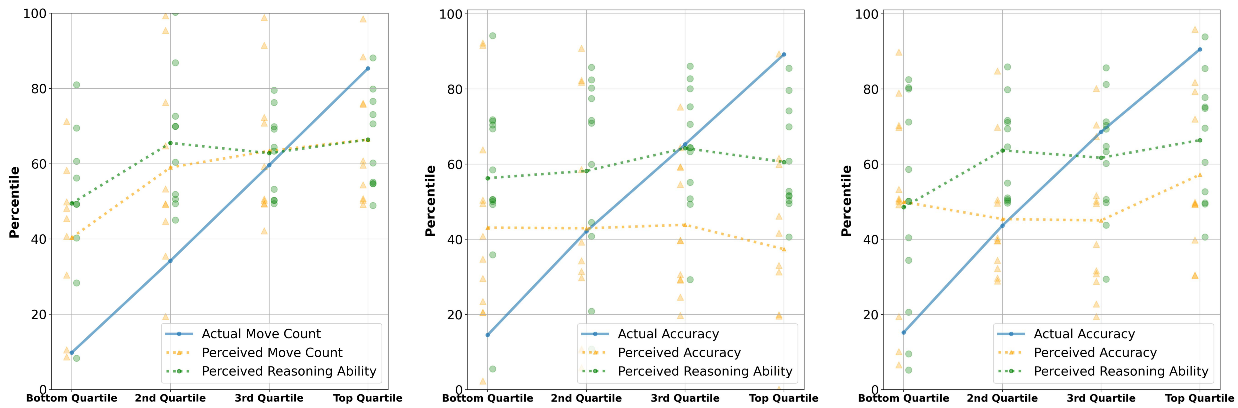
Unmasking Dunning-Kruger Effect in Visual Reasoning & Judgment

Mengyu Chen, Yijun Liu, and Emily Wall



(a) Sliding Puzzle Game (b) Categorization Task: Car (c) Categorization Task: Credit

Fig. 1: DKE results from two studies across three tasks show actual vs. perceived performance percentiles on the X and Y axes, respectively. The blue line shows a baseline if participants perceived their performance accurately. The yellow dotted line depicts participants' perceived performance percentile in a sliding puzzle game (a) and a categorization task using cars (b) and credit (c). The green color depicts participants' perceived reasoning ability. Across all three tasks, we observe the canonical curves indicative of DKE: bottom quartile performers tend to overestimate their performance, while top quartile performers tend to underestimate performance.

Abstract—The Dunning-Kruger Effect (DKE) is a metacognitive phenomenon where low-skilled individuals tend to overestimate their competence while high-skilled individuals tend to underestimate their competence. This effect has been observed in a number of domains including humor, grammar, and logic. In this paper, we explore if and how DKE manifests in visual reasoning and judgment tasks. Across two online user studies involving (1) a sliding puzzle game and (2) a scatterplot-based categorization task, we demonstrate that individuals **are** susceptible to DKE in visual reasoning and judgment tasks: those who performed best underestimated their performance, while bottom performers overestimated their performance. In addition, we contribute novel analyses that correlate susceptibility of DKE with personality traits and user interactions. Our findings pave the way for novel modes of bias detection via interaction patterns and establish promising directions towards interventions tailored to an individual's personality traits. All materials and analyses are in supplemental materials: https://github.com/CAV-Lab/DKE_supplemental.git.

Index Terms—Cognitive Bias, Dunning Kruger Effect, Metacognition, Personality Traits, Interactions, Visual Reasoning

1 Introduction

Imagine that two colleagues, Bob and Jane, are tasked with analyzing visualizations of model output for financial forecasting at their company. Bob, despite his limited experience in data analysis, confidently takes charge of interpreting the financial charts. He gravitates towards a line chart with smoothed trends, which he interprets as conclusive evidence of growth. His confident interpretation presents an incomplete, overly simplified view of the data, masking critical fluctuations and anomalies indicated by other visualizations of the model output. This could potentially lead to misguided strategic decisions based on an incomplete understanding of market dynamics. On the other hand, Jane, with a strong background in financial analysis, focuses on a scatterplot that details individual data points representing key financial metrics with confidence intervals around each prediction produced by the forecasting model. Her interpretation highlights variability and outliers, providing a more comprehensive view crucial for informed decision-making, suggesting a more cautious approach to future investments. Despite the precision of her analysis, Jane lacks confidence, believing that her nuanced interpretation might not be the best to base decisions on.

