterogeneity in sociological research with an application to neighbourhood choice. Sociol. Methods Res. 00491241231186657.

Louviere, J., 2000. Stated Choice Methods: Analysis and Applications. Cambridge University Press.

Maimon, O., Rokach, L., 2005. Data Mining and Knowledge Discovery Handbook, vol. 2, (no. 2005), Springer.

Mariel, P., Hoyos, D., Meyerhoff, J., Czajkowski, M., Dekker, T., Glenk, K., Jacobsen, J.B., Liebe, U., Olsen, S.B., Sagebiel, J., et al., 2021. Environmental Valuation with Discrete Choice Experiments: Guidance on Design, Implementation and Data Analysis. Springer Nature.

Martín-Baos, J.Á., López-Gómez, J.A., Rodriguez-Benitez, L., Hillel, T., García-Ródenas, R., 2023. A prediction and behavioural analysis of machine learning methods for modelling travel mode choice. Transp. Res. Part C: Emerg. Technol. 156, 104318.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Van Ruijven, T., Lo, J., Kortmann, R., Wenzler, I., 2014. The research and evaluation of serious games: Toward a comprehensive methodology. Br. J. Educ. Technol. 45 (3), 502–527.

McFadden, D., 1974. The measurement of urban travel demand. J. Public Econ. 3 (4), 303–328.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. J. Appl. Econometrics 15 (5), 447–470.

Medler, B., Magerko, B., 2011. Analytics of play: Using information visualization and gameplay practices for visualizing video game data. Parsons J. Inf. Mapp. 3 (1), 1–12.

Meijer, S., 2009. The Organisation of Transactions: Studying Supply Networkd Using Gaming Simulation. Wageningen University and Research.

Meißner, M., Musalem, A., Huber, J., 2016. Eye tracking reveals processes that enable conjoint choices to become increasingly efficient with practice. J. Mark. Res. 53 (1), 1–17.

Meyerhoff, J., Oehlmann, M., Weller, P., 2015. The influence of design dimensions on stated choices in an environmental context. Environ. Resour. Econ. 61, 385–407.

Michael, D.R., Chen, S.L., 2005. Serious Games: Games That Educate, Train, and Inform. Muska & Lipman/Premier-Trade.

Olejniczak, K., Newcomer, K.E., Meijer, S.A., 2020. Advancing evaluation practice with serious games. Am. J. Eval. 41 (3), 339–366.

Ortelli, N., Hillel, T., Pereira, F.C., de Lapparent, M., Bierlaire, M., 2021. Assisted specification of discrete choice models. J. Choice Model. 39, 100285.

Ortúzar, J.d., Willumsen, L.G., 1978. Learning to manage transport systems. Traffic Eng. Control. 19, 236–239.

Ortúzar, J.d., Willumsen, L.G., 1982. GUTS, un juego de planificación del transporte urbano. EURE 9, 85–96.

Páez, A., Boisjoly, G., 2022. Discrete Choice Analysis with R. Springer.

Parady, G., Ory, D., Walker, J., 2021. The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. J. Choice Model. 38, 100257.
Paz, A., Arteaga, C., Cobos, C., 2019. Specification of mixed logit models assisted by an optimization framework. J. Choice Model. 30, 50–60.
Plass, J.L., Homer, B.D., Kinzer, C.K., 2015. Foundations of game-based learning. Educ. Psychol. 50 (4), 258–283.
Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., Podsakoff, N.P., 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J. Appl. Psychol. 88 (5), 879.
Poplin, A., 2012. Playful public participation in urban planning: A case study for online serious games. Comput. Environ. Urban Syst. 36 (3), 195–206.
Rawitsch, D., 1978. Oregon trail. Creat. Comput. 4, 132–139.
Rodrigues, F., Ortelli, N., Bierlaire, M., Pereira, F.C., 2020. Bayesian automatic relevance determination for utility function specification in discrete choice models. IEEE Trans. Intell. Transp. Syst. 23 (4), 3126–3136.
Rose, J.M., Bliemer, M.C., 2009. Constructing efficient stated choice experimental designs. Transp. Rev. 29 (5), 587–617.
Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., et al., 2018. Many analysts, one data set: Making transparent how variations in analytic choices affect results. Adv. Methods Pr. Psychol. Sci. 1 (3), 337–356.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366.
Smith, S.P., Blackmore, K., Nesbitt, K., 2015. A meta-analysis of data collection in serious games research. Serious Games Anal.: Methodol. Perform. Meas., Assess. Improv. 31–55.
Snow, E., Jacovina, M., Varner, L., Dai, J., McNamara, D., 2014. Entropy: A stealth measure of agency in learning environments. In: Educational Data Mining 2014. Citeseer.
Squire, K., 2006. From content to context: Videogames as designed experience. Educ. Res. 35 (8), 19–29.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press.
Tukey, J.W., 1977. Exploratory Data Analysis. Reading/Addison-Wesley.
Van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., Walker, J., 2022. Choice modelling in the age of machine learning-discussion paper. J. Choice Model. 42, 100340.
Van Dijk, E., De Dreu, C.K., 2021. Experimental games and social decision making. Annu. Rev. Psychol. 72 (1), 415–438.
Vecchio, G., 2024. Serious games and transport justice: Examining redistributive issues through classification and dictator games. J. Transp. Heal. 36, 101817.
Walker, J.L., Li, J., 2007. Latent lifestyle preferences and household location decisions. J. Geogr. Syst. 9, 77–101.
Wang, B., Gao, L., Juan, Z., 2017. Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. IEEE Trans. Intell. Transp. Syst. 19 (5), 1547–1558.
Wang, S., Mo, B., Zhao, J., 2020. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. Transp. Res. Part C: Emerg. Technol. 112, 234–251.
Wicherts, J.M., Veldkamp, C.L., Augusteijn, H.E., Bakker, M., Van Aert, R.C., Van Assen, M.A., 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Front. Psychol. 7, 1832.
Willumsen, L.G., Ortúzar, J.d., 1985. Intuition and models in transport management. Transp. Res. Part A: Policy Pr. 19 (1), 51–57.
Xu, J., Min, J., Hu, J., 2018. Real-time eye tracking for the assessment of driver fatigue. Heal. Technol. Lett. 5 (2), 54–58.
Zaki, M.J., 2001. SPADE: An efficient algorithm for mining frequent sequences. Mach. Learn. 42, 31–60.
Zhang, Y., Paquette, L., 2023. Sequential pattern mining in educational data: The application context, potential, strengths, and limitations. In: Educational Data Science: Essentials, Approaches, and Tendencies: Proactive Education Based on Empirical Big Data Evidence. Springer, pp. 219–254.