Putkonen, Aini; Nioche, Aurélien; Tanskanen, Ville; Klami, Arto; Oulasvirta, Antti

How Suitable Is Your Naturalistic Dataset for Theory-based User Modeling?

Published in:
UMAP2022 - Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization

DOI:
10.1145/3503252.3531322

Published: 07/04/2022

Document Version
Publisher's PDF, also known as Version of record

How Suitable Is Your Naturalistic Dataset for Theory-based User Modeling?

Aini Putkonen
aini.putkonen@aalto.fi
Aalto University
Finland

Aurélien Nioche
nioche.aurelien@gmail.com
Aalto University
Finland

Ville Tanskanen
ville.tanskanen@helsinki.fi
University of Helsinki
Finland

Arto Klami
arto.klami@helsinki.fi
University of Helsinki
Finland

Antti Oulasvirta
antti.oulasvirta@aalto.fi
Aalto University
Finland

ABSTRACT

Theory-based, or "white-box," models come with a major benefit that makes them appealing for deployment in user modeling: their parameters are interpretable. However, most theory-based models have been developed in controlled settings, in which researchers determine the experimental design. In contrast, real-world application of these models demands setups that are beyond developer control. In non-experimental, naturalistic settings, the tasks with which users are presented may be very limited, and it is not clear that model parameters can be reliably inferred. This paper describes a technique for assessing whether a naturalistic dataset is suitable for use with a theory-based model. The proposed parameter recovery technique can warn against possible over-confidence in inferred model parameters. This technique also can be used to study conditions under which parameter inference is feasible. The method is demonstrated for two models of decision-making under risk with naturalistic data from a turn-based game.

CCS CONCEPTS

• **Human-centered computing** → **User models**; *HCI theory, concepts and models.*

KEYWORDS

naturalistic data, parameter recovery, user modeling, theory-based models, risky choice

ACM Reference Format:
Aini Putkonen, Aurélien Nioche, Ville Tanskanen, Arto Klami, and Antti Oulasvirta. 2022. How Suitable Is Your Naturalistic Dataset for Theory-based User Modeling?. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22), July 4–7, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3503252.3531322

UMAP '22, July 4–7, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9207-5/22/07.
https://doi.org/10.1145/3503252.3531322

1 INTRODUCTION

This paper contributes to attempts to use theory-based, so-called white-box, models as user models. We use the term "theory-based model" for an instantiation of a theory that explains cause-and-effect relationships between inputs and outputs with interpretable parameters. In contrast, data-driven, or "black-box," models, while they may be grounded in theory, learn a relationship between features. Data-driven models are widely used in user modeling but can lack explainability [12]. Theory-based models of human behavior have succeeded in explaining decision-making in various controlled contexts [e.g., 16, 22, 29] and, therefore, could help to increase interpretability in user modeling.

We address a discrepancy between how theory-based models are developed and how they should be deployed as user models. Theory-based models, especially the models of human decision-making examined in this paper, are generally developed under controlled settings (i.e., the researcher can choose the experimental design). However, user models often get applied in situations where only non-experimental, or naturalistic, data are available. Naturalistic datasets can take many forms (e.g., clickstream data and game logs), but their defining feature is that a user has produced data without experimental oversight, doing so in scenarios that arise in the natural course of system use. This discrepancy produces a need to assess whether a theory-based model can be fitted to a naturalistic dataset.

This paper presents *a parameter recovery technique* [17, 33] for user models. Parameter recovery refers to the study of conditions in which model parameters can be reliably inferred. Our technique can be used for two purposes: to assess how suitable a naturalistic dataset is for a given theory-based user model and, secondly, to study requirements for reliable application of a model. It can help practitioners to avoid misinterpretation of their modeling results.

The proposed technique should be implemented in a systematic modeling workflow. Consider a flow in which a model is picked, it gets fitted to data, and its performance is evaluated in an iterative process (Box's loop [5] in Figure 1A). Naturalistic datasets require special attention in the modeling workflow. A naturalistic dataset, in the case of models of decision-making, has two components: 1) n tasks of choosing from m discrete options (task data) and 2) users' actual choices in the tasks (choice data). This split is applicable to user modeling in a relatively general manner, since users frequently make such choices as picking which method to use in a sub-task or

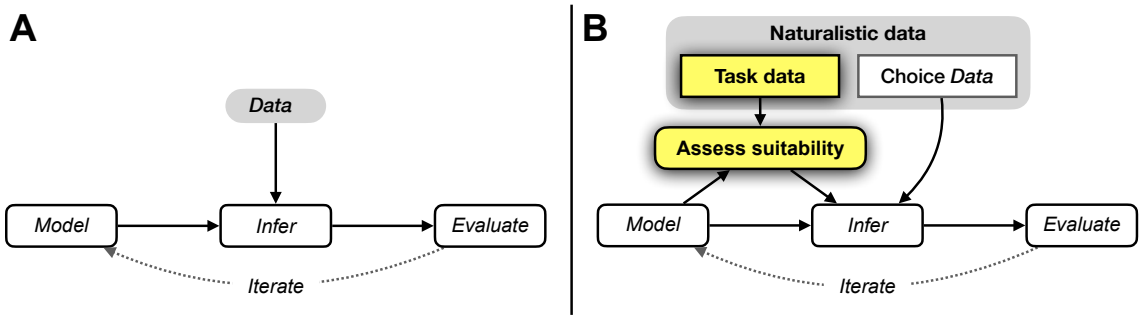

Figure 1: Before model fitting, one should assess a naturalistic dataset for suitability, since it may be uninformative for the given model. A. Typical steps in iterative modeling (Box's-loop adaptation based on Blei's work [5]). B. Our suggested modeling workflow, including a parameter recovery step to assess the suitability of naturalistic data. This considers that the data include information on tasks and users' choices in these tasks. The tasks' suitability for the model should be assessed before fitting.

how to configure a system [19]. For instance, mobile-game players may be given a task of choosing one tool among multiple options presented to them in an in-app shop, or a user could choose an item from a menu. Crucially, in naturalistic settings these tasks emerge from users' interactions with the system and are not designed to interrogate a specific theory. Given a theory-based model, we need to assess the suitability of the task data for said model (see Figure 1B). In contrast, the task quality in controlled experiments is (or should be) checked *before* data collection [17, 33]. We suggest using parameter recovery, *after* data collection, to assess whether tasks in an existing dataset are suitable for a theory-based model.

The parameter recovery technique for assessing the suitability of naturalistic data consists of two steps: a simulation and an assessment. In a parameter recovery simulation, an artificial user behaving in line with known "generating" parameters of a model is simulated in the same tasks as the real users. Then, the chosen model is fitted to the simulated choice data. Parameter recovery is adequate if the correlation between the generating and inferred parameters is sufficient, which we measure with the Pearson correlation coefficient in this paper.

Here, we demonstrate the proposed technique via data from *Blade Runner: Rogue*, a turn-based game from Next Games. In the game, a player chooses a skill to use to attack an enemy in short battles. Our results, from assessing two classic models of decision-making via the log data from 510 players, indicate that only 6% of the players considered were given tasks rich enough for adequate parameter recovery for one of the two models tested. These results imply that parameter inference for naturalistic datasets should be exercised with caution, and that checking the parameter recovery is crucial. Finally, we discuss how the tasks' properties may result in inadequate parameter recovery.

This paper makes the following contributions:

- We present a parameter recovery technique for assessing whether a theory-based model can be fitted to a naturalistic dataset.
- We demonstrate our approach by inferring players' attitudes to risk from game logs, illustrating how to avoid misinterpretation due to an unsuitable naturalistic dataset.

