



HAL
open science

Decoupling Judgment and Decision Making: A Tale of Two Tails

Başak Oral, Pierre Dragicevic, Alexandru Telea, Evanthia Dimara

► **To cite this version:**

Başak Oral, Pierre Dragicevic, Alexandru Telea, Evanthia Dimara. Decoupling Judgment and Decision Making: A Tale of Two Tails. IEEE Transactions on Visualization and Computer Graphics, 2023, 10.1109/TVCG.2023.3346640 . hal-04354869v2

HAL Id: hal-04354869

<https://hal.science/hal-04354869v2>

Submitted on 23 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Decoupling Judgment and Decision Making: A Tale of Two Tails

Başak Oral, Pierre Dragicevic, Alexandru Telea, and Evanthia Dimara

Abstract—Is it true that if citizens understand hurricane probabilities, they will make more rational decisions for evacuation? Finding answers to such questions is not straightforward in the literature because the terms “*judgment*” and “*decision making*” are often used interchangeably. This terminology conflation leads to a lack of clarity on whether people make suboptimal decisions because of inaccurate judgments of information conveyed in visualizations or because they use alternative yet currently unknown heuristics. To decouple judgment from decision making, we review relevant concepts from the literature and present two preregistered experiments (N=601) to investigate if the task (judgment vs. decision making), the scenario (sports vs. humanitarian), and the visualization (quantile dotplots, density plots, probability bars) affect accuracy. While experiment 1 was inconclusive, we found evidence for a difference in experiment 2. Contrary to our expectations and previous research, which found decisions less accurate than their direct-equivalent judgments, our results pointed in the opposite direction. Our findings further revealed that decisions were less vulnerable to status-quo bias, suggesting decision makers may disfavor responses associated with inaction. We also found that both scenario and visualization types can influence people’s judgments and decisions. Although effect sizes are not large and results should be interpreted carefully, we conclude that judgments cannot be safely used as proxy tasks for decision making, and discuss implications for visualization research and beyond. Materials and preregistrations are available at https://osf.io/ufzp5/?view_only=adc0f78a23804c31bf7dd9385cb264f.

Index Terms—Cognition, Decision Making, Judgment, Psychology, Visualization



1 INTRODUCTION

Imagine a user browsing a weather app on their phone that shows forecast data. They ask themselves two questions: *i)* How likely is it that there will be heavy rainfall this afternoon? *ii)* Should I carry an umbrella on my way to the doctor’s appointment? These questions seem related, so it is fair to assume that our *judgment* of the forecast probability should directly influence our *decision* to carry an umbrella. Yet, even if we assume that everyone can correctly derive the probability estimates from the visualization, other factors, such as our bag size, the umbrella weight, or our tolerance to mild rainfall, may influence our decision. More life-threatening decisions, such as if to evacuate a city based on the forecast of a hurricane strike or wildfire, are more arduous and suffer from cognitive biases [1]. Stressful situations can further introduce time pressure leading to inaccurate assessment of the visualized data [2].

It is rather unclear if visualization users make suboptimal decisions because they perform inaccurate judgments of probabilities, or they use alternative yet currently unknown heuristics. More research is needed to explain *why* people make seemingly suboptimal decisions. One barrier to answering this question is that researchers tend to use the terms judgment and decision interchangeably. For example, in a visualization study, participants were asked to decide, as Red Cross managers, on issuing blankets to help alpacas survive harsh weather conditions [3]. Although the task was a decision, the interpretation of task accuracy was associated

with the ability of participants to judge the probability of temperatures falling below 32 °F. Other experiments, while aiming to focus on studying human decisions, exposed participants to visual judgments, such as finding the maximum average bar length [4] or finding a growth trend [5]. The notions of judgment and decision making are often conflated in a single task which sometimes measures people’s decision performance and other times their ability to judge numerical estimates. More importantly, the initial goal of the study is not always in line with the underlying task, which makes the results hard to interpret.

To some extent, this conflated use of judgments and decisions is justifiable – a judgment is a fact-based question that is easier to test than a decision. Yet, recent research gives preliminary evidence that people may need to use different visualizations for judgment and decision tasks. Kale *et al.* [6] presented participants with four different visualizations (quantile dotplots, density plots, HOPs, intervals), in a fantasy sports game. The task was to judge how probably a participant’s team would score with or without a new player, as well as to decide whether to pay for the new player or not. They found that, while participants elicited the best decisions with densities and intervals, they judged better with quantile dotplots. Although the goal of the paper was not to compare judgments with decisions, the fact that the best visualization for judgment did not necessarily lead to better decisions reveals the need to decouple these two tasks. So, to clarify the confusion of judgment and decision making, it is important to separate judgment and decision tasks and to investigate to what extent they relate to each other.

In this paper, we study the interplay between judgment and decision making in the context of data visualization. We first review how judgment and decision making have been conceptualized and studied in both visualization (Sec. 2) and

- B. Oral was with Utrecht University.
E-mail: e.oral@uu.nl
- P. Dragicevic was with Inria Bordeaux.
- A. Telea, and E. Dimara were with Utrecht University.

Manuscript received January 10, 2023; revised September 29, 2023.

psychology research (Sec. 3). We next empirically study and discuss to what extent a judgment task (*i.e.*, picking the optimal alternative based on expected values estimated visually) is a sufficient behavior predictor in a decision making task (*i.e.*, to buy (or not) a player in fantasy sports). We examine judgments and decisions in-between conditions by casting participants in a series of either judgment or decision tasks with different visualizations (quantile dotplot, density plot, probability bar) and scenarios (sports, humanitarian). All our experimental material, data, analyses, and preregistrations can be found at the OSF link in the abstract.

2 RELATED WORK

2.1 Judgment in Visualization

Visualization is used for decades to support various tasks that involve human *judgments* for data exploration, analysis, and presentation. Judgment determines the need to build visualization systems: “*If a fully automatic solution has been deemed to be acceptable, then there is no need for human judgment, and thus no need for you to design a vis tool.*” [7].

Visual judgments have been studied at the lower levels of processing, aiming to understand what people *perceive* when looking at a visualization [8]. Examples include judging color discriminability [9], line orientation [10], contrast [11], motion [12], groupings [13], average positions [14], extrema [15], and correlation patterns [16].

Higher-level judgments have been studied when looking at how users make sense of the displayed data – that is, how they interpret the visualization to come to conclusions about the shown data [7], [17]. Examples include probability judgments using intervals, density plots, quantile dotplots, and tables, which show sampled distributions of one [3], [6], [18] or several attributes [19]; or more complex tasks such as assessing causality [20], missing data [21], information credibility [22] or sensemaking [23].

Human judgments are further studied in the context of critical applications of visualizations, like medical diagnosis [24] and weather forecasts [25]. Other judgment task examples include gauging the maintainability of a software system by visualizing its dependency graph [26]; assessing traffic congestion of vessel fleets by visualizing their movements over space and time [27], and assessing the group structure of multidimensional samples to infer how easily classifiable a dataset is by machine learning [28].

One can easily argue that most data visualizations involve a form of judgment – indeed, if users were not able to, or interested in, making judgments about the depicted data values therein, that visualization would be useless.

2.2 Decision Making in Visualization

Following some authors, a key goal of visualization is to help “*the decision maker to discover what should be said and done*” [29], and that decision making is the most important reason of “*why visualization is important*” [30]. Visualization research invested much effort to study how users make decisions with visualized data [1], [31]. Decision tasks that have been studied include binary choices [6], such as to perform a humanitarian action or not [3] or *multi-attribute choice tasks* between many options [32]. Other examples are choosing a

holiday hotel based on price, hotel quality, landscape interest, and security level [32] or choosing an investment based on company traits such as leadership ability, proprietary technology, market conditions, and competitor strength [19]. More examples of multi-attribute choice tasks cover time-interval choices (*e.g.* when to arrive to catch a bus [33]) and group decision making [34]. Decision making has been further studied for critical applications such as finance [35], hiring [36], and engineering optimization decisions [37].

Decision making is less studied than judgment in visualization. A recent review showed that decision tasks were involved in only 6% of quantitative and 4% of qualitative evaluations of visualization tools for decision support, further attributed to the lack of relevant guidance from decision theory in visualization literature [38]. Decision tasks are also omitted by visualization task taxonomies, including low-level and high-level ones [38]. Most real-world visualization applications involve domain experts who make judgments with data and almost never decision makers who make decisions based on data [39]. Visualizations seemed to be used mostly to communicate a small part of the decisions already made by humans [39] or by AI systems [40], [41]. Two recent reviews on the use of visualization for explainable AI (XAI) [41], [42] concluded that researchers study how to reveal the inner workings of a model (operations and outputs) and not how end-users ultimately involve the model in their decision making process. Meanwhile, the analysis of 940 tools reported by professional decision makers who described their work practices revealed the lack of a decision making tool that can support them through all steps of their decision making process, [39].