Both Bob and Jane exemplify a cognitive bias known as Dunning-Kruger Effect (DKE). In the seminal paper titled "Unskilled and Unaware of It", Kruger and Dunning describe a phenomenon in which the people who perform the worst on various knowledge tests have an inflated perception of their abilities [42]. Bottom-quartile performers believed that their performance was above average, while those in the top quartile underestimated their performance relative to their peers [42]. This lack of realization about one's own skill reflects a metacognitive deficit, i.e., a lack of "knowing what we know" and "knowing what we don't know" [2].

DKE can have numerous consequences. In the opening example, Bob's tendency to express uninformed views likely prevents other colleagues with more fruitful perspectives from participating. This phenomenon may also affect organizations, in which the most capable people may not be the ones making decisions; instead, those with the greatest self-perceived ability (and often lesser actual skill) take precedence. Hence, the social consequences of this bias can lead to larger systemic problems [59]. Namely, DKE can lead to situations wherein true expertise may not reach the decision-making table, dominated instead by those who may be unaware of their own lack of proficiency.

We posit that DKE also can critically affect visual reasoning and judgment tasks. People with limited knowledge in the domain of inter-

• *Mengyu Chen, Yijun Liu, and Emily Wall are with Emory University. E-mails: {mengyu.chen, yijun.liu, emily.wall}@emory.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

est or in visual data analysis practices may be prone to overestimating their ability to accurately reason with data supported by visualizations. The consequences of DKE in visual reasoning and judgment tasks could result in people confidently reporting on flawed analyses or drawing incorrect conclusions. Likewise, highly skilled users might underestimate their abilities, which can lead to a lack of confidence in their interpretations or decisions. This could cause second-guessing of sound analyses or potentially missing important insights or trends in the data.

In this paper, **we demonstrate that DKE can be replicated in visual reasoning and judgment across two experiments.** In the first study, DKE can be observed in participants' visual reasoning abilities using a sliding puzzle game. In the second study, DKE can be observed in participants' visual judgment through use of an interactive scatterplot to categorize data. Furthermore, our work builds on a prior body of work in interaction analysis that suggest relationships between (i) interactive behaviors and personality traits [3] and (ii) interactive behaviors and cognitive biases [65]. Specifically, we demonstrate that those who are most susceptible to DKE (i.e., have the biggest gap between their actual and perceived performance) tend to have higher values for the conscientiousness personality trait in the puzzle game (Study 1) and the car task (Study 2) and perceive they have greater domain familiarity in the car task and credit task (Study 2). In addition, those who performed at the extreme high and low ends in both studies exhibited some distinct patterns of interaction.

To our knowledge, this work is the first of its kind to address *metacognitive deficits* in visual reasoning and judgment. By demonstrating the presence of DKE in visual reasoning and judgment tasks, we provide further evidence of its pervasive impact. In addition, understanding *how* the bias manifests through interactive strategies and how it relates to personality traits paves the way towards personalized bias detection and targeted mitigation strategies.

2 Related Work

Our work is contextualized amongst several areas of prior work, including investigations of DKE in other settings, making inferences from user interactions, and efforts to understand biases in visualizations.

2.1 Dunning-Kruger Effect

In the Cognitive Science community, the term **bias** refers to errors that occur when people make decisions using "rules of thumb" or heuristics [35, 36, 61]. Despite being generally efficient [24, 25], these biases may lead to ineffective or wrong decisions. For DKE, in particular, even among highly educated communities (e.g., physicians [11], pilots [50], reviewers and editors [33]), people exhibit a compromised ability to accurately assess their own skills.

In Kruger and Dunning's seminal work [42], they attributed DKE to a lack in metacognitive abilities; that is, insufficient knowledge about one's own knowledge [22]. The effect has been uncovered in many settings involving medical resident training [54], debate team performance [20], beginning aviators [56], gun owners' knowledge of firearms [20], and tournament players in "Texas Hold'em" poker and chess [18], among others. Additionally, recent work examined DKE in the context of nuclear weapons, English grammar and logical reasoning, and considered personality and cognitive characteristics, in which the neuroticism trait has been linked to underconfidence, leading to increased underprecision [57]. Across all of these contexts, the canonical observation for DKE persists: that those who perform the worst tend to overestimate their performance, while those who perform the best tend to underestimate their performance.

We build upon these findings to determine the presence of DKE in two visual tasks and investigate the correlation between these tasks, personality traits, and interaction strategies.