The proposed technique is relevant for any application where naturalistic data can be split into tasks and choices. We focus on two specific models in this paper, but other types that may be applied in user modeling include memory models [2] for intelligent tutoring systems and ambiguity models [16] to describe novice users.

The body of the paper is structured thus: We describe related work in Section 2. The proposed technique is detailed in Section 3 and applied to the two models of decision-making in Section 4. Results are discussed in Section 5.

2 RELATED WORK

With this section, we briefly review types of models used in user modeling, with special attention to theory-based models that describe behavior via interpretable parameters. We then turn to the broader literature on statistical modeling and discuss recent attempts to rethink modeling workflows for higher reliability.

2.1 Theory-based models and user modeling

User models can be viewed as representing user behavior via parameters inferred from interactions with a system [1]. In (purely) data-driven models these parameters are learned from data, whereas in models we refer to as theory-based the parameters are determined *a priori* to have specific interpretations. In the past few decades, using machine-learning models to learn the parameters of a user model has been popular; however, this approach comes with limitations, among them computational complexity, lack of scrutability, and a need for large datasets [12, 32]. On the other hand, a theory (in the social-sciences context) is intended to provide a coherent, rational, and plausible explanation of cause and effect in a given phenomenon [3]. The benefit of incorporating theory-based insights into user models has been discussed in contexts including recommender systems [10] and gaming [25]. Theories of human behavior and cognition have been extensively studied in experimental psychology, cognitive science, and economics, largely in well-controlled settings [15]. Since user models are often employed in naturalistic settings, theory-based approaches' applicability in these domains requires attention.

2.2 Modeling workflows and guidelines

Recently, the idea of implementing a workflow has gained attention in efforts to improve the reliability of modeling for statistics and data science [14], and similar iterative guidelines have been proposed in the context of cognitive science [17, 33]. In representing a systematic approach to implementing a computational model, these workflows or guidelines are aimed at better model building and parameter inference [13, 33]. Among the general challenges in modeling are the uncertainty in the choice of model [13] and assessment of the model's reliability at describing latent (i.e., non-observable) processes [17, 33]. Modeling workflows and guidelines combat these issues by considering steps such as simulations or predictions, comparison of models, and parameter recovery, depending on the problem's specifics and the estimate types used [13, 17, 33]. However, these steps in modeling are not exactly new; before attempts to draw them together into a systematic workflow or a set of guidelines, they existed as implicit knowledge passed from modeler to modeler [17]. Despite their benefits, such iterative workflows are not widely reported in the user modeling literature.

2.3 Parameter recovery

Parameter recovery is often discussed as an element of a modeling workflow [14] or modeling guidelines [17, 33]. It is used to assess the parameter estimates produced with a chosen model and estimation technique for accuracy and consistency with a given dataset [17]. One runs a parameter recovery simulation by fitting a model to data, generated with known parameters of the same model, to inspect whether the generating parameters are close to the inferred ones. This technique is considered good practice in cognitive modeling, and it also serves as a bug check for the software employed for parameter inference [17, 33]. With non-linear models, running a parameter recovery simulation is important because it can reveal details about their parameter-estimation properties [17, 26]. Parameter recovery is closely related to *identifiability*. Parameter identifiability, redundancy, estimability [11], and symmetry [21] all involve the same problem of inferring the generating parameter vector uniquely from data. There are many reasons for poor parameter recovery in a given problem, with identifiability of the model's parameters being one of them.

Among the properties examined previously in parameter recovery studies are stimulus quality and the chosen estimation procedure. Broomell et al. [8] argue that current modeling practices often neglect discussion of the "model–stimulus relationship" – that is, of the impact of the chosen stimuli on parameter-estimation and model-comparison results. Sloman et al. [26] and Broomell and Bhatia [7] attest to a great influence of the experimental design on the quality of parameter estimates, from demonstrating it in the context of models of decision-making. Similarly, Toubia et al. [28] used parameter recovery to optimize experimental design. On the other hand, some studies examine parameter recovery from the perspective of the estimation procedure; for instance, Nilsson et al. [20] propose favoring a hierarchical Bayesian parameter-estimation procedure over the more common maximum-likelihood estimation, for better recovery. The work described has often taken models of decision-making under risk as the model class (namely, using cumulative prospect theory, introduced further on). However, parameter recovery has been explored also in such contexts as models of inter-temporal choice [4] and cognitive decision models [30]. The focus in the work reviewed is largely on datasets obtained in controlled experiments. To the best of our knowledge, parameter recovery is not widely used in user modeling, and its application with naturalistic datasets is a less explored topic.

3 WALK-THROUGH OF THE PROPOSED PARAMETER RECOVERY TECHNIQUE

This section describes the proposed parameter recovery technique in detail. The starting point is a wish to employ a theory-based model to explain user behavior. Our goal is to evaluate whether this model can be used with the data at hand.

Prerequisites. The proposed recovery technique entails a minimum requirement for application: *the dataset must contain information about both choices and tasks*. For instance, assume we have mobile-game logs containing information about tasks wherein players choose a tool to purchase from the in-app shop. The options (the various tools) and choices (the tool actually chosen) are recorded. Hence, both task and choice data are available. This condition would not be fulfilled if only the choices each player made were recorded, without any contextual information on the tasks.

Modeling workflow. Parameter recovery is applied as a part of the following modeling workflow (see Figure 2):

(1) The naturalistic dataset is split into task and choice data for each user. In the example of a purchase from the game's in-app shop, this translates to splitting the log data into tasks (options, the tools the players chose from) and choices (the tool chosen).
(2) Choice data are set aside for later. Each user's task datum is evaluated, in light of some model of interest, with parameter recovery. For instance, we could assume that users in the game example may be modeled via an arbitrary decision-making model with a parameter λ. We would take a sequence of tasks from one user and simulate choices of tools with different values of λ. Then we would assess parameter recovery for the simulations. The next section provides detailed explanation of parameter recovery.
(3) A pre-specified assessment criterion is used to assess which users were presented with tasks that are informative enough for the given model. In the game case, we would be interested in those players presented with options "rich enough" to be modeled by means of the chosen decision-making model.
(4) Now choice data are considered. The given model is fitted to the choice data of only the users who passed step 3.

Parameter recovery simulations and assessment. The proposed parameter recovery technique consists of simulating and assessing the recovery. Parameter recovery involves only the task data (as opposed to choice data). The suitability of the task data is measured by simulating the choices that a model with known parameters produces and studying whether we can recover the same parameters. The metric of interest is the distance between two sets of parameters: one used to generate choices and one inferred from those choices. If parameter recovery is inadequate, it is unlikely

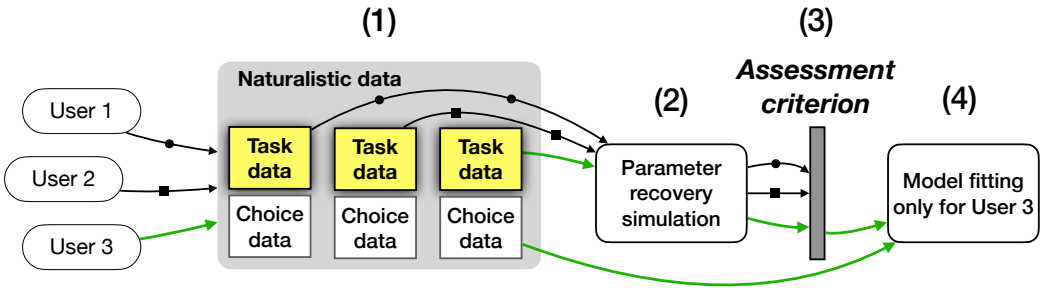

Figure 2: The modeling workflow for assessing the data's suitability for the model of decision-making. The naturalistic dataset is split into task and choice data for each user (1). Then, the data for each user are evaluated via parameter recovery (2). A pre-specified assessment criterion indicates whether the tasks are informative enough for the given model (3). The model is fitted to choice data only for those users filtered past step 3 (4).