2.3 Conflating Judgment and Decision Making

One barrier to studying decision making effectively is that judgments and decisions are terms not well-defined and separated. Visualization papers almost never formally define, or even casually describe, decision making [38]. Some authors consider decision making as a subpart of the high-level judgment task of sensemaking, [43]; other authors see decision making as a broad expansive task that often contains a sensemaking subtask [38]; yet other authors view decision making and sensemaking as distinct tasks which should be supported by different visual analytics tools [44]. Visual analytics literature often uses the term decision making to refer to algorithmic and not human decisions [45].

This terminology confusion becomes more evident in empirical studies where the terms decision and judgment are used interchangeably. For example, participant responses have been interpreted as “decisions”, though participants performed low-level visual judgment tasks, like spotting the higher bar between two alternatives [15] and spotting the bar with the maximum average length between two bar sets [4]. When investigating if hypothetical outcome plots (HOPs) can facilitate people’s judgment of trends, participants were asked to play the role of a newspaper editor whose job was to make a “decision” on the headline of a growth trend in the job market [5]. The term decision was again used to refer to a visual judgment, although in that case, the confusion could have been influenced by the response type format – the so-called “two-alternative forced choice” (2AFC)– and not by

the nature of the task. 2AFC response formats though can be used for both judgment and decision tasks without changing the nature of the task. In another experiment, participants were asked to perform a decision task – to issue (or not) blankets to protect alpacas from cold [3]. Yet, responses were this time interpreted as the ability of participants to make “judgments” on the probability of temperatures falling below 32 °F [3]. This conflation of terminology can also explain why the vast majority of visualization tools designed for decision support used solely judgment tasks in their evaluation [38].

We are not aware of visualization studies that explicitly compare how people perform judgments and decisions. Two studies give some evidence on whether people are more accurate with judgments or decisions, as side notes within different experimental goals: Kale *et al.* [6] found that judgments and decisions can be better supported by different visualizations. Quantile dotplots supported better the judgment of the probability of superiority; interval and density plots supported better decision accuracy. Dimara *et al.* [46] studied how various narratives affect response accuracy, asking participants to select a house in a scatterplot based on price and size. Participants performed the task as a real estate agent, providing a judgment for their client; or as decision makers deciding which house to buy. Results showed that decision narratives elicited less accurate responses than judgment narratives. Interpreting this finding is hard. Both tasks were factually identical; an accurate response was a house that is not dominated by other houses that are both cheaper and bigger, and this accuracy criterion was explicitly communicated to both groups. In a realistic task, one possible interpretation could be that the decision group had subjective preferences of other salient features of the house, *e.g.*, visual appearance. Yet, in this study, no other information was given besides size and price. Findings of the above two studies suggest that factors such as the visualization type or the task framing can elicit performance differences in judgment and decision making.

To conclude, judgment and decision making are two terms conflated and two tasks which we do not yet know how they relate to each other.

3 JUDGMENT AND DECISION MAKING IN PSYCHOLOGY

There is often confusion about what to call judgment or decision (see Sec. 2). To clarify this, we discuss definitions of judgment and decision making tasks in Sec. 3.1; we review empirical studies from psychology and economics in Sec. 3.2.

3.1 Judgment & Decision Making Definitions

The American Psychological Association (APA) dictionary of psychology defines judgment as “*the capacity to recognize relationships, draw conclusions from evidence, and make critical evaluations of events and people; or the ability to determine the presence or relative magnitude of stimuli.*” Meanwhile, they define decision making as “*the cognitive process of choosing between two or more alternatives, ranging from the relatively clear cut to the complex.*” [47]. Extending the APA definitions, statistical decision theory [48] posits that both judgments and subsequent decisions involve accounting for a certain level of uncertainty. However, unlike decisions, judgments do not require individuals to choose among alternatives [49].

Consistent with the APA definition, Diederich and Busemeyer [50] classify five judgment categories: (1) drawing conclusions from evidence (“Will you be in class on time?”; “What are the chances that you will get the attractive position you applied for?”); (2) making critical evaluations (“How much do you actually like that class?”); (3) value judgments (“A is interesting, beautiful, or better than or similar to B”); (4) category judgment (judging class membership), and (5) probability or quantity estimation. Meanwhile, Fischhoff and Broomell [51] define decision making as a process that consists of three parts: (1) judgment, *i.e.*, the prediction of outcomes, (2) preference, *i.e.*, how people weigh those outcomes, and (3) choice, *i.e.*, how people combine judgments and preferences to make a decision. In line with Fischhoff and Broomell’s definition [51], in visualization research, Dimara and Stasko [38], and Oral *et al.* [52], argue that, unlike a judgment, a decision task should at the very least include a choice stage, following Simon’s decision stages [53].

Despite the operational clarification on the need of a choice stage, the underlying mechanisms differentiating judgment from decision making remain elusive. Eberhard [49] concurs that decision making implies choosing actions with subsequent consequences, unlike judgment, which assesses alternatives without an obligation to act. Neuroimaging studies support this view by revealing activation in motor regions of the brain during decision tasks [54]. This view is further reinforced by interviews with decision makers and analysts, as a participant summarizes: “A decision maker has to live with their decision while an analyst can just say what the best thing is and walk away!” [39]. It is also shown that being an actor (*vs.* an observer) enhances the sense of control and induces responsibility, attributes more pronounced in decision making [55]. Beyond these factors, decisions pertain to the future [56], involve past experiences and personal identity [56], and stronger emotions [57] than judgments.

In summary, while decision making shares similarities with judgment, it embodies four distinguishing features: (I) it requires a choice among alternatives, implying a loss of the remaining alternatives, (II) it is future-oriented, (III) it is accompanied with overt or covert actions, and (IV) it carries a personal stake and responsibility for outcomes. The more of these features a judgment has, the more “decision-like” it becomes. When a judgment has all four features, it no longer remains a judgment and becomes a decision. This operationalization offers a fuzzy demarcation between judgment and decision making, in the sense that it does not draw a sharp line between the two concepts, but instead specifies the attributes essential to determine the extent to which a cognitive process is a judgment, a decision, or somewhere in-between [58], [59]. We will use this operationalization in the rest of our article.

3.2 Judgment & Decision Making Studies

When designing experiments, psychology researchers appear more consistent with the aforementioned definitions when using the terms judgment and decision. For instance, cultural psychology explicitly identifies judgment as the evaluation of the source of one’s behavior (*i.e.*, causal attribution), and it identifies as a decision whether to cooperate with or compete against a party (*i.e.*, conflict decision) [60]. However, we also observed that some psychology papers confuse

judgment with decision making, just as in visualization research. For example, van Norman *et al.* [61] identified the visual judgment of a trend on whether an intervention improved students' performance or not as a decision.

As noted in Sec.2.3, with its inherent *choice* framing, the use of AFC format sometimes leads researchers to categorize judgment tasks as decision tasks. However, in psychophysics [62], AFC is commonly used for low-level perceptual judgments like reporting the brightest stimuli [63]. Conversely, in behavioral economics, AFC is used for making decisions under uncertain conditions [64]. Therefore, unlike the factors discussed in Sec.3.1, the AFC format alone should not define the task as judgment or decision making.

The psychology literature illuminated our discussion providing distinct definitions and more consistent use of the terms judgment and decision making. The domain of decision making under risk [65] further provides insights into how to measure judgment and decision accuracy. This domain jointly studies judgments and decisions. Researchers studied judgment by measuring people's understanding of risks, *e.g.*, perceived risk probabilities of different options [66], conjointly with decision making by measuring risk behavior, *e.g.*, willingness to pay for a product that reduces the risk of getting a disease or injury [66], [67]. Findings suggest that people tend to pay more for risk-reducing products when the risk is shown by icon arrays instead of numerical formats, claiming that icon arrays twist risk perception by highlighting the people at risk [68]. In contrast, other studies showed icon arrays to improve judgments of health risk [69]. On the other hand, Wu *et al.* [70] showed that making accurate probability judgments did not improve search decisions. Likewise, although people can judge doing something as morally wrong, they may decide to behave in a way that is not consistent with their judgment [55].

Although judgment and decision accuracy have not been explicitly contrasted in any of these studies, we saw that is plausible that judgments can influence decisions in all possible directions. We saw accurate judgments which improve or do not improve decisions, as well as inaccurate judgments that even lead to better decisions. It thus remains unclear if judgment accuracy can guarantee better subsequent decisions. One possible confounding factor in all previous studies is that judgments and decisions were tested in within setups. So, we cannot exclude the possibility that asking a person to perform a judgment before a decision can itself influence the decision behavior.

To conclude, we are still not aware of any attempt that explicitly studies the difference in response accuracy between judgment and decision making.

4 RESEARCH HYPOTHESES

To empirically investigate the interplay between judgment and decision making, we formulated the following hypothesis which was preregistered prior to data collection at the OSF link given at the end of the abstract.