2.2 Making Inferences from User Interactions

In this work, we aim to probe reasoning processes pertaining to DKE among two user groups exhibiting extreme task performance. Prior work on analytic provenance emphasizes the importance of understanding a user's reasoning process through their interactions with visual interfaces to perform analytical tasks [49, 53]. Previous research in the visual analytics and human computer interaction communities has laid a foundation for logging, storing, and interpreting a user's interactions and activities. Cowley et al., for instance, documented low-level events in the Glass Box system such as copy/paste, mouse clicks, and window activations [9], while others such as Willett et al. used historical interaction data to refine interfaces [71]. Others including Gomez and Laidlaw [28], Battle and Heer [1], Dou et al. [16], and Brown et al. [3] focus on predicting and recovering higher level reasoning processes. Our work leverages these prior insights on analytic provenance, with the goal of learning how interactions may correlate with DKE.

Additionally, existing research proposes that individual personality traits can serve as predictors of proficiency, specifically speed and accuracy when performing tasks [30, 76]. Interactive strategies such as basic navigation of zoom-in, zoom-out, and pan interactions was shown to correlate with locus of control, neuroticism, and extraversion in a "Where's Waldo"task [3].

We are inspired by these efforts to assess the extent to which we may observe relationships between personality traits, interactive strategies, and DKE, aiming to ultimately enable early personalized interventions by understanding individual tendencies and patterns.

2.3 Human Bias in Visualization

A growing body of work in visualization related to other forms of bias informs our efforts to explore DKE. For instance, Wall et al. defined metrics to quantify signals of bias from interactive behavior [64, 66]. Other metrics have been introduced to similarly capture concepts such as analytic focus [75] and exploration pacing and uniqueness [21]. Some such metrics have been associated with, e.g., selection bias [29] or anchoring bias [65]. Other researchers have replicated a variety of other cognitive biases in visual analytics. For instance, Xiong et al. demonstrated that existing knowledge or beliefs affect individuals' interpretations of charts and communication with visualizations (the curse of knowledge) [72], while Cho et al. demonstrated the anchoring effect in a visual analytic tool by priming [7].

Overarching the study of individual biases are efforts to create characteristic frameworks or taxonomies [14, 67] and identify strategies to mitigate biases [13, 48, 68, 69]. In addition to the explicit work on cognitive biases in visualization discussed above, there are other efforts that inform this work on metacognition in the visualization community. For instance, numerous studies utilize measures of self-reported confidence, e.g., [34, 39, 46, 52], which is a critical feature for measuring DKE (particularly in evaluating the gap between perceived and actual performance). Further, Kim et al. illustrated that people's interpretations of visualizations resemble Bayesian updating [40], a process integral to metacognition as it involves the reflection and adjustment of one's knowledge base.

3 General Method

We conducted two complementary user studies that cover visual reasoning and judgment tasks. In the first study (Section 4), participants arranged tiles in a 15-puzzle game, and in the second study (Section 5), participants completed a data categorization task using an interactive scatterplot. These tasks engage spatial reasoning and pattern recognition skills [37, 70] as well as interactivity [74], which are critical for making sense of data in visualizations [74]. In this section we outline the general method and hypotheses that are common to both studies.

3.1 Procedure

After providing informed consent, participants began with the 20-item Big Five Personality inventory [15], which produces a score from 4 to 20 for five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Participants then completed the main task (15 puzzle game in Study 1; data categorization in Study 2), during which we logged their interactions with the interface. Two attention check questions were interspersed. Participants were compensated at $10/hour based on the estimated completion time for each study.

Afterward, participants were asked to estimate their performance relative to their peers as a percentile along two dimensions: solution optimality (Study 1) or accuracy (Study 2), and reasoning ability (both). This estimation was done in a post-survey with questions such as *'how efficiently (fewer moves is better) do you think you did relative to your peers?'* A higher percentile means that the participant perceived that they performed better than their peers. Participants were also asked to rate their familiarity with the puzzle game (Study 1) or data domain (Study 2), e.g., *'To what degree do you consider yourself familiar with the puzzle game?'* on a 5-point Likert scale (1: not at all familiar, 5: extremely familiar). We asked participants about their domain familiarity after the task to prevent any influence on their confidence or perceived expertise before the task. This approach aimed to gather unbiased performance data and assess the correlation between perceived familiarity and actual/perceived performance without pre-task bias.

3.2 Hypotheses

Across the two user studies, we hypothesized that:

- **H1** **DKE in Visualization.** We will replicate DKE within the context of visualization. Specifically, less competent individuals will overestimate their task performance relative to peers, while competent individuals will underestimate their corresponding performance percentile.
- **H2** **Performance and Interactions.** To our knowledge, no previous work has specifically explored whether there are indicative behaviors associated with the DKE. Hence, we hypothesize that there will be detectable differences in interactive strategies used by individuals who are more and less competent.
- **H3** **Interactions and Personality.** People with different personalities will display different interactive strategies.
- **H4** **Personality and Δ Performance.** The difference between actual and perceived performance (Δ Performance) quantifies the magnitude of metacognitive miscalibration in DKE. We hypothesize that there will be correlations between personality traits and Δ Performance.
- **H5** **Δ Performance and Domain.** Previous studies have demonstrated that personal familiarity with a particular event or outcome tends to boost comparative optimism [47]. Hence, we hypothesize that people's overestimation of their performance will be positively associated with their familiarity in the respective domains.