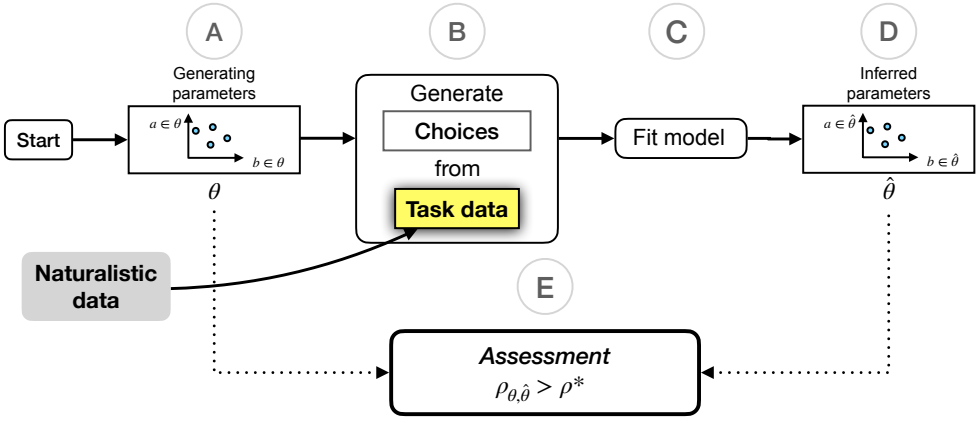

Figure 3: Parameter recovery simulation. A. Models of decision-making with known generating parameters θ are considered. B. Choices are generated via the model of decision-making with known parameters over a set of tasks. Prior work has often used parameter recovery with experimental tasks [17, 33, etc.]. We propose extending the technique to be used with naturalistic data. C, D. The chosen model is fitted to the choices, and $\hat{\theta}$ is obtained. E. Closeness of θ and $\hat{\theta}$ is measured through Pearson correlation coefficient ρ. Note that other assessment criteria may be used as well.

that we can recover the ground-truth model – the parameters that describe how users produce choice data.

Here, we consider a parameter recovery simulation including the following steps (for other descriptions, see the literature [4, 17, 33]). We take a model of decision-making with known generating parameters θ (Figure 3, pane A) and generate choices for a set of tasks by using this model (Figure 3, pane B). Our implementation employs 100 distinct combinations of parameters for each model, which is the number used by Ballard et al. [4]. Then, the candidate model is fitted to these choices (pane C) to obtain inferred parameters $\hat{\theta}$ (pane D). These inferred parameters $\hat{\theta}$ should be close to the generating parameters θ, assessed against a set criterion (pane E). The criterion considered here is the Pearson correlation coefficient $\rho_{\theta,\hat{\theta}}$ between θ and $\hat{\theta}$. We set some threshold ρ^* that $\rho_{\theta,\hat{\theta}}$ should exceed. We use $\rho^* = 0.7$, which can be thought of as a large correlation coefficient in contexts of human behavioral data [18]. The exact value of ρ^* depends on the dataset, and there may be a tradeoff between the quality of parameter inference and the number of users accepted for model fitting. Due to random sampling of the generating parameters θ, some variation may exist in the number of users accepted. For models with a larger number of parameters, visualizing the results may be more challenging, but the computation of the metric to either reject or accept a model still applies. We use maximum-likelihood estimation as a goodness-of-fit metric when fitting parameters, using an L-BFGS-B optimization algorithm [9] (SciPy library) for generating the samples. The parameter-inference method used in parameter recovery should match that used in model fitting. We note, however, that these specific choices are not central to the proposed technique, and the principle expressed can be applied with other metrics and inference methods.

4 APPLICATION: MODELS OF DECISION-MAKING UNDER RISK

Our demonstration of the proposed parameter recovery technique applies models of decision-making under risk to a naturalistic game dataset. We start by describing the corresponding tasks, risky-choice problems. Then we outline two candidate theories: probabilistic cumulative prospect theory (P-CPT) and probabilistic expected utility theory (P-EUT). We illustrate in two ways how properties of the data may give rise to inadequate parameter recovery: 1) via a demonstration with a naturalistic dataset and 2) by analyzing the task properties with an artificial dataset.

4.1 Problem formulation

We focus on risky-choice problems, where a user chooses from options that are lotteries; that is, each option is a probability distribution over rewards x. Consider a choice problem where a user has to choose among m options L_i, where $i = 1, 2, ..., m$ (see Figure 4). Each option L_i is a lottery with two outcomes: a high outcome x_{high} at probability $p_{x_{\text{high}}}$ and a low one x_{low} at probability $p_{x_{\text{low}}}$. We assume users calculate a subjective expected utility (SEU) for each lottery, which guides their decision and depends on the model considered. In our calculation of SEU, the outcomes are normalized with respect to the highest outcome a user observes when using the application. We assume player behavior to be probabilistic. Here, we consider both P-EUT and P-CPT as candidates for modeling.

4.2 Candidate models

4.2.1 Choice stochasticity. We assume that the users' decisions are probabilistic. Users will choose the lottery with the highest SEU most of the time, but occasionally they explore other options (some prior work has taken a similar approach [6, 24, 27, 33]). Scholars have observed that people's decisions may vary between very similar tasks and that probabilistic models explain behavior better than deterministic approaches [24]. The probability of a user choosing each lottery is calculated via:

$$p_{\text{choice}}(L_i) = \frac{e^{\beta \text{SEU}(L_i)}}{\sum_{j=1}^{m} e^{\beta \text{SEU}(L_j)}} \tag{1}$$

which is a softmax function where m is the number of lotteries and $\beta \in [0, 100]$ is the inverse temperature parameter, with values close to 0 indicating random behavior and values close to 100 indicating deterministic behavior. When β is close to 0, the chances of a user choosing each lottery are nearly equal, whereas when β is close to 100 the user will almost certainly choose the lottery with the highest SEU. The SEU values depend on whether we model the user behaving in line with expected utility theory (EUT) vs. cumulative prospect theory (CPT).

4.2.2 Probabilistic expected utility theory. Individuals following the explanation of EUT, a normative theory, obey certain axioms of decision-making, whereby they choose as if they were maximizing their *expected utility* (EU) [31], the utility being a quantity that represents the satisfaction or pleasure of the individual. We use a probabilistic form (P-EUT), with a constant relative risk-aversion utility function u expressed as:

$$u(x) = x^{1-\alpha} \tag{2}$$

where $x \in \mathbb{R}$ is a possible outcome and $\alpha \in [-3, 0.75]$ is the coefficient of relative risk aversion. The bounds are chosen such that they produce symmetric utility functions around the line of equality ($y = x$). The higher α is, the more risk-averse the individual. If $\alpha = 0$, the individual is risk-neutral. The expected utility for a lottery is calculated as:

$$EU(L) = u(x_{\text{low}}) \cdot (1 - p_{\text{high}}) + u(x_{\text{high}}) \cdot p_{\text{high}} \tag{3}$$

We assume that the individuals' behavior is probabilistic, following Equation 1, where SEU=EU.