Hr : Performance on a judgment task with a data visualization is not a good proxy for performance with the corresponding decision task with the same data visualization.

Hypothesis Hr was assessed through 5 sub-hypotheses:

- Hr₁** : Decision accuracy overall differs from judgment accuracy. This hypothesis does not assume directionality because there are conflicting trends in related works (discussed in Secs.2.3 and 3).
- Hr₂** : Decision accuracy and judgment accuracy differ in their sensitivity to the underlying scenario. This hypothesis is motivated by work suggesting that the prospective death of an animal elicited irrational responses in decision accuracy (*e.g.*, always issuing blankets to the Alpacas regardless of the weather forecast probabilities) [3].
- Hr₃** : There is a specific visualization where decisions elicit different accuracy than judgments.
- Hr₄** : There is a specific visualization and scenario where decisions elicit different accuracy than judgments.
- Hr₅** : The decision task differs from the judgment task in its vulnerability to the sub-optimal heuristic strategies "risk-aversion" and "status-quo" biases. This hypothesis is motivated by research showing decisions are hindered by loss aversion and avoidance of changing the person's current state (status-quo) [31].

Our hypotheses were hierarchically structured. The judgment task is not expected to be a sufficient proxy for the decision task [Hr] because either of the following: (a) decisions elicit overall different accuracy from judgments [Hr₁]; (b) decision accuracy overall differs from judgment accuracy in its sensitivity to the underlying scenarios [Hr₂]; (c) decision accuracy is better facilitated by different visualizations [Hr₃]; (d) there is at least a specific visualization and scenario where decisions elicit different accuracy from judgments [Hr₄]; (e) decision task differs from judgment task in its vulnerability to suboptimal heuristic strategies, *e.g.*, risk-aversion and status-quo biases [Hr₅]. Support for at least one of the sub-hypotheses [Hr₁ - Hr₅] provides support for the main hypothesis [Hr]. However, the larger the number of the supported sub-hypotheses are, the stronger the understanding of the nature of unsuitability of a judgment task as a proxy for decision task will be. Before we provide the statistical hypotheses (Sec.6.7) which map to the above research hypotheses, we detail our experiment design.

5 DESIGN RATIONALE: JUDGMENT VS. DECISION

To investigate **Hr**, we assigned participants to two tasks: judgment and decision. We next outline our design choices for these tasks based on insights from Secs.2.3, 3.1, and 3.2.

5.1 Differences Between Tasks

Question Type: As illustrated in Fig. 2, the sole distinction between the two tasks lay in their respective *framings*: For the judgment task, we asked the question "What is the best option?" (an observation framing); for the decision task, we asked "What do you choose?" (an action-oriented framing). This framing manipulation is almost identical to the narrative study by Dimara *et al.* [46] discussed in Sec.2.3, which designated participants as either real-estate analysts who provide judgments or house buyers who make decisions. Drawing from the insights in Sec.3.1, we hypothesized that the action-oriented framing (decision

feature **III**) of the decision task would elicit a heightened sense of control, responsibility (decision feature **IV**), and consequently emotion, while the judgment framing would position participants more as analytical observers.

5.2 Identical Design Choices Between Tasks

In order to ensure validity in comparing the judgment and decision tasks and minimize confounding variables, we maintained uniformity in the remaining parameters:

Response Type: Both tasks employed the AFC format. As mentioned in Sec. 3.2, this format is apt for both judgment-related psychophysics and economic decision making. Unlike the binary choices (2AFC) used in studies [3], [6], we adopted 5AFC to increase the resolution of the accuracy score, which would otherwise be limited to 0 or 1 (a 50% chance to identify the best outcome by pure chance). Both tasks featured identical option sets: one optimal, one risk-averse, one status-quo (Hr_5), and two alternative options, detailed in Sec. 6.3.

Response Accuracy: We measured the accuracy of both tasks using the same expected value metric, as detailed in Sec. 6.6.

Task Context: All participants viewed three visualizations (Hr_3) across two scenarios (Hr_2): one on sports and another, emotionally charged, on children in war. We designed these scenarios to maintain a consistent emotional influence in both tasks, regardless of question framing. See Sec. 6.2 for scenario details and Sec. 6.1 for the visualizations.

Incentives and Instructions: We compensated all participants and encouraged “as accurate as possible” responses. While many experimental designs, particularly in economics, introduce extrinsic motivation through performance-based incentives to emulate real-life decisions, we designed our scenarios to foster intrinsic motivation, consistent with psychology study methodologies. We opted against performance-based incentives due to concerns raised in [6] preventing disparities between the judgment and decision tasks that might skew our comparative analysis. Guiding participants explicitly toward a “correct” way of answering would counter our objective of capturing innate strategies in judgments and decisions. In real-world situations, the distinct incentive structures for genuine decisions and judgments would make direct comparisons even more challenging. Thus, although we trained participants on probabilities, visual interpretation, and cost-profit trade-offs, we refrained from priming them with specific “correctness” benchmarks or linking them to incentives other than the ones implied in the task framing.

6 EXPERIMENT 1

This experiment tested our hypothesis Hr via sub-hypotheses $Hr_1 - Hr_5$. We exposed participants to visualizations: quantile dotplot, density plot, and probability bar, asking questions involving probability and cost/profit estimations. We examined the influence of the task (judgment task *vs.* decision task), scenario (humanitarian *vs.* sports), and visualization on response accuracy. All materials and preregistration details are at the OSF link in the abstract.

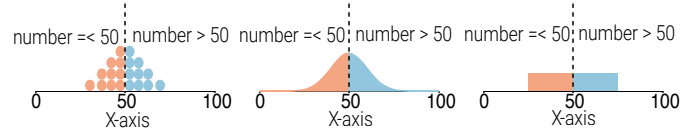


Fig. 1. The three visualizations used in both experiments: quantile dotplot (left), density plot (middle), and probability bar (right).

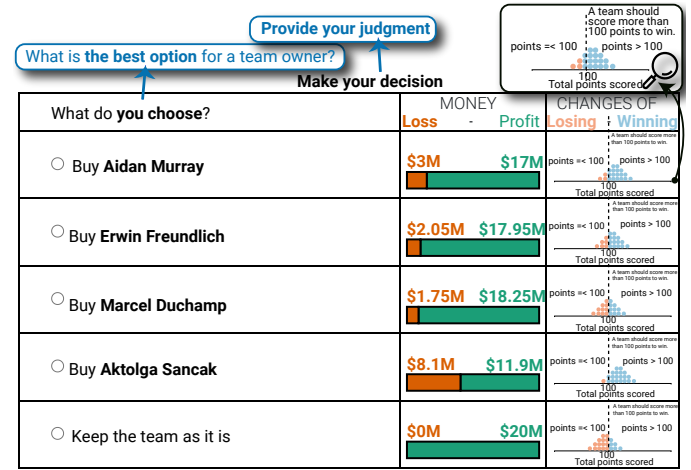


Fig. 2. Example of the decision task condition with the sports scenario and quantile dotplot visualization. The judgment task version was identical, except for modifications in the title and question, highlighted in blue.

6.1 Visualizations

We utilized three uncertainty visualizations: quantile dotplot, density plot, and probability bar, as shown in Fig. 1. Our choices were based on studies that explored decision making and probability estimates with uncertainty visualizations [3], [6]. Unlike [3], [6], which used intervals, we opted for probability bar visualization. While the dot count in quantile dotplot and area in density plot convey probability information, interval bar widths don’t directly represent probabilities but denote 95% of possible outcomes. We therefore adapted the interval’s horizontal bars to incorporate exact probabilities, ensuring they align in width with quantile dotplot and density plot.

6.2 Scenarios and Stimuli

We showed two scenarios, sports and humanitarian (Hr_2). In the sports scenario, adapted from Kale *et al.*, [6], a team owner can buy players to create a team. Each player has a cost (shown with red bars in Fig. 2), but if they help their team to win, the team gets a profit (shown with green bars in Fig. 2). The winning probability of a player, displayed in blue in Fig. 1, is contingent upon a certain threshold of points the team needs to score for victory (*e.g.*, 100). The decision task was to choose which player to buy (Fig. 2). The judgment task was to indicate which player is the best for the team owner (Fig. 2 blue highlights). Both tasks had 5 alternative responses, 4 different players, or the option to do nothing (for Hr_5). The best response was the one that maximizes the expected value (see Sec. 6.6). Colors were chosen with ColorBrewer to ensure safety for color-blind viewers.