4 Study 1: Visual Reasoning with 15-Puzzle Game

We selected a puzzle game as our first task because of its relevance to spatial reasoning [10] and pattern recognition [6] that are key to reasoning with visualizations [62, 63]. Importantly, it also serves as a relatively simple task, characterized by few interactive elements, which can be a valuable starting point to first verify the relevance of DKE in visualization prior to exploring more complex interactions and tasks. The goal of this study is to (1) examine if DKE exists in the context of visual reasoning, (2) investigate users' interactive behaviors when performing the task, and (3) examine if their personality traits are indicative of this bias. To realize the three goals, we designed a pre-registered experiment.

4.1 Experimental Setup

Task & Interface. The primary view of the 15-puzzle game, as depicted in Figure 2, features a 4x4 grid with 15 numbered tiles and one empty space, allowing tiles to be moved by dragging them into the empty slot (A). The goal of the puzzle is to rearrange tiles in ascending numerical order (1, 2, ..., 15) in the *least number of moves possible*. Below the board, a move counter increments by one with each move (B), allowing participants to track their total number of moves. Participants were informed at the beginning of the study that the back button was disabled to prevent reversing the move count. To guarantee comparability, all participants started the game with the same initial configuration (as shown in Figure 2). Users' interactions (including tiles clicked, positions they moved from/to, and time stamps of each movement) while performing the task were recorded for analysis.

Task Difficulty. Every possible state of a 4x4 board is solvable in 0 to 80 moves [4]. We conducted preliminary pilot studies to calibrate the appropriate task difficulty. These pilot studies involved a variety of puzzle sizes, ranging from the simpler 8-puzzle (3 by 3 layout) to a more complex 24-puzzle (5 by 5 layout). These trials were instrumental in gauging performance across a spectrum of difficulties, leading us to settle on the standard 15-puzzle (4 by 4 layout) configuration. We chose an initial configuration that could be solved in 10 moves, as determined by the A* algorithm [31], indicating a moderate difficulty level.

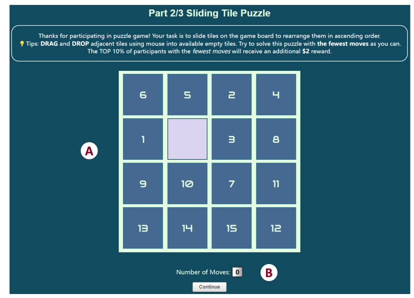


Fig. 2: 15-puzzle interface with (A) the primary puzzle and (B) move counter. The tiles shown represent the initial board configuration used in Study 1.

Recruitment. We conducted a power analysis of pilot data with 18 participants who completed the 15-puzzle using a similar experimental setup. Our minimum target sample size was 36 participants to obtain .8 power to detect a medium effect size of .25 at the standard $\alpha = .05$. We recruited 48 participants in total. Participants were compensated $2.50 for the study which had an estimated duration of 15 minutes (actual median completion time of 7.5 minutes). Prolific allowed participants to spend a maximum of 56 minutes on the task; no participants timed out. Participants were incentivized with an additional $1 performance bonus if they completed the puzzle in the top 10% based on lowest move count. The performance bonus was awarded to 23 people who completed the puzzle in the optimal number of 10 moves. No data were excluded from analyses due to failed attention checks in this study; however, 9 were discarded due to (i) data loss resulting in missing logs (5 people), and (ii) erroneous logs resulting from refreshing the browser which reset the task (4 people). This may lead to a skew in the data (e.g., poor performers with high move counts may be the ones who were more likely to refresh). In total *we collected and used data from 39 individuals* in our forthcoming analysis.

Participant Demographics. Among the 39 participants who completed the study, 18 identified as female, 20 identified as male, and 1 preferred not to disclose. The majority of participants (25) hold a college degree (Bachelor's, Master's, or Doctorate), while 10 have some college or Associate's degree, and 4 have a High School diploma. Participants were on average 34.87 years ($SD = 11.71$) of age. Participants rated familiarity with the solution of the puzzle game on average 2.23 out of 5 (with 33 participants reporting a familiarity level of 3 or lower).