4.2.3 Probabilistic cumulative prospect theory. CPT is an extension of EUT, a descriptive theory that assumes that people weight probabilities non-linearly [29]. The original CPT is deterministic and assumes choice of the lottery with the highest *cumulative prospect value* (CPV), whereas we use a probabilistic CPT (P-CPT).

We use the utility function presented in Equation 2, alongside the following probability-weighting function [23]:

$$w(p) = e^{-(-ln\ p)^{\gamma}} \tag{4}$$

Here, $\gamma \in [0.001, 0.999]$ is the probability-weighting parameter. The closer it is to 0, the more the function takes on an inverse S shape. The CPV for a lottery is then obtained by means of:

$$CPV(L) = u(x_{\text{low}}) \cdot (1 - w(p_{\text{high}})) + u(x_{\text{high}}) \cdot w(p_{\text{high}}) \tag{5}$$

We assume that the players' behavior is probabilistic, according to Equation 1, where SEU=CPV. The three forms of the functions, for equations 1, 2, and 4, follow those suggested by Stott [27].

4.3 Modeling game players by using models of decision-making under risk

This subsection describes how parameter recovery can be used to assess an existing naturalistic dataset obtained from an interactive system. Using data collected in a turn-based game where each player is given a set of risky-choice problems as tasks, we determine which users were shown tasks with sufficient parameter recovery for model fitting.

4.3.1 Task description. Our work focuses on *Blade Runner: Rogue*, a non-deterministic tactical turn-based game developed by Next Games. Figure 5A presents a screenshot of a task (a risky-choice problem) from the game. In a typical gaming session, the player controls a set of characters in several short turn-based battles. Each of the battles comprises several risky-choice problems. In turns, the player will choose a skill to use and a target to attack with a predetermined character. Each of these skill–target combinations can be represented as a lottery, where the user receives a reward with a certain probability. The user interface shows the player details of the characters' relative strengths, the targets, and the skills. We assume that the player can use this information to infer the probability distributions of outcomes for the various lotteries (an abstraction is presented in Figure 5B). Game logs from anonymous users in this game inform inferences of attitudes to risk.

4.3.2 Data preparation. The raw dataset, obtained from a gaming company, contained information about 510 users playing the game during a specific time interval. That interval was picked on the basis of restrictions related to the game version, chosen to guarantee that

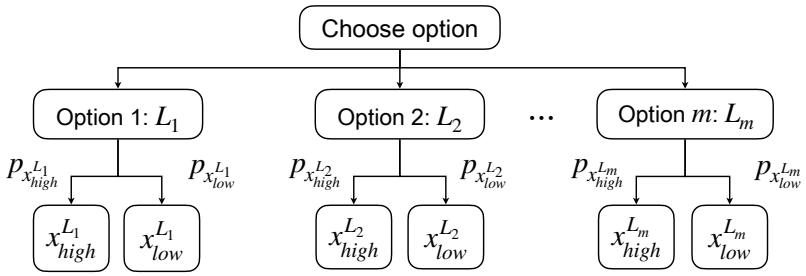

Figure 4: Parameter recovery demonstrated via risky-choice problems. With these tasks, the user chooses from m options L_i, where $i = 1, 2, ..., m$. Each option has a high outcome x_{high} and a low outcome x_{low}, at probabilities $p_{x_{\text{high}}}$ and $p_{x_{\text{low}}}$, respectively.

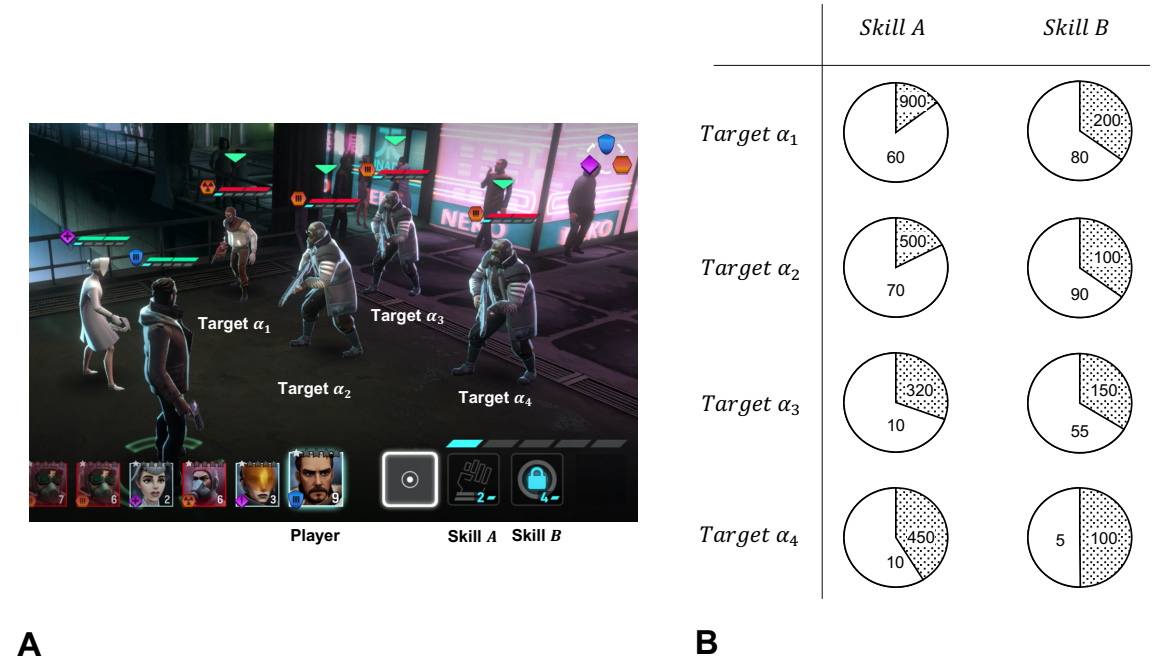

Figure 5: Illustration of applying the proposed technique via data from the game *Blade Runner: Rogue*. A. A screenshot of a task in the game, where a player chooses a skill to attack a target with (screenshot courtesy of Next Games). B. Illustration of one example task: a player has to choose from among eight lotteries, where each combination of a target and skill is a lottery (illustrated as a pie chart with rewards overlaid). We assume the player can construct these lotteries via information available from the user interface. The player chooses one lottery.

all data came from the same version of the game. The game data were anonymized. In all, 58 users' data were removed because of incompleteness. Our analysis used the data of 452 users in total, corresponding to 452 unique sets of tasks these users saw during game play. We had no control over the tasks in these sets. The information contained in this dataset was converted into a structure for tasks and choices. Each task was a risky-choice problem where the player has to choose from among m lotteries, with 2 outcomes in each (see Figure 4). Each choice picked a skill and a target for an attack (one of the pie charts in Figure 5B). The outcome is damage from a "normal" hit (x_{low}) and a "critical" hit (x_{high}), inflicted on the target in game-specific health units. A critical hit deals out greater damage but occurs at lower probability. Approximations for the probability distributions of the lotteries in each task were extracted by means of expert knowledge of the game mechanics. Information about the attacking character, target, skill, and relative strengths were used to reverse-engineer a stochastic function for determining an approximation for the damage each target–skill combination yields. A proxy for the damage observed was available in the dataset. This was compared to the estimates obtained via reverse-engineering. If both of the possible approximations of outcomes (x_{low} and x_{high}) from the chosen lottery deviated by more than 25% from the outcome observed, the task was discarded, to account for inaccuracies in the reverse-engineered estimates. In