For the second scenario, we sought a more affective context than sports, such as aiding Alpacas in Castro *et al.* [3] and humanitarian visualizations [71]. However, it was essential that the storyline supports the same components as the sports scenario, including 5 alternatives, cost-profit

considerations, probability trade-offs, and a do-nothing option. We adopted a narrative involving a war between two fictitious countries, Syldavia and Borduria (names from the Adventures of Tintin). Orphans in Syldavia need essential goods. Syldavia can either wait for aid from neighboring countries without cost (do-nothing option) or fund one of four humanitarian organizations to help the orphans. Each organization has a transportation fee, charged from Syldavia's limited resources. If they succeed in delivering, the kids receive the goods (profit). Each organization has its own delivery probability before borders close, as each player has varied the probability of the win in Fig. 2. If unsuccessful, Syldavia's money is wasted. Again, the decision was to choose one of the 4 organizations or do nothing, and the judgment was to indicate the best option for Syldavia. Except for text, the stimulus of humanitarian was identical to the sports scenario. Detailed experiment instructions are in the supplementary materials.

6.3 Dataset

We generated synthetic data for probability and cost/profit information. We selected 33% and 90% as the probability endpoints; 33% and 90% were always for status-quo and risk-averse alternatives, respectively. Since the brain represents probability on the log-odds scale [72], we converted the endpoints into log-odds units. Then, we sampled on this logit-transformed scale using linear interpolation between the endpoints. We created three difficulty levels, *i.e.*, easy, medium, and hard. At the *easy* level, in addition to status quo and risk-averse alternatives, we added the best alternative and two alternatives dominated by the best alternative. While one dominated alternative had the same cost and profit value as the best alternative, its probability of winning was lower. The other dominated alternative had the same probability but more cost and less profit compared to the best. At the *medium* level, in addition to status quo and risk-averse alternatives, we added the best alternative, a second alternative, and an alternative dominated by both the best and the second alternatives. At the *hard* level, in addition to the status quo and risk-averse alternatives, we added the best alternative and two others. At the easy and medium levels, accuracy scores are 0, 0.25, 0.50, 0.75, and 1 for the risk-averse, status-quo, alternative 1, alternative 2, and best alternative, respectively. On the other hand, at the hard level, the accuracy of different alternatives changed. The best alternative and alternative 2 still have accuracy scores of 1 and 0.75. But, the status-quo, alternative 1, and risk-averse alternatives have accuracy scores of 0, 0.25, and 0.50, respectively. All cost and profit values were randomly selected and fixed considering the defined probability and accuracy values. The dataset is available at the OSF link.

6.4 Procedure

Participants completed a consent form and underwent training on the visualizations involved, density plot, probability bar, and quantile dotplot, and cost/profit bars. During the training, they tackled two types of tasks for each visualization. The first task involved identifying the alternative with the highest probability depicted by the visualization among five alternatives. The second task was selecting the probability range depicted by the visualization. Finally, we

instructed participants on how to read cost and profit bars by explicitly stating that they should consider the length of the green bar in conjunction with the probability of success and the length of the orange bar in conjunction with the probability of failure (e.g., see intro1-page74 under experiment_screens on OSF Supplementary materials)

Post-training, participants performed either judgment or decision tasks, based on their assigned condition, with all scenarios and visualization types. They then completed a demographics questionnaire, including gender, age, education level, and country of residence, an Adaptive Berlin Numeracy Test (BNT), [73], and answered two optional questions about their strategy and comments. On average, the experiment lasted approximately 30 minutes. Details on the procedure are in the supplementary materials.

6.5 Experiment Design

The experiment utilized a mixed design. Scenario (sports and humanitarian) and visualization (density plot, probability bar, and quantile dotplot) were within-subjects independent variables, ensuring participants encountered all scenarios and visualizations. The task (judgment task or decision task) was a between-subjects independent variable; thus, a participant, randomly assigned, either provided judgments or made decisions. The motivations for choosing a between-subjects design is detailed in Sec. 3.2. The order for within-subjects conditions was entirely randomized, while difficulty sequence remained fixed (easy, medium, hard as discussed in Sec. 6.3). Overall, each participant tackled 18 trials (3 visualizations x 2 scenarios x 3 difficulty levels).

6.6 Measures

We used the following measures:

Accuracy as the proportion of the difference between the expected value (EV) of a participant's choice (among five different alternatives) and the choice with the worst possible EV, relative to the difference between the choices having the best and worst possible EVs:

$$Accuracy = \frac{|EV_{choice} - EV_{worst}|}{|EV_{best} - EV_{worst}|}.$$

To illustrate this with an example: assume EVs are placed on a numerical axis. The accuracy score is calculated based on how far the EV of the chosen option is from the worst EV, scaled by the range of the numerical axis (*i.e.*, the difference between the best and worst values). In this case, if a participant chooses the best option, the accuracy score is 1 while it is 0 if the worst option is chosen. The EV value of the options in the middle (*i.e.*, the other three options) is adjusted in such a way that each option is equidistant from the adjacent ones. This results in a set of possible accuracy scores: $\{0, 0.25, 0.5, 0.75, 1\}$, as explained in Sec. 6.3.

Time of completion in seconds.

Risk Literacy with the Adaptive Berlin Numeracy Test [73].

6.7 Statistical Hypotheses

We next translate our research hypotheses (Sec. 4) to the following statistical (thus, testable) hypotheses:

H_1 : The mean accuracy score of the decision task across scenarios and visualizations is measurably different than the mean accuracy score of the judgment task.

- H₂ : The mean accuracy score of the decision task is measurably different than the mean accuracy score of the judgment task with scenario S , where $S \in \{\text{sports, humanitarian}\}$.
- H₃ : The mean accuracy score of the decision task is measurably different than the mean accuracy score of the judgment task with visualization V , where $V \in \{\text{quantile dotplot, density plot, probability bar}\}$.
- H₄ : The mean accuracy score of the decision task is measurably different than the mean accuracy score of the judgment task in scenario S and visualization V , where $S \in \{\text{sports, humanitarian}\}$, and $V \in \{\text{quantile dotplot, density plot, probability bar}\}$.
- H₅ : In the decision task, the “status quo” alternative is chosen measurably more or less frequently than it is done in the judgment task– and the same for the “risk-averse” alternative in the decision task *vs.* the judgment task. Both expectations are on “more.”

6.8 Participants

We secured ethics, privacy, and data management approval for our study from our department’s respective committees. Participants were sourced from the Prolific crowdsourcing platform (platform specifics: all countries, standard sample, prior minimum approval rate: 95%, payment: 6 pounds/hour). Our target sample size was $N = 300$ (around 150 per task) following previous studies, [3], [6]. Of 162 participants in the judgment task condition who consented, 16 quit before completing, and 4 were excessively slow (see preregistered exclusion criteria). From 171 participants in the decision task condition who consented, 13 didn’t finish. Thus, our final count was 300 participants ($N = 142$ for judgment task and $N = 158$ for decision task), with 62% male and 51% aged 18-30, representing 24 countries, primarily Europe (main countries: UK 33%, Poland 14%, South Africa 11%).

6.9 Results of Experiment 1

We planned and preregistered all analyses (see OSF link in abstract) prior to data collection. We used an *estimation approach* to statistical reporting, *i.e.*, we base all our inferences on confidence intervals reported graphically, seeing statistical evidence as lying on a continuum rather than being binary [74]–[76]. For guidance on reading graphs with confidence intervals and relate them to p-values, see [77], [78].

All our confidence intervals (CIs) are 95% BCa bootstrap confidence intervals [79]. Our CIs are not corrected for multiplicity. Thus, since we test five hypotheses and some of them break down into multiple statistical hypotheses (up to 6), any isolated finding must be taken as tentative. As a reminder, consistent with our statistical hypotheses H₁ – H₄ (Sec. 6.7), we expect to find accuracy differences between the judgment task and the decision task across scenarios and visualizations. Also, consistent with H₅, we expect the percentage of status-quo and risk aversion biases to differ across the judgment task and the decision task.