4.2 Results

Analysis of DKE is based on a comparison of actual and perceived performance percentiles, particularly for the top quartile performers and bottom quartile performers. 23 participants completed the task in the optimal number of 10 moves. Thus for the forthcoming analysis on DKE, we defined quartiles of actual task performance primarily based on minimal move count, further differentiated by task completion time as an additional performance indicator. Participants spent on average 3.18 minutes (min: 0.22, max: 27.23) to solve the puzzle.

4.2.1 H1: DKE in Visual Reasoning

To test **H1**, we assigned a percentile ranking for each participant based on their actual move count, then by time spent to complete the puzzle .

As Fig 1(a) illustrates, bottom quartile participants whose actual move counts (blue) ranked in the 10^{th} percentile on average, placed themselves around the 40^{th} percentile (yellow). In the top quartile, however, participants whose actual performance fell in the 85^{th} percentile grossly underestimated their move count compared to their peers to be in the 65^{th} percentile on average. We found a statistically significant discrepancy between the actual and perceived percentiles for both the bottom quartile ($t = -4.11, p < 0.01$) and the top quartile ($t = 4.62, p < 0.01$). The misjudgment of performance among the two extreme quartiles was still observed even though perceived performance was significantly correlated with actual performance ($r(37) = 0.36, p = 0.025$). This indicates that although there is a general alignment between individuals' perceptions of their performance and their actual performance, those at the very high and very low ends of the spectrum tend to have distorted views of their own performance. Thus, we find support for **H1**, consistent with DKE.

Discussion of Results. Our results indicated that DKE is observable in the context of visual reasoning in the sliding puzzle game. We also considered possible confounds, e.g., that participants who performed the fewest movements might not spend the least time completing the puzzle, as they may be more likely to spend more time strategizing before moving. However, further exploratory analysis suggests otherwise. On average, the top performers took 0.47 minutes to 'think' before making their first move (the time between loading the puzzle game page and initiating the first move), compared to 1.57 minutes for the bottom group. In addition, the top performers averaged only 0.29 minutes ($SD = 0.14$) elapsed between the first and last move, whereas the bottom group took much longer with 6.65 minutes ($SD = 6.41$).

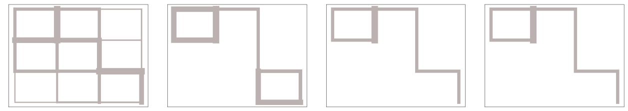
We also considered **time spent** as a measure of success (*'how quickly (less time is better) do you think you did relative to your peers'*), rather than move count, and nonetheless observed a similar pattern of DKE (see details in the supplemental materials). While the extent of the bias varied slightly between the two measures of success (move count vs. time spent), the overarching trend was consistent: participants at both ends of the skill spectrum showed discrepancies between their perceived and actual performances. Likewise, we also considered how performance compares to individuals' general perceived **reasoning ability** (*'how well do you think your reasoning ability compares to your peers'*) (Fig 1a, green). We again found a comparable result among top ($t = 4.54, p < 0.01$) and bottom ($t = -5.15, p < 0.01$) quartile. This reinforces the prevalence of DKE across multiple measures of task success and general reasoning ability. Details are provided in the supplemental materials.

4.2.2 H2: Performance and Interactions

To test **H2**, we visually examined the interactions of participants in four quartiles using lines overlaid on the puzzle grid (Fig. 3). Line thickness is proportional to the number of times a tile was moved in the given direction. For example, a horizontal mark in the top right portion of the figure signifies that participants moved tiles left-right or right-left into the top rightmost grid cell of the puzzle. The thicker the line, the more frequently that path was taken. The line widths are normalized based on each participant's actual move counts to ensure the view of the interactions is not dominated by participants who performed significantly more moves than others in their quartile.

Because 23 participants achieved the optimal solution, Figure 3 reveals the same movement paths for the third ($n_{Q_3} = 10$) and top ($n_{Q_4} = 9$) quartile groups, reflecting a single unique optimal solution. Low-skilled participants (Q_1) tended to randomly explore the board to find a solution, compared to their higher-skilled peers. We thus find support for **H2** via visual inspection, that there are detectable differences in interactive strategies used by individuals who are more and less competent.

Discussion of Results. The interaction analysis highlights differences in the interactions of participants across varying skill levels during the puzzle task. However, the ability to discern differences in strategies can be complicated by the large number of participants who achieved an optimal solution. Had we chosen a more complex puzzle configuration



(a) Q_1 (b) Q_2 (c) Q_3 (d) Q_4

Fig. 3: Interaction strategies by four participant groups (Q_1 = lowest performers and Q_4 = highest performers).