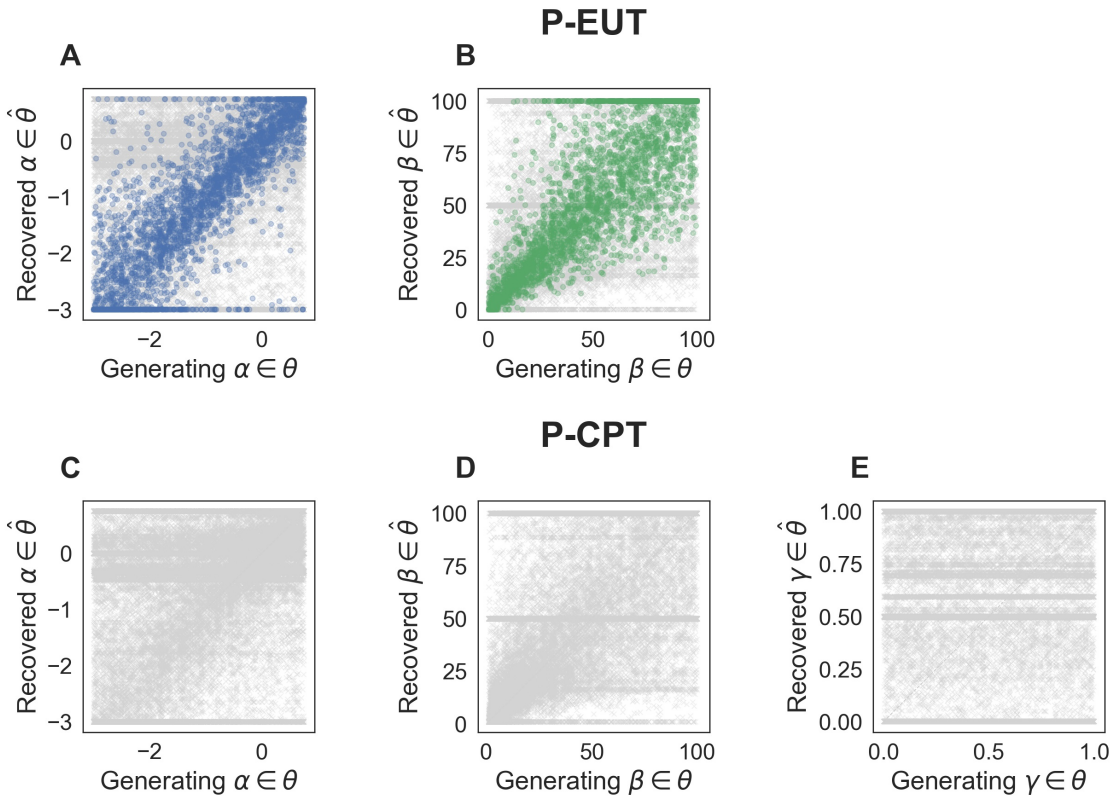

Figure 6: Parameter recovery results with naturalistic game data. Each marker corresponds to a model parameter used to generate choices. The colored markers correspond to parameters for generating choices for those players whose task data passed the parameter recovery threshold ($\rho_{\theta,\hat{\theta}} > 0.7$). A–B. Parameter recovery is sufficient for the task data presented to 6% of the players when they are assumed to behave in accordance with probabilistic expected utility theory (P-EUT). C–E. With probabilistic cumulative prospect theory (P-CPT), no users received tasks with sufficient parameter recovery.

addition, all deterministic tasks wherein only one option was presented to the player were removed. Some inaccuracies arose from game features not captured in the lottery estimates (e.g., healing of characters and damage inflicted over time). The distribution of the number of tasks per player was right-skewed. Even though the modeling did not encompass all tasks presented to the players in the game, this long right tail suggests that many players tend to play it for only a few rounds.

4.3.3 Results. Parameter recovery was poor for a large proportion of the naturalistic dataset (this is visible as large discrepancies between the generating and recovered parameters in Figure 6). Only 6% of the players' task sets exceeded the threshold $\rho_{\theta,\hat{\theta}} > 0.7$ when modeled via P-EUT. No players' task sets passed this threshold when modeled with P-CPT, a result most likely related to the small range of outcome probabilities for "critical" damage $p_{x_{\text{high}}}$ in the raw dataset. We analyze possible reasons for the poor parameter recovery in the following Subsection 4.4.

4.3.4 Implications for parameter inference. Since only 6% of the players were given tasks for which parameter recovery sufficed when P-EUT is used as the candidate model (i.e., $\rho_{\theta,\hat{\theta}} > 0.7$), we consider model fitting for these players alone. On visual inspection, P-EUT seems insufficiently descriptive of the behavior of some players – namely, those individuals for whom β is low, who seem to have been choosing randomly (horizontal line in Figure 7B). However, since we verified the quality of parameter recovery for the tasks shown to these users, the issues with the model's ability to describe player behavior can be put down to modeling considerations other than the tasks presented.

4.4 Examining task properties to understand inadequate parameter recovery

This portion of the paper illustrates how restrictions that arise in naturalistic settings (e.g., with the game data considered in Subsection 4.3) can affect parameter recovery. We demonstrate this by means of a parameter recovery simulation using an artificial dataset. The choice problems generated mimic the structure of those used for the game data in Subsection 4.3. An artificial dataset grants us full control over the modeling inputs. Specifically, we show how the following properties affect parameter recovery: 1) number of observations, 2) range of outcome probabilities, and 3) outcome

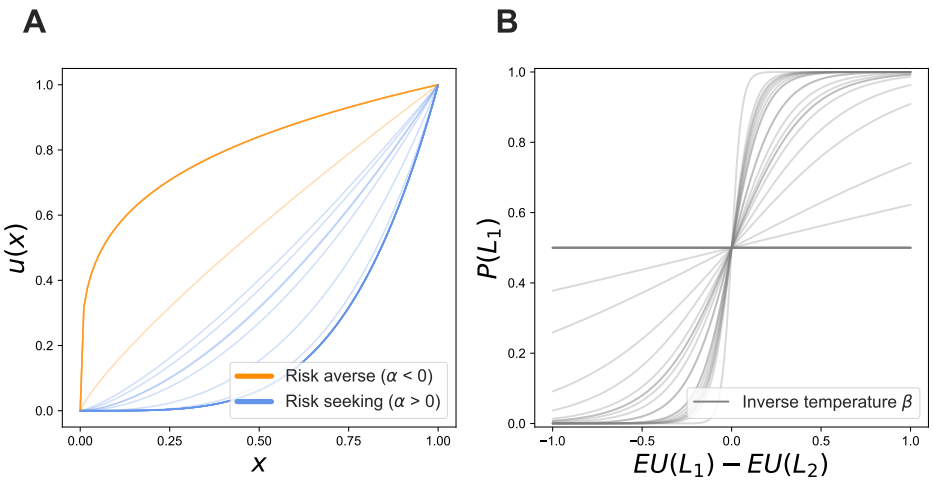

Figure 7: Parameter inference results for the 31 game players presented with tasks where parameter recovery is sufficient for probabilistic expected utility theory (P-EUT). A. Utility functions ($u(x) = x^{1-\alpha}$) plotted for point estimates of the risk-aversion parameter α for each player. Concave (orange) utility functions describe a risk-averse and convex (blue) a risk-seeking player. B. The inverse temperature parameter β's effect on the probability of choosing lottery L_1, when there are two options L_1 and L_2 with subjective expected utility EU. Lower values of β correspond to more stochastic behavior (horizontal line).

magnitudes. The results presented in this section were produced with a Jupyter Notebook available online.[1]

4.4.1 Task description. The tasks we consider are simplified versions of the game presented in Subsection 4.3: risky-choice problems with two lotteries to choose from in each (i.e., we have lotteries where $m = 2$ in a task represented in Figure 4). All told, there are n_{choices} choice problems. Even though the game data presented in Subsection 4.3 contain tasks for which the number of lotteries $m > 2$, we constrain the number of options to simplify computation. Each lottery has two potential outcomes, high and low reward (x_{high} and x_{low}, respectively). We confine ourselves to positive rewards, to match the structure of the tasks in the aforementioned naturalistic dataset. The higher reward occurs with probability p_{high}. We employed P-EUT and P-CPT as candidate user models. The n_{choices} choice problems considered can be thought of as task data collected for one user. As noted in Subsection 4.3, we observed that for a large proportion of the users, the task data were not informative enough. This artificial setup enables us to examine potential reasons for the poor parameter recovery.