The main results of experiment 1 are shown in Fig. 3 (left column). The point estimates and CIs for differences (in the ellipses on Fig. 3) were neither planned nor preregistered and were computed later to facilitate the interpretation of statistical evidence. To further assist the interpretation of effect magnitudes, we also chose to report standardized

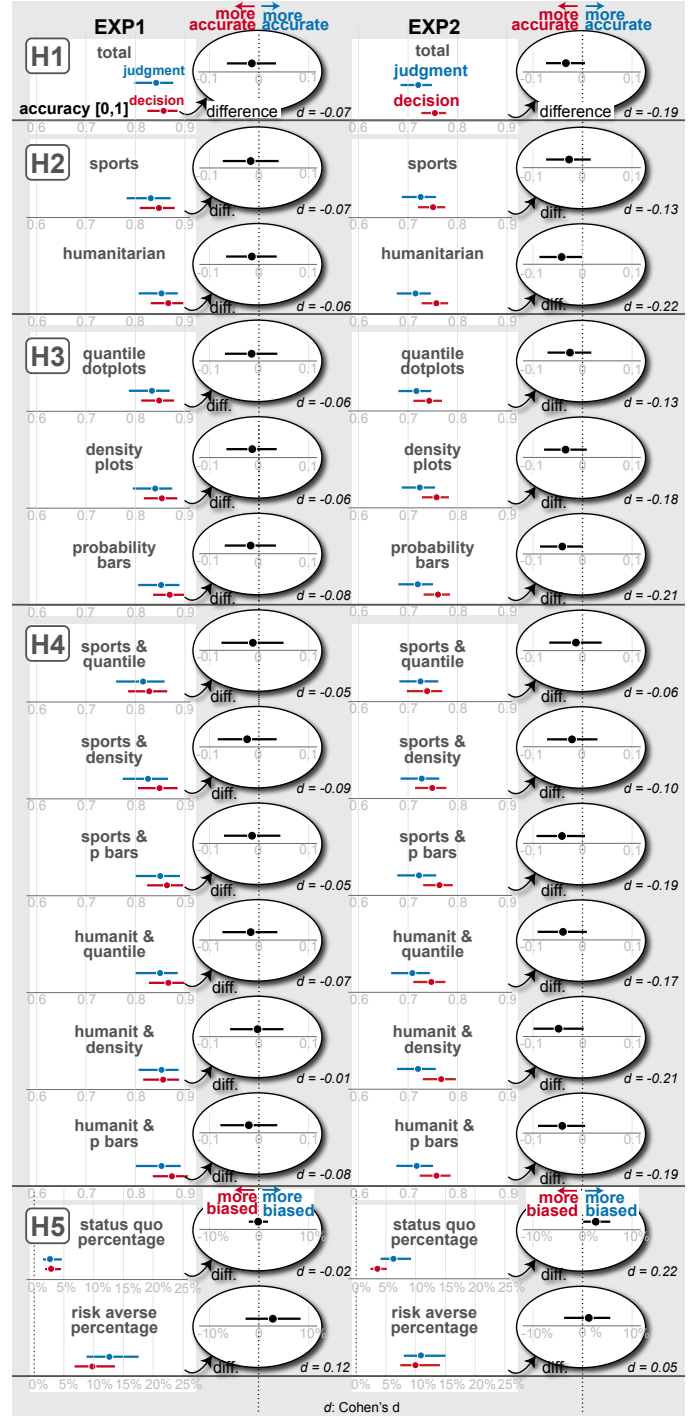


Fig. 3. Results for H₁ - H₅ showing CI of average accuracy and bias percentage per condition for experiments 1 (left) and 2 (right). Ellipses depict the difference (black CI) between average judgment (blue CI) and decision (red CI) accuracies (for H₁ - H₄) and bias percentage (H₅). Cohen’s d values for each difference are bottom-right of the ellipses. Chart titles (e.g., total, sports) indicate the specific condition.

effect sizes (Cohen’s d), of which computation was also neither planned nor preregistered.

For H₁, we compared the mean accuracy score between the judgment task and the decision task. Results are reported graphically in Fig. 3, column EXP1, row H1. We found no evidence for a difference between judgment accuracy and decision accuracy. Likewise H₁, the results were also inconclusive for hypotheses from H₂ to H₅ (see corresponding rows in Fig. 3 for visual comparison).



Fig. 4. Accuracy scores (x axis) vs. participant count (y axis) colored by response time for both experiments. Log-transformed time ranges: fast (0.52-1.43], medium (1.43-1.74], slow (1.74-3.04], corresponding to minutes: (1.68-4.18], (4.18-5.70], (5.70-20.90)].

6.10 Additional Analyses

6.10.1 Response time and task accuracy

We measured the possible correlation between response time and task accuracy to see whether the speed of responses can relate to what people decide or judge. Correlation analysis showed that there is some evidence for a positive trend between response time and both judgment accuracy, $r = 0.30$, CI [0.12, 0.44], and decision accuracy, $r = 0.18$, CI [0.03, 0.31]. Yet, point estimates for the r values were small.

6.10.2 Risk literacy and task accuracy

We examined how risk literacy, intended to capture people’s ability to accurately calculate and understand probabilities, relates to judgment accuracy and decision accuracy. We found no clear evidence between risk literacy and judgment accuracy ($r = 0.09$, CI [-0.08, 0.24]) and between risk literacy and decision accuracy ($r = 0.19$, CI [0.04, 0.33]). Yet, if there is any, it is more likely that more accurate decisions can be associated with higher risk literacy.

6.11 Summary of Experiment 1

The results were inconclusive for all our hypotheses on whether there is a performance difference (or not) between conditions. We found no evidence of judgment accuracy-decision accuracy differences across scenarios, visualizations, or any specific scenario-visualization pair. Also, we found no evidence for status-quo and risk-aversion bias percentage differences between conditions.

7 EXPERIMENT 2

The lack of observed judgment accuracy – decision accuracy difference in experiment 1 could be, we believe, a ceiling effect, as many accuracy scores were equal to 1 (see raw accuracy scores, Fig. 4 top). Hence, we conducted a second experiment increasing task difficulties [C1]. Given participants’ comments as “Yes, the study wasn’t randomized so after a while it was easy to remember which options I chose previously

even though they were shuffled around or the scenario was changed” and “The study was enjoyable, but I would say I found the sections checking understanding at the start were a bit longer than I think they needed to be”, we increased the options’ variety across trials [C2] and shortened the tutorial length [C3].

7.1 Scenarios, Tasks and Stimuli

Experiment 2 was identical to experiment 1, including hypotheses and design, except for the changes described next. Scenarios, tasks, and stimuli were identical except for a few small but important changes in experiment instructions: To ensure that the judgment and decision framing did not escape participants’ attention, we emphasized it further by changing “What is the best option for Syldavia?” to “Consider that this is your judgment. What is the best option for Syldavia?”, and “What is the best option for a team owner?” to “Consider that this is your judgment. What is the best option for a team owner?” for the judgment task. We also changed “What do you choose?” to “Consider that this is your decision. What do you choose?” for the decision task. Also, we changed “Next you will be asked to provide judgments.” to “Next you will be asked to provide your own judgments.” and “Next you will be asked to make decisions.” to “Next you will be asked to make your own decisions.” To avoid shortening the experiment duration if participants felt they needed to make computational calculations on the side, we also clarified that “This experiment is not a math test. No calculations are required on your side, but your answer should reflect your judgment/decision as fast and accurately as possible.”

7.2 Dataset

The dataset generation algorithm was identical to experiment 1 (Sec. 6.3), including difficulty levels (easy, medium, hard). However, there was no option with the same probability or cost/profit value, making the comparison of options harder than experiment 1 [C1]. Also, the best alternative did not always have the second-best probability; for example, it had the third and fourth-best probabilities at medium and hard levels, respectively. In addition, another difference was that we created unique probability and cost/profit pairs for each trial and that we selected three sets, *i.e.*, Set1:[25%, 90%], Set2:[30%, 90%], Set3:[33%, 90%] for the probability endpoints to increase the variety of trials [C2]. Like in experiment 1, we transformed the endpoints of each set into log odds and applied linear interpolation to sample three other probability values between the endpoints of each set. The dataset is available at the OSF link at the end of the abstract.

7.3 Procedure

The procedure was identical to experiment 1 (Sec. 6.4), except for decreasing the length of training [C3]. Participants answered only the question of finding the highest probability among 5 alternatives until they found the correct answer or up to 5 trials in total, and did not fill a literacy test.

7.4 Measures

The measures were identical to experiment 1 (Sec. 6.6), except removing the risk literacy measure to reduce duration given the lack of observed relation between risk literacy and accuracy (Sec. 6.10.2).

7.5 Participants

We again recruited participants via Prolific. Sample specifications, payment, planned sample size, and exclusion criteria were as in experiment 1 (Sec. 6.8). We got valid data from 301 participants ($N = 146$ for the judgment task; $N = 155$ for the decision task), 61% male, 56% aged between 18-30, from 26 countries (top ones: UK 32%, USA 15%, Poland 10%, and South Africa 10%).

7.6 Results of experiment 2

All analyses were planned and preregistered (see OSF link in abstract) prior to data collection. Fig. 3 (right) shows the results. Analysis methods, computations, and their explanation are identical to experiment 1.

7.6.1 Statistical hypothesis H_1

We compared the mean accuracy score of the judgment task and the decision task (see Fig. 3, column EXP2, row H1). We found suggestive evidence for a difference between judgment accuracy and decision accuracy, more specifically, that decision accuracy is higher than judgment accuracy.

7.6.2 Statistical hypothesis H_2

We compared the mean accuracy score of the judgment task and the decision task across scenarios. In the sports scenario (see Fig. 3, column EXP2, row H2, sports), we did not find evidence that judgment accuracy and decision accuracy differ. However, for the humanitarian scenario (see Fig. 3, column EXP2, row H2, humanitarian), we found suggestive evidence for a difference between judgment accuracy and decision accuracy, more specifically, that decision accuracy is higher than judgment accuracy in the humanitarian scenario.