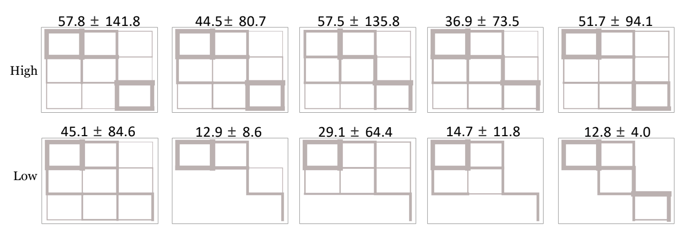
or one that had multiple optimal solutions, we may have been able to observe greater diversity in strategies among the top performers. We explore this further in the more difficult scatterplot categorization task in Study 2 (Section 5).

4.2.3 H3: Personality and Interactions

To test **H3**, we visualized participants' movement patterns similarly to the analysis for **H2**, but stratified by low and high scores for personality traits rather than low and high task performance. We analyze each individual personality trait independently, consistent with prior work on personality traits and DKE [57] and personality traits and interactions [3]. We confirmed that the distribution of scores for all five personality traits were normally distributed using a Shapiro-Wilk test [58] with all p values > 0.05. According to [27], participants' personality trait scores are considered 'average' if they fall within one-half standard deviation of the mean. Accordingly, we categorize the middle 40% of scores as average, with each tail (30%) representing high and low values for each personality trait. Figure 4 illustrates movement paths with average move counts by varied personality groups.

By visual inspection, we observe that participants with higher scores for each of four personality traits (conscientiousness, extraversion, agreeableness, and neuroticism) tended to explore the entire puzzle grid more evenly to find a solution and generally with higher move counts, whereas those with a lower score often left blank areas in certain segments of the grid. However, a significant difference was observed in move count between individuals with high and low scores for conscientiousness only ($u = 105, p = 0.01$).

These preliminary findings could indicate that individuals with lower scores might lean towards more precise interaction strategies and potentially achieve optimal task performance, although these trends are not statistically significant for four traits. Overall, we find mixed support for **H3**, with observable differences in interaction strategies for agreeableness, conscientiousness, extraversion, and neuroticism, and less clear differences for openness.



(a) O (b) C (c) E (d) A (e) N

Fig. 4: Movement path triggered by different personality traits with move counts Mean ± SD.

Discussion of Results. Our results suggest that participants scoring high in conscientiousness, extraversion, agreeableness, and neuroticism appeared to evenly explore the puzzle grid, possibly indicating a more exhaustive or trial-and-error approach to problem-solving. In contrast, participants with lower scores in these traits exhibited more selective interaction with the grid. One potential explanation for the blank areas in their movement paths could indicate a more contemplative approach where participants might have spent time strategizing before making moves which could reflect a more cautious or measured approach to problem-solving. The mixed support for **H3** underscores the complexity of the relationship between personality and problem-solving

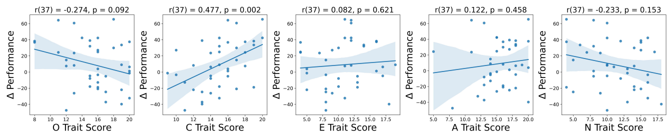


Fig. 5: Correlation between personality traits and ΔPerformance = Estimated Percentile − Actual Percentile).

behaviors and emphasizes the need for further research to unpack these dynamics.

4.2.4 H4: Personality and Δ Performance

To test **H4**, we computed Pearson correlation [51] for each of five personality traits compared to their respective disparity between actual and perceived performance. We use difference in performance as a proxy for susceptibility to DKE, e.g., people with a larger magnitude of Δ Performance are more prone to metacognitive miscalibration. Figure 5 depicts personality trait scores (x-axis, ranging from 4 to 20), and Δ Performance (y-axis, Δ = Estimated Percentile − Actual Percentile). When there is a positive/negative slope, it suggests a trend in over-/under-estimation of performance relative to the personality trait, while a flatter slope reflects more accurate perception of performance independent of personality traits.

Only Conscientiousness (C) was observed to have a significant effect on perceived performance ($r(37) = 0.48, p < 0.01$), implying that individuals with higher conscientiousness scores tend to exhibit a greater magnitude of overestimation of task performance. Thus, we find weak support for **H4**.

Discussion of Results. We observe that individuals high in conscientiousness, known for their meticulousness and strong commitment to task completion, may exhibit an optimistic attitude in assessing their capabilities and achievements [55, 60]. This optimism could stem from lofty personal standards and goals, leading to a self-view that matches their ideal performance.

While we found a statistically significant correlation for conscientiousness, we note that this result may be sensitive to how we define top and bottom performers. For example, altering our selection method to a random choice of 9 from the 23 participants who attained the optimal move count as the top quartile—instead of further differentiating by task completion time—eliminated the statistical significance for all personality traits (see details in Supplemental Materials). This suggests that the significant association with conscientiousness may not be robust and could be dependent on the performance metrics we adopted.