4.4.2 Data generation. Generating four versions of the artificial dataset described, we restrict different properties of the tasks (number of choices n_{choices}, probability of the highest reward resulting p_{high}, and range of rewards x_{high} and x_{low}) in order to mirror features of the lotteries encountered by the players in the naturalistic game data considered in Subsection 4.3. The artificial datasets generated are identical apart from the dimension that is modified. The exact (uniform) distributions from which we sampled data are described in the leftmost column in Figures 8 and 9.

[1] See https://github.com/aalto-ui/umap22-parameter-recovery.

4.4.3 Results. Then, we pass the resulting artificial datasets through the parameter recovery simulations described in Figure 3. When the tasks are not rich enough, the recovery results deteriorate as measured by means of the Pearson correlation coefficient ρ between generating and recovered parameters. The two parameters in P-EUT pass the $\rho^* > 0.7$ threshold for the original reference dataset and when the range of probabilities is restricted (see Figure 8, panes A and C). The three P-CPT parameters pass the parameter recovery check only in the reference dataset (see Figure 9, pane A).

4.4.4 Implications for parameter recovery with naturalistic datasets. Again, we observed that only 6% of users' task data passes the parameter recovery tests presented in Subsection 4.3 only for P-EUT, and our simulation with artificial data shows that the dataset's properties can lead to bad parameter recovery. Parameter recovery proves sufficient for all P-EUT parameters for two of the datasets considered: the original reference data and with outcome probabilities restricted. In contrast, P-CPT parameters pass the parameter recovery test only in the case of the original reference data. This result illustrates that P-EUT, as the simpler model, imposes restrictions on fewer dimensions of the dataset. The probability-weighting function used in P-CPT (Equation 4) creates a further requirement that the outcome probabilities have to fulfill. Indeed, the game data encompassed a limited range of outcome probabilities, which may explain why none of the players' task data showed sufficient parameter recovery. On the other hand, a limited number of observations leads to poor parameter recovery for both models. Users often played through only a small set of tasks in the game data considered in Subsection 4.3; this could be another reason for insufficient parameter recovery.

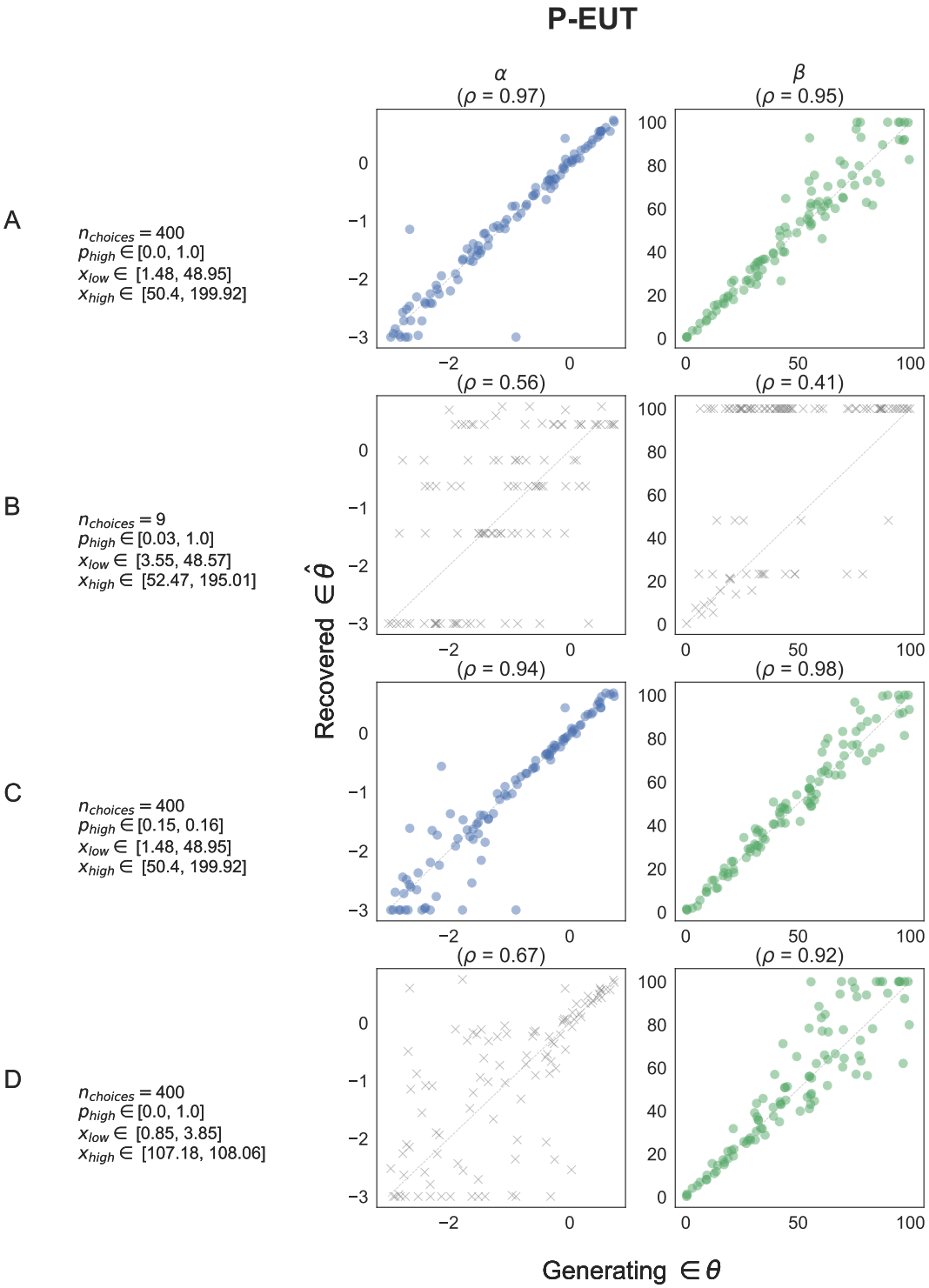

Figure 8: Parameter recovery simulation results for artificial datasets with probabilistic expected utility theory (P-EUT). Data are generated for n_{choices} choice problems with two outcomes (x_{high} and x_{low} with probabilities p_{high} and $1 - p_{\text{high}}$). The correlation between generating and recovered parameters is worse when certain properties of the dataset are limited. Those parameters with low correlation (i.e., $\rho_{\theta,\hat{\theta}} < 0.7$) are indicated with gray crosses. Several datasets were used: original reference data (A), number of observations limited (B), range of probabilities limited (C), and range of outcomes limited (D).