7.6.3 Statistical hypothesis H_3

We compared the mean accuracy score of the judgment task and the decision task across visualizations, *i.e.* quantile dotplots, density plots, and probability bars. We found suggestive evidence for a difference between judgment accuracy and decision accuracy in probability bars, but the evidence is weaker for the remaining visualizations (see Fig. 3 column EXP2, row H3).

7.6.4 Statistical hypothesis H_4

We compared the mean accuracy of the judgment task and the decision task across scenarios and visualization pairs. Results were inconclusive for the sports scenario, except when the probability bars were used (see Fig. 3, column EXP2, row H4). For the humanitarian scenario, we found suggestive evidence for a difference across all three visualizations (see Fig. 3, column EXP2, row H4).

7.6.5 Statistical hypothesis H_5

We compared the mean status quo and risk-aversion percentage for the judgment task and the decision task. We did not find evidence for a difference between the judgment task and the decision task on risk-aversion percentage (see Fig. 3, column EXP2, row H5, risk averse percentage). In contrast, we found suggestive evidence for a difference between the judgment task and the decision task on status quo percentage. Specifically, there was more status-quo bias in the judgment task than in the decision task (see Fig. 3, column EXP2, row H5, status quo percentage).

7.7 Additional Analyses

We investigated if there is a correlation between judgment accuracy and decision accuracy and response time and found some evidence for a moderate positive correlation, $r = 0.29$, CI [0.12, 0.43]. We also found some evidence that increasing response time may lead to more accurate judgments, $r = 0.20$, CI [0.02, 0.36]. Overall, this suggests that it is more likely to observe more accurate judgments and decisions with increasing time on the task.

7.8 Participant Self-reported Strategies

In Experiment 2, we found suggestive evidence for a difference between judgment and decision accuracy. To help explain our findings, we analyzed participants' self-reported strategies through a qualitative analysis. We identified high-level strategies and keywords related to judgment and decision making, including variants like "judge", "decide", and "choose". A comprehensive breakdown of these strategies is available in the supplementary materials at the OSF link.

Many participants in both judgment (61 of 146) and decision (62 of 155) tasks either did not answer or reported having no strategy. The most common strategy reported by both judgment (50 of 146) and decision (58 of 155) participants involved seeking for a balance between the cost/profit and probability. The textual analysis did not reveal differences between judgment and decision responses: 17 judgment participants mentioned decision-related terms, while 19 decision participants did so. It appeared that these references could simply relate to the AFC response choice format rather than being indicative of a high-level strategy.

7.9 Summary of Experiment 2

Consistent with H_1 , we found suggestive evidence that, overall, decisions are more accurate than judgments. For H_2 to H_4 , we noted in Sec. 6.9 that, given the relatively large number of statistical hypotheses, each individual finding should be interpreted as tentative. However, the results in the second column of Fig. 3 are remarkably consistent with our results for H_1 , with CIs being either inconclusive or providing evidence that decision accuracy is more accurate. Thus there is converging evidence. Importantly, we obtained different results for our conditions (*e.g.*, the evidence is slightly stronger for the humanitarian scenario than for the sports scenario), but we cannot conclude that the effect differs across conditions [80], [81]. Regarding H_5 , Fig. 3 suggests that the higher accuracy for the decision task might be driven by a stronger status quo bias in the judgment task, indicating participants were more inclined to take action in the decision task. Finally, through exploratory qualitative analysis, we noted that participants did not appear aware of following a strategy that differentiates between judgment and decision.

8 DISCUSSION

Experiment 1 did not provide evidence of a difference between judgment accuracy and decision accuracy (Sec. 6.9). Experiment 2, with a slightly different design (including increasing the task difficulty, see Sec. 7), gave supportive evidence suggesting that judgment accuracy is not a good proxy for decision accuracy. We discuss these findings next.

First, **we found suggestive evidence that decision accuracy was higher than judgment accuracy.** This outcome

contrasts with Dimara *et al.*'s findings [46] (see Sec. 2.3). In their study, both narratives involved an optimal selection among 20 options. Participants in the decision making narrative, asked to choose a house for themselves, were less accurate than participants in the analytic narrative, asked to identify a top real estate deal for clients. Three factors might explain this discrepancy: (1) different numbers of alternatives (5 in our work *vs.* 20 in [46]); (2) differences between the nature of the scenarios (humanitarian and sports *vs.* real estate); and (3) different task complexities. Previous studies showed that the decision task gets harder with more alternatives, *e.g.*, 10 or 20 in [82]. Secondly, a scenario's emotional value can affect how people approach a decision task (see the humanitarian and sports scenarios). Finally, on task complexity, it appears that the judgment task (s) in Dimara *et al.* are not too hard, as they only require spatial comparisons of different points. In contrast, the judgment task in our work require mental aggregation of extracted probabilities from uncertainty visualizations with cost/profit values to make a rough trade-off analysis – which is less straightforward and prone to more errors.

Secondly, **in the humanitarian scenario, we found suggestive evidence that decision accuracy was higher than judgment accuracy.** If a scenario effect indeed exists, it might stem from the emotional value associated with the humanitarian scenario. Previously, Castro *et al.* [3] analyzed decision strategies and showed that emotions associated with alpacas in danger made some subjects act irrationally, *e.g.*, by constantly issuing blankets to alpacas. Contrary to this, our findings suggest that emotions might not always be detrimental (see also [83]). Feeling responsible about deciding to help orphan kids might have sharpened participants' focus on pertinent information, aiding subsequent mental calculations. For example, some participants reported that they exclusively calculated EVs for the humanitarian scenario: *"I just tried to balance the profits/help for the kids with a probability of winning/the goods arriving on time."*

Thirdly, **a stronger status-quo bias was potentially more prevalent in the judgment task than in the decision task.** While status-quo bias has been identified as a cognitive bias for decision tasks [31], our results indicate its presence also in judgment tasks. Research suggests that individuals tend to favor the status-quo especially when faced with challenging tasks [84]. The strong bias in our judgment task might stem from its inherent difficulty, given it demands mentally combining probabilities with cost/benefit evaluation. Conversely, the diminished status-quo bias in the decision task might be attributed to participants feeling a direct accountability for outcomes, driving them to take immediate actions.

Among the three visualizations, **we found suggestive evidence for a difference only with probability bars, indicating that decision accuracy was higher than judgment accuracy.** This finding is consistent with Kale *et al.* [6] study in that interval visualizations prompted better decisions.

Furthermore, **we clearly observed that experiment 2 was harder than experiment 1**, as can be traced from the distribution of raw accuracy scores (see Fig. 4), even though our goal was not to compare the two experiments. This confirms that the changes we made, *e.g.*, unique trials, and personalized task wording increased task difficulty. However, although the two experiments differed in their outcomes

(inconclusive *vs.* decision accuracy > judgment accuracy), there was no evidence of differences in effects between experiments. Therefore, we cannot conclude that the changes we made to experiment 1 amplified the difference between the decision task and the judgment task [80], [81].

Additionally, designing an experiment to equitably compare judgment and decision making tasks was intricate. To enable such comparison, we simplified these tasks based on key decision making features discussed in previous research (see Sec. 3.1). Although we observed distinct trends in accuracy differences, this simplified design, which might explain why participants' self-reported strategies were not as informative (see Sec. 7.8), cannot reveal the underlying reasons for those differences. We encourage future research to explore these decision features and use more elaborated tasks to uncover the heuristics that decision makers apply.

Nevertheless, our findings suggest that even with simple judgment and decision tasks—those requiring only basic calculations on a few attributes—decision accuracy appears higher than judgment accuracy, while the Dimara *et al.* [46] study suggested the reverse. Yet our findings align with both Kale *et al.* [6] and Dimara *et al.* [46] studies on the need to decouple judgment and decision making performance. We thus propose four essential action points for visualization research: (1) establishing a shared *definition* of what constitutes judgment and what decision making; (2) investigating *task context* familiarity and valence to counteract potential biases; (3) developing *scales* to measure and categorize difficulty levels for judgment tasks and decision tasks, facilitating comparisons across studies and enabling the calibration of task difficulty; and (4) developing *metrics* that assess a complex system's capacity to aid decision making. This diverges from current evaluations which use analytic task performance as a proxy for decision making ability or solely measure task-time. For the latter, one could adapt preference elicitation methodologies [32] to objectively measure users' decision accuracy by their subjective preferences.

9 CONCLUSION

The conjoined study of judgment and decision making has a long history, much older than the visualization field [85]–[87]. Our work contributed to the decoupling of judgment and decision making in the following ways. We identified inconsistencies in terminology and consequently misinterpretations on whether people make suboptimal decisions because of or despite inaccurate judgments of information. To enhance clarity, we analyzed, compared, and contrasted relevant concepts, experiment designs, and findings from the literature. To the best of our knowledge, we conducted the first experiments that investigated judgments and decisions explicitly as distinct, yet direct-equivalent tasks. Contrary to our expectations and trends observed in previous research, we found decisions to be more accurate in affective scenarios and less vulnerable to the status-quo bias, suggesting that decision makers may disfavor responses associated with inaction. We conclude that judgments cannot be safely used as proxy tasks for decision making.