4.2.5 H5: Δ Performance and Domain

To test **H5**, we investigated the correlation between self-reported domain familiarity (from 1-5) and the manifestation of DKE. The Pearson correlation revealed no significant correlation between the two variables ($r(37) = -0.035, p = 0.83$). Thus, we find no support for **H5**.

Discussion of Results. Our results suggest that the tendency for people to overestimate their performance is not correlated with their familiarity with the sliding puzzle game. These results contribute to the ongoing discourse about the complex nature of self-assessment in cognitive tasks and highlight the interplay between self-perception and actual skill levels in various domains [8, 22].

5 Study 2: Categorization with Interactive Scatterplot

In Study 1, we replicated DKE in the context of a puzzle game, in which bias was measured as a function of perceived optimality of achieving the solution. We explore a complementary task in Study 2, where participants categorize data points in an interactive scatterplot. Bias in this study is measured as a function of the perceived accuracy of categorization. We build upon Study 1 by seeking to replicate DKE in a task that increases the complexity of interactions supported and gets closer to a realistic judgment task facilitated by visualizations . We present results of a pre-registered within-subjects study exploring DKE in the context of two categorization tasks in the domains of cars and credit where participants engage in interactive labeling [38].

5.1 Experimental Setup

Dataset & Tasks. Participants completed two tasks (order counterbalanced) in different domains. We used datasets from the domains of car type and credit score level as the general public usually has a reasonable degree of familiarity with these topics. For the car task, participants were asked to assign one of three types (SUV, Sedan, Minivan) to each point by comparing the provided statistics for each car, such as engine size and fuel economy (see details in the Supplemental Materials). Similarly, for the credit task, participants were asked to assign one of the three levels (Good, Standard, Poor) to each point based on credit-related traits for each person, such as number of credit cards and credit history age (see details in the Supplemental Materials).

These domains allow us to understand (1) the generalizability of DKE in this task beyond a single domain, and (2) varying levels of task complexity, facilitating the investigation of how these differences influence task performance [5].

To guarantee that the classes are distinctly separable based on the attributes of the data, we selected 30 points from each dataset and selected a subset of attributes to describe each data point: 6 attributes for the car task and 8 attributes for the credit task. This served as a proxy for task difficulty, where a pilot study with n = 12 participants confirmed (t = 4.69, $p < 0.01$) that the credit task was more difficult ($\mu_{accuracy} = 0.33$) than the car task ($\mu_{accuracy} = 0.47$).

Interface. We used an interactive scatterplot system in which the primary view displays the 30 points that represent individual cars or bank customers (Fig 6 (A)). Hovering on a point shows details about the particular car or customer in a tooltip (B). To label a point in the scatterplot, participants can click the appropriate category button (C) then click the respective points in the scatterplot. The x- and y-axes can be changed to represent any of the attributes through a drop down menu (D). Task instructions and interface guidance are presented in a tooltip when participants hover over the help button (E). As with the 15-puzzle game in Study 1, interactions with the system were logged including time stamped records of click, hover, and axis interactions. To ensure data quality, we required participants to classify at least 90% of data points (≥ 27) before proceeding to the next task.

Procedure. Participants were first presented with a practice task using a dog breeds dataset (categorize the dogs by breed: Bernedoodle, Shih Tzu and American Bulldog based on attributes such as amount of shedding, size, etc.) to become comfortable with the interface prior to completing the main tasks. Prior to beginning each task, participants needed to select the default attributes that would be displayed on the x- and y-axes for the initial visualization, to avoid biasing participants to use any particular attributes in their decision making. Additionally, we prefaced the task with an additional visualization literacy assessment including 7 multiple-choice questions (raw scores ranged from 0 to 7) specific to scatterplots adopted from VLAT [43] in the preliminary survey to ensure that participants could accurately interpret the scatterplot visualization.

Recruitment. We initially recruited 48 participants through the Prolific crowdsourcing platform based on a power analysis of pilot data with 12 participants who completed the categorization task using a similar experimental setup. Our minimum target sample size was 44 participants to obtain .8 power to detect a medium effect size of .25 at the standard $\alpha = 0.05$. Participants were initially compensated \$3.50 for the study which had an estimated duration of 20 minutes (actual median completion time of 25 minutes led to an adjustment of payment to \$4.85). Prolific allowed participants to spend a maximum of 67 minutes on the task; three participants timed out and their data was subsequently excluded. Participants were incentivized with an additional \$1 performance bonus respectively if they (1) completed a visualization literacy assessment survey with the highest correctness in the top 5% and (2) completed the categorization task with accuracy in the top 5%. 13 participants earned performance bonuses: 7 for top performance on