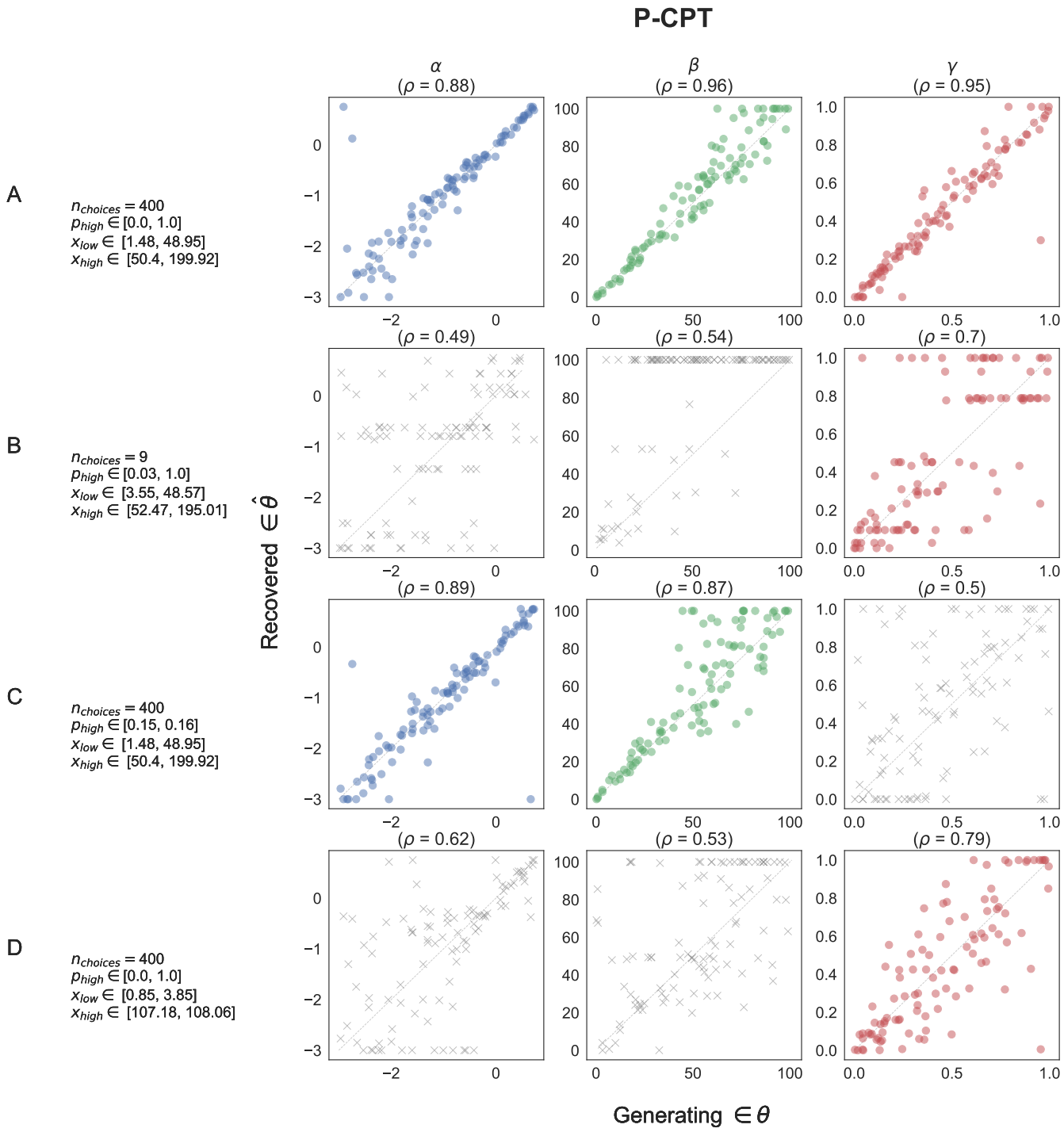

Figure 9: Parameter recovery simulation results for artificial datasets with probabilistic cumulative prospect theory (P-CPT). We generated data in n_{choices} choice problems with two outcomes (x_{high} and x_{low} and probabilities p_{high} and $1 - p_{\text{high}}$). Fewer datasets pass the parameter recovery test for P-CPT than P-EUT (see Figure 8). Gray crosses mark those parameters for which the correlation between generating and inferred parameters is low (i.e., $\rho_{\theta,\hat{\theta}} < 0.7$). The following datasets were included: original reference data (A), number of observations limited (B), range of probabilities limited (C), and range of outcomes limited (D).

5 DISCUSSION

We have presented a parameter recovery technique for assessing whether a theory-based model should be fitted to a naturalistic dataset. By doing so, we contribute to wider application of theory-based models in user modeling, which could offer several benefits. For instance, theory-based models provide a more interpretable alternative to data-driven techniques and are a better way of describing human behavior when fewer data are available [22]. That said, because theory-based models often are built in controlled experiments [15], their application in naturalistic settings requires special care since the tasks presented to users of interactive systems seldom are designed with a particular model in mind. We suggest that parameter recovery is a way for verifying that invalid conclusions are not drawn about users.

The demonstration here presents the proposed technique's application for a naturalistic dataset containing information about 510 players from a mobile game. In this application, the data comprise, as tasks, risky-choice problems where the user must choose from lotteries with two outcomes in each. Our results from applying two theory-based models for this dataset (P-EUT and P-CPT) suggest that we should fit only one model (P-EUT) for 6% of the players, since the other users were given tasks with insufficient parameter recovery. This result indicates that model fitting should be exercised with caution with naturalistic data, in that a strikingly low percentage of the players received tasks with sufficient parameter recovery. Hence, we recommend that those working with datasets of such a nature always validate their content by using parameter recovery before fitting a model. Verifying parameter recovery aids in interpreting the model-fitting results: if the inferred parameters seem to be insufficient in explaining user behavior, this may be attributed to modeling considerations other than the tasks in which the data were produced (e.g., the choice of model).

In our further examination of how limitations in properties of the tasks may lead to inadequate parameter recovery, we generated artificial data containing tasks with restrictions similar to what naturalistic datasets can exhibit. Of relevance for this paper were restrictions to the range of observations, outcomes, and outcome probabilities of risky-choice problems in which a user has to choose between two lotteries where each has two potential outcomes. Restricting these properties led to, in most cases, inadequate parameter recovery. Examining the game data in light of these observations suggests that parameter recovery failed for a large proportion of the users since the players were given tasks with a very limited range of outcome probabilities.

The implementation chosen displays certain limitations that could be addressed in future work. We used point estimates obtained via maximum likelihood estimation in assessing the recoverability of model parameters. Also, hierarchical Bayesian approaches [20], in which the inference at population level is handled at the same time as the inference at the level of individuals, could contribute to a more reliable inference process. In addition, future work could extend the proposed technique by exploring assessment criteria other than the Pearson correlation coefficient.

6 CONCLUSION

Theory-based models provide a good foundation for user models, thanks to their interpretability. However, theory-based inference is only as good as the data, since these models are often developed in controlled conditions. Attention should be given to the tasks users are presented with in naturalistic settings, on account of the lack of experimental control. Hence, we have proposed a parameter recovery technique for assessing how suitable a dataset is for a given theory-based model. The technique can increase (or decrease) confidence in parameter-inference results. Its results for a real-world game, which showed a strikingly low rate of application for the two models, lead us to recommend always checking parameter recovery before application in naturalistic datasets.

ACKNOWLEDGMENTS

This work was supported by Business Finland (MINERAL project), the Finnish Center for Artificial Intelligence (FCAI) and Academy of Finland projects Human Automata (Project ID: 328813) and BAD (Project ID: 318559). We would also like to thank Next Games for their assistance.