Yet studying how, when, where, and why decision and judgment tasks differ remains an open question. Future research is needed to understand the currently unknown heuristics that decision makers use.

REFERENCES

- [1] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci, "Decision making with visualizations: a cognitive framework across disciplines," *CRPI*, vol. 3, pp. 29:1–29:25, 2018.
- [2] L. Cheong, S. Bleisch, A. Kealy, K. Tolhurst, T. Wilkening, and M. Duckham, "Evaluating the impact of visualization of wildfire hazard upon decision-making under uncertainty," *IJGIS*, vol. 30, no. 7, pp. 1377–1404, 2016.
- [3] S. C. Castro, P. S. Quinan, H. Hosseinpour, and L. Padilla, "Examining effort in 1d uncertainty communication using individual differences in working memory and nasa-tlx," *IEEE TVCG*, vol. 28, no. 1, pp. 411–421, 2022.
- [4] N. Jardine, B. D. Ondov, N. Elmqvist, and S. Franconeri, "The perceptual proxies of visual comparison," *IEEE TVCG*, vol. 26, no. 1, pp. 1012–1021, 2020.
- [5] A. Kale, F. Nguyen, M. Kay, and J. Hullman, "Hypothetical outcome plots help untrained observers judge trends in ambiguous data," *IEEE TVCG*, vol. 25, no. 1, pp. 892–902, 2019.
- [6] A. Kale, M. Kay, and J. Hullman, "Visual reasoning strategies for effect size judgments and decisions," *IEEE TVCG*, vol. 27, no. 2, pp. 272–282, 2021.
- [7] T. Munzner, *Visualization analysis and design*. CRC Press, 2014.
- [8] M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafrin, "A design space of vision science methods for visualization research," *IEEE TVCG*, vol. 27, no. 2, pp. 1117–1127, 2021.
- [9] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss, "Colorgical: Creating discriminable and preferable color palettes for information visualization," *IEEE TVCG*, vol. 23, no. 1, pp. 521–530, 2017.
- [10] J. Talbot, J. Gerth, and P. Hanrahan, "An empirical model of slope ratio comparisons," *IEEE TVCG*, vol. 18, no. 12, pp. 2613–2620, 2012.
- [11] S. Mittelstädt, A. Stoffel, and D. A. Keim, "Methods for compensating contrast effects in information visualization," *Comput. Graphics Forum*, vol. 33, no. 3, pp. 231–240, 2014.
- [12] R. Veras and C. Collins, "Saliency deficit and motion outlier detection in animated scatterplots," in *Proc. CHI*, pp. 541:1–541:12, ACM, 2019.
- [13] S. Haroz and D. Whitney, "How capacity limits of attention influence information visualization effectiveness," *IEEE TVCG*, vol. 18, no. 12, pp. 2402–2410, 2012.
- [14] C. Xiong, C. R. Ceja, C. J. Ludwig, and S. Franconeri, "Biased average position estimates in line and bar graphs: Underestimation, overestimation, and perceptual pull," *IEEE TVCG*, vol. 26, no. 1, pp. 301–310, 2020.
- [15] A. Patil, G. Richer, C. Jermaine, D. Moritz, and J.-D. Fekete, "Studying early decision making with progressive bar charts," *IEEE TVCG*, vol. 29, no. 1, pp. 407–417, 2023.
- [16] L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking visualizations of correlation using weber's law," *IEEE TVCG*, vol. 20, no. 12, pp. 1943–1952, 2014.
- [17] A. C. Telea, *Data visualization: Principles and practice*. CRC Press, 2014.
- [18] P. Kayongo, G. Sun, J. Hartline, and J. Hullman, "Visualization equilibrium," *IEEE TVCG*, vol. 28, no. 1, pp. 465–474, 2022.
- [19] Y. Zhang, R. K. Bellamy, and W. A. Kellogg, "Designing information for remediating cognitive biases in decision-making," in *Proc. CHI*, pp. 2211–2220, ACM, 2015.
- [20] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri, "Illusion of causality in visualized data," *IEEE TVCG*, vol. 26, no. 1, pp. 853–862, 2020.
- [21] H. Song and D. A. Szafrin, "Where's my data? evaluating visualizations with missing data," *IEEE TVCG*, vol. 25, no. 1, pp. 914–924, 2019.
- [22] J. Schwarz and M. Morris, "Augmenting web pages and search results to support credibility assessment," in *Proc. CHI*, pp. 1245–1254, ACM, 2011.
- [23] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan, "Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs," *IEEE TVCG*, vol. 24, no. 1, pp. 340–350, 2018.
- [24] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister, "Evaluation of artery visualizations for heart disease diagnosis," *IEEE TVCG*, vol. 17, no. 12, pp. 2479–2488, 2011.
- [25] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann, "Time-hierarchical clustering and visualization of weather forecast ensembles," *IEEE TVCG*, vol. 23, no. 1, pp. 831–840, 2017.
- [26] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE TVCG*, vol. 12, no. 5, pp. 741–748, 2006.
- [27] N. Willems, H. Van De Wetering, and J. J. Van Wijk, "Visualization of vessel movements," *Comput. Graphics Forum*, vol. 28, no. 3, pp. 959–966, 2009.
- [28] P. E. Rauber, A. X. Falcao, and A. C. Telea, "Projections as visual aids for classification system design," *Information Visualization*, vol. 17, no. 4, pp. 282–305, 2018.
- [29] J. Bertin, *Semiology of graphics: Diagrams, networks, maps*. Esri Press, 1983.
- [30] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: Foundations, techniques, and applications*. A.K. Peters, 2015.
- [31] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE TVCG*, vol. 26, no. 2, pp. 1413–1432, 2020.
- [32] E. Dimara, A. Bezerianos, and P. Dragicevic, "Conceptual and methodological issues in evaluating multidimensional visualizations for decision support," *IEEE TVCG*, vol. 24, no. 1, pp. 749–759, 2018.
- [33] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay, "Uncertainty displays using quantile dotplots or cdfs improve transit decision-making," in *Proc. CHI*, pp. 144:1–144:12, ACM, 2018.
- [34] E. Hindalong, J. Johnson, G. Carenini, and T. Munzner, "Abstractions for visualizing preferences in group decisions," *ACM HCI*, vol. 6, no. CSCW1, pp. 49:1–49:44, 2022.
- [35] A. Savikhin, R. Maciejewski, and D. S. Ebert, "Applied visual analytics for economic decision-making," in *Proc. VAST*, pp. 107–114, IEEE, 2008.
- [36] A. Narechania, A. Coscia, E. Wall, and A. Endert, "Lumos: Increasing awareness of analytic behavior during visual data analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 1009–1018, 2022.
- [37] L. Cibulski, J. Schmidt, and W. Aigner, "Reflections on visualization research projects in the manufacturing industry," *IEEE CG&A*, vol. 42, no. 2, pp. 21–32, 2022.
- [38] E. Dimara and J. Stasko, "A critical reflection on visualization research: Where do decision making tasks hide?," *IEEE TVCG*, vol. 28, no. 1, pp. 1128–1138, 2022.
- [39] E. Dimara, H. Zhang, M. Tory, and S. Franconeri, "The unmet data visualization needs of decision makers within organizations," *IEEE TVCG*, vol. 28, no. 12, pp. 4101–4112, 2022.
- [40] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *Proc. CHI*, pp. 559:1–559:12, ACM, 2019.
- [41] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim, "A survey of human-centered evaluations in human-centered machine learning," *Comput. Graphics Forum*, vol. 40, no. 3, pp. 543–568, 2021.
- [42] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The state of the art in enhancing trust in machine learning models with the use of visualizations," *Comput. Graphics Forum*, vol. 39, no. 3, pp. 713–756, 2020.
- [43] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of International Conference on Intelligence Analysis*, vol. 5, pp. 2–4, 2005.
- [44] K. Sedig, P. Parsons, and A. Babanski, "Towards a characterization of interactivity in visual analytics," *J. Multim. Process. Technol.*, vol. 3, no. 1, pp. 12–28, 2012.
- [45] IEEE Visualization, "Analytics and decisions area model," 2022. <http://ieevis.org/year/2022/info/call-participation/area-model>.
- [46] E. Dimara, A. Bezerianos, and P. Dragicevic, "Narratives in crowd-sourced evaluation of visualizations: A double-edged sword?," in *Proc. CHI*, pp. 5475–5484, ACM, 2017.
- [47] A. P. Association., "APA Dictionary of Psychology (2nd ed.)," 2015.
- [48] J. Berger, *Statistical decision theory: foundations, concepts, and methods*. Springer Science & Business Media, 2013.
- [49] K. Eberhard, "The effects of visualization on judgment and decision-making: a systematic literature review," *Management Review Quarterly*, vol. 73, no. 1, pp. 167–214, 2023.
- [50] A. Diederich and J. R. Busemeyer, *Judgment and decision making*, ch. 24, pp. 660–683. John Wiley & Sons, Inc., 2012.

- [51] B. Fischhoff and S. B. Broomell, "Judgment and decision making," *Annual Review of Psychology*, vol. 71, no. 1, pp. 331–355, 2020.
- [52] E. Oral, R. Chawla, M. Wijkstra, N. Mahyar, and E. Dimara, "From information to choice: A critical inquiry into visualization tools for decision making," *IEEE TVCG*, 2023. Preprint.
- [53] H. A. Simon, *The new science of management decision*. Harper & Brothers, 1960.
- [54] N. J. Wispinski, J. P. Gallivan, and C. S. Chapman, "Models, movements, and minds: bridging the gap between decision making and action," *Ann. N. Y. Acad. Sci.*, vol. 1464, no. 1, pp. 30–51, 2020.
- [55] N. Gold, B. D. Pulford, and A. M. Colman, "Do as i say, don't do as i do: Differences in moral judgments do not translate into differences in decisions in real-life trolley problems," *Journal of Economic Psychology*, vol. 47, pp. 50–61, 2015.
- [56] G. Karlsson, "A phenomenological psychological study of decision and choice," *Acta Psychologica*, vol. 68, no. 1, pp. 7–25, 1988.
- [57] C. Pletti, L. Lotto, G. Buodo, and M. Sarlo, "It's immoral, but i'd do it! psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations," *Br. J. Psychol.*, vol. 108, no. 2, pp. 351–368, 2017.
- [58] J. A. Hampton, "Concepts as prototypes," vol. 46 of *Psychology of Learning and Motivation*, pp. 79–113, Academic Press, 2006.
- [59] L. Decock and I. Douven, "What is graded membership?," *Noûs*, vol. 48, no. 4, pp. 653–682, 2014.
- [60] K. Savani, J. Cho, S. Baik, and M. W. Morris, *Culture and judgment and decision making*, vol. 2, ch. 16, pp. 456–477. John Wiley & Sons, Ltd, 2015.
- [61] E. R. Van Norman, P. M. Nelson, J.-E. Shin, and T. J. Christ, "An evaluation of the effects of graphic aids in improving decision accuracy in a continuous treatment design," *J. Behav. Educ.*, vol. 22, pp. 283–301, 2013.
- [62] G. A. Gescheider, *Psychophysics: the fundamentals*. Psychology Press, 2013.
- [63] D. B. Liston and L. S. Stone, "Saccadic brightness decisions do not use a difference model," *JOV*, vol. 13, no. 8, pp. 1:1–1:10, 2013.
- [64] S. K. Lynn, J. B. Wormwood, L. F. Barrett, and K. S. Quigley, "Decision making from economic and signal detection perspectives: development of an integrated framework," *Frontiers in psychology*, vol. 6, pp. 952:1–952:14, 2015.
- [65] V. H. Visschers, R. M. Meertens, W. W. Passchier, and N. N. De Vries, "Probability information in risk communication: a review of the research literature," *Risk Anal.*, vol. 29, no. 2, pp. 267–287, 2009.
- [66] H. F. Chua, J. F. Yates, and P. Shah, "Risk avoidance: Graphs versus numbers," *M&C*, vol. 34, pp. 399–410, 2006.
- [67] E. R. Stone, J. F. Yates, and A. M. Parker, "Effects of numerical and graphical displays on professed risk-taking behavior," *JEP:A*, vol. 3, no. 4, pp. 243–256, 1997.
- [68] E. R. Stone, W. R. Sieck, B. E. Bull, J. F. Yates, S. C. Parks, and C. J. Rush, "Foreground: background salience: Explaining the effects of graphical displays on risk avoidance," *Organizational Behavior and Human Decision Processes*, vol. 90, no. 1, pp. 19–36, 2003.
- [69] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer, "Using icon arrays to communicate medical risks: Overcoming low numeracy," *Health Psychol.*, vol. 28, no. 2, pp. 210–216, 2009.
- [70] C. M. Wu, B. Meder, F. Filimon, and J. D. Nelson, "Asking better questions: How presentation formats influence information search," *JEP:LMC*, vol. 43, no. 8, pp. 1274–1297, 2017.
- [71] L. Morais, Y. Jansen, N. Andrade, and P. Dragicevic, "Can anthropographics promote prosociality? a review and large-sample study," in *Proc. CHI*, pp. 611:1–611:18, ACM, 2021.
- [72] H. Zhang and L. T. Maloney, "Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition," *Frontiers in Neuroscience*, vol. 6, pp. 1:1–1:14, 2012.
- [73] E. T. Cokely, M. Galesic, E. Schulz, S. Ghazal, and R. Garcia-Retamero, "Measuring risk literacy: The berlin numeracy test," *Judgment and Decision Making*, vol. 7, pp. 25–47, 2012.
- [74] G. Cumming, "The new statistics: Why and how," *Psychol. Sci.*, vol. 25, no. 1, pp. 7–29, 2014.
- [75] P. Dragicevic, *Fair statistical communication in HCI*, pp. 291–330. 2016.
- [76] L. Besançon and P. Dragicevic, "The continued prevalence of dichotomous inferences at CHI," in *Proc. CHI EA*, pp. alt14:1–alt14:11, ACM, 2019.
- [77] G. Cumming and S. Finch, "Inference by eye: Confidence intervals and how to read pictures of data," *Am. Psychol.*, vol. 60, no. 2, pp. 170–180, 2005.
- [78] M. Krzywinski and N. Altman, "Error bars: The meaning of error bars is often misinterpreted, as is the statistical significance of their overlap," *Nature Methods*, vol. 10, pp. 921–923, 2013.
- [79] K. N. Kirby and D. Gerlanc, "BootES: An R package for bootstrap confidence intervals on effect sizes," *Behav. Res.*, vol. 45, pp. 905–927, 2013.
- [80] A. Gelman and H. Stern, "The difference between "significant" and "not significant" is not itself statistically significant," *The American Statistician*, vol. 60, no. 4, pp. 328–331, 2006.
- [81] S. Nieuwenhuis, B. U. Forstmann, and E.-J. Wagenmakers, "Erroneous analyses of interactions in neuroscience: A problem of significance," *Nat. Neurosci.*, vol. 14, pp. 1105–1107, 2011.
- [82] B. Dy, N. Ibrahim, A. Poorthuis, and S. Joyce, "Improving visualization design for effective multi-objective decision making," *IEEE TVCG*, vol. 28, no. 10, pp. 3405–3416, 2022.
- [83] G. Lufityanto, C. Donkin, and J. Pearson, "Measuring intuition: Nonconscious emotional information boosts decision accuracy and confidence," *Psychol. Sci.*, vol. 27, no. 5, pp. 622–634, 2016.
- [84] S. M. Fleming, C. L. Thomas, and R. J. Dolan, "Overcoming status quo bias in the human brain," *Proc. Natl. Acad. Sci.*, vol. 107, no. 13, pp. 6005–6009, 2010.
- [85] D. Kahneman, "Article commentary: Judgment and decision making: A personal view," *APS*, vol. 2, no. 3, pp. 142–145, 1991.
- [86] D. Kahneman, "A perspective on judgment and choice: mapping bounded rationality," *Am. Psychol.*, vol. 58, no. 9, pp. 697–720, 2003.
- [87] G. Gigerenzer and W. Gaissmaier, "Heuristic decision making," *Annual review of psychology*, vol. 62, pp. 451–482, 2011.



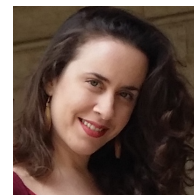
Başak Oral earned an MA in Cognitive Psychology from Bogazici University, Turkey, in 2021 and, since November 2021, has been enrolled at Utrecht University as a PhD candidate with a primary research focus on multi-criteria decision making (MCDM) with interactive data visualizations.



Pierre Dragicevic is a permanent Research Scientist at Inria Bordeaux, France. He is interested in humanitarian data visualization, physical and immersive visualizations, judgment and decision-making with visualizations, research transparency and statistical communication, as well as design spaces and conceptual frameworks.



Alexandru Telea received his PhD (2000) in Computer Science from the Eindhoven University of Technology. He was assistant professor in visualization and computer graphics at the same university (until 2007) and then full professor of visualization at the University of Groningen. Since 2019 he is full professor of visual data analytics at Utrecht University. His interests include high-dimensional visualization, visual analytics, and image-based information visualization.



Evanthia Dimara is Tenured Assistant Professor at Utrecht University. Her fields of research are Information Visualization and Human-Computer Interaction. Her focus is on decision making – how to help people make unbiased and informed decisions alone or in groups.