REFERENCES

[1] Robert B. Allen. 1997. Mental Models and User Models. In *Handbook of Human–Computer Interaction* (2nd ed.), Marting G. Helander, Thomas K. Landauer, and Prasad V. Prabhu (Eds.). North-Holland, Amsterdam, 49–63. https://doi.org/10.1016/B978-044481862-1.50069-8

[2] John R. Anderson, Michael Matessa, and Christian Lebiere. 1997. ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention. *Human–Computer Interaction* 12, 4 (1997), 439–462. https://doi.org/10.1207/s15327051hci1204_5

[3] Carol S. Aneshensel. 2013. Introduction to Theory-based Data Analysis. In *Theory-Based Data Analysis for the Social Sciences* (2nd ed.). Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781506335094

[4] Timothy Ballard, Ashley Luckman, and Emmanouil Konstantinidis. 2020. How meaningful are parameter estimates from models of inter-temporal choice? https://doi.org/10.31234/osf.io/mvk67

[5] David M. Blei. 2014. Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annual Review of Statistics and Its Application* 1, 1 (2014), 203–232. https://doi.org/10.1146/annurev-statistics-022513-115657

[6] David D. Bourgin, Joshua C. Peterson, Daniel Reichman, Thomas L. Griffiths, and Stuart J. Russell. 2019. Cognitive Model Priors for Predicting Human Decisions. In *Proceedings of the 36th International Conference on Machine Learning*. International Machine Learning Society, Long Beach, CA., 8984–8992.

[7] Stephen B. Broomell and Sudeep Bhatia. 2014. Parameter Recovery for Decision Modeling Using Choice Data. *Decision* 1, 4 (2014), 252–274. https://doi.org/10.1037/dec0000020

[8] Stephen B. Broomell, Sabina J. Sloman, Leslie M. Blaha, and Julia Chelen. 2019. Interpreting Model Comparison Requires Understanding Model–Stimulus Relationships. *Computational Brain & Behavior* 2, 3–4 (Dec. 2019), 233–238. https://doi.org/10.1007/s42113-019-00052-z

[9] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* 16, 5 (Sept. 1995), 1190–1208. https://doi.org/10.1137/0916069

[10] Li Chen, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, Francesco Ricci, and Giovanni Semeraro. 2013. Human Decision Making and Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 3, 3, Article 17 (Oct. 2013), 7 pages. https://doi.org/10.1145/2533670.2533675

[11] Diana Cole. 2020. *Parameter Redundancy and Identifiability.* CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781315120003

[12] Paul De Bra. 2017. Challenges in User Modeling and Personalization. *IEEE Intelligent Systems* 32, 5 (2017), 76–80. https://doi.org/10.1109/MIS.2017.3711638

[13] Andrew Gelman. 2004. Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics* 13, 4 (2004), 755–779. https://doi.org/10.1198/106186004X11435

[14] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. Bayesian Workflow. https://doi.org/10.48550/ARXIV.2011.01808

[15] Robert L. Goldstone and Gary Lupyan. 2016. Discovering Psychological Principles by Mining Naturally Occurring Data Sets. *Topics in Cognitive Science* 8, 3 (2016),

548–568. https://doi.org/10.1111/tops.12212

[16] Yoram Halevy. 2007. Ellsberg Revisited: An Experimental Study. *Econometrica* 75, 2 (2007), 503–536. https://doi.org/10.1111/j.1468-0262.2006.00755.x

[17] Andrew Heathcote, Scott D. Brown, and Eric-Jan Wagenmakers. 2015. An Introduction to Good Practices in Cognitive Modeling. In *An Introduction to Model-based Cognitive Neuroscience*, Birte U. Forstmann and Eric-Jan Wagenmakers (Eds.). Springer, New York, NY, 25–48. https://doi.org/10.1007/978-1-4939-2236-9_2

[18] J. Hemphill. 2003. Interpreting the Magnitudes of Correlation Coefficients. *The American Psychologist* 1, 58 (2003), 78–79. https://doi.org/10.1037/0003-066X.58.1.78

[19] Anthony Jameson. 2012. Choices and Decisions of Computer Users. In *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition*, Julie A. Jacko (Ed.). CRC Press, New York, NY, 77–91.

[20] Håkan Nilsson, Jörg Rieskamp, and Eric-Jan Wagenmakers. 2011. Hierarchical Bayesian Parameter Estimation for Cumulative Prospect theory. *Journal of Mathematical Psychology* 55, 1 (2011), 84–93. https://doi.org/10.1016/j.jmp.2010.08.006

[21] Robert Nishihara, Thomas Minka, and Daniel Tarlow. 2013. Detecting Parameter Symmetries in Probabilistic Models. https://doi.org/10.48550/ARXIV.1312.5386

[22] Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. 2021. Using Large-Scale Experiments and Machine Learning to Discover Theories of Human Decision-making. *Science* 372, 6547 (2021), 1209–1214. https://doi.org/10.1126/science.abe2629

[23] Drazen Prelec. 1998. The Probability Weighting Function. *Econometrica* 66, 3 (1998), 497–527. https://doi.org/10.2307/2998573

[24] Jörg Rieskamp. 2008. The Probabilistic Nature of Preferential Choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34 (Dec. 2008), 1446–1465. https://doi.org/10.1037/a0013646

[25] Carlos Pereira Santos, Kevin Hutchinson, Vassilis-Javed Khan, and Panos Markopoulos. 2019. Profiling Personality Traits with Games. *ACM Trans. Interact. Intell. Syst.* 9, 2–3, Article 11 (March 2019), 30 pages. https://doi.org/10.1145/3230738

[26] Sabina J. Sloman, Stephen B. Broomell, and Timothy Kusuma. 2020. Diagnosing pervasive issues with parameter estimation. https://cognitivesciencesociety.org/cogsci20/papers/0467/index.html

[27] Henry Stott. 2006. Cumulative Prospect Theory's Functional Menagerie. *Journal of Risk and Uncertainty* 32 (March 2006), 101–130. https://doi.org/10.1007/s11166-006-8289-6

[28] Olivier Toubia, Eric Johnson, Theodoros Evgeniou, and Philippe Delquié. 2013. Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters. *Management Science* 59, 3 (March 2013), 613–640. https://doi.org/10.1287/mnsc.1120.1570

[29] Amos Tversky and Daniel Kahneman. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty* 5, 4 (1992), 297–323. https://doi.org/10.1007/978-3-319-20451-2_24

[30] Don van Ravenzwaaij, Gilles Dutilh, and Eric-Jan Wagenmakers. 2011. Cognitive Model Decomposition of the BART: Assessment and Application. *Journal of Mathematical Psychology* 55, 1 (2011), 94–105. https://doi.org/10.1016/j.jmp.2010.08.010

[31] John Von Neumann and Oskar Morgenstern. 1990. The Notion of Utility. In *Theory of Games and Economic Behavior* (3rd ed.). Princeton University Press, Princeton, NJ, 15–31. original-date: 1944.

[32] G. I. Webb, M. J. Pazzani, and D. Billsus. 2001. Machine Learning for User Modeling. *User Modeling and User-adapted Interaction* 11 (2001), 19–29. https://doi.org/10.1023/A:1011117102175

[33] Robert C. Wilson and Anne G. E. Collins. 2019. Ten Simple Rules for the Computational Modeling of Behavioral Data. *eLife* 8 (Nov. 2019), Article e49547. https://doi.org/10.7554/eLife.49547