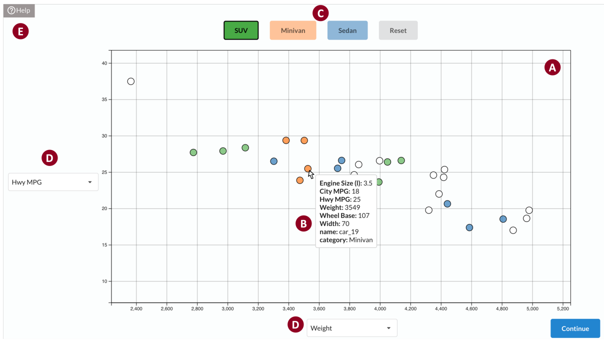


Fig. 6: Interactive scatterplot (A) that shows tooltips on hover (B). It also features category buttons for labeling (C), x- and y-axis dropdowns (D) and a help reminder of interface mechanics (E).

the visualization literacy assessment, 3 for the highest accuracy in the car task, and another 3 for the highest accuracy in the credit task.

No data were excluded from analyses due to failed attention checks in this study. However, we excluded data from participants with a visualization literacy score below 3 out of 7, deviating from our initial pre-registration plan of measuring DKE as a function of visualization literacy. This decision was based on the realization that a fundamental grasp of visualization literacy is a crucial prerequisite for meaningfully measuring DKE. Including participants with poor visualization literacy could undermine the study's integrity, akin to measuring DKE through a literature test presented in a language unfamiliar to the participants. Finally, two data points were excluded as outliers (outside 1.5 times the Interquartile Range). In total we used data from 46 individuals in our forthcoming analyses.

Participant Demographics. 12 participants identified as female, 33 identified as male, and 1 identified as non-binary. The majority of participants (31) hold a college degree (Bachelor's, Master's, or Doctorate), while 9 have some college or Associate's degree, and 6 have a High School diploma. Participants were on average 30.60 years old (SD = 9.48). After excluding individuals who achieved a scatterplot literacy score of less than 3 out of 7, the remaining participants achieved average scores of 5.18 (min = 3, max = 7) after applying the correction-for-guessing method [12, 23]. Participants reported average familiarity of 2.57 out of 5 for cars (35 participants rated 3 or lower) and 2.98 for credit (29 participants rated 3 or lower).

5.2 Results

In this study, we define performance based on categorization accuracy. Top performers (top quartile) are those individuals who achieved the highest categorization accuracy (largest number of points labeled correctly), while bottom performers (bottom quartile) are those who achieved the lowest categorization accuracy. To evaluate DKE, we compare participants' actual categorization accuracy to their perceived accuracy relative to their peers. Participants achieved on average 46.7% accuracy ($SD = 18.3\%$) in the car task and 32.59% accuracy ($SD = 10.2\%$) in the credit task. Participants spent approximately 3.47 minutes on the car task and 5.36 minutes on the credit task.

5.2.1 H1: DKE in Visual Reasoning

Consistent with the findings of DKE in the sliding puzzle game in Study 1, we likewise observe DKE for both the car and credit categorization tasks (Figure 1 (b and c, respectively)). Participants in the bottom quartile of the car task (Figure 1 (b), blue), with an average accuracy in the 15^{th} percentile, overestimated their accuracy (42^{th} percentile, Figure 1 (b), yellow) relative to their peers ($t = -3.36, p < 0.01$). Conversely, participants in the top quartile, who scored in the 90^{th} percentile on average, significantly underestimated their accuracy (37^{th} percentile) relative to their peers ($t = 6.19, p < 0.01$). A congruent pattern was observed in the credit task (Figure 1 (c)) as well, where bottom quartile participants, scoring in the 15^{th} percentile (Figure 1 (c), blue) on average, overestimated their accuracy (50^{th} percentile, Figure 1 (c), yellow) relative to their peers ($t = -4.27, p < 0.01$), while top quartile participants, scoring in the 90^{th} percentile on average underestimated their accuracy (57^{th} percentile) relative to their peers ($t = 4.95, p < 0.01$). This finding supports **H1**, less competent individuals overestimate their performance relative to peers, while competent individuals underestimate their performance.

Discussion of Results. These findings offer an empirical understanding of the relationship between self-assessment and actual performance in both low-skilled and high-skilled participants in an interactive scatterplot categorization task across two domains, contributing to the growing body of knowledge that DKE is a generalized pattern rather than a task-specific anomaly. Similar to Study 1, we considered an individual's general perceived **reasoning ability** relative to their peers (Fig 1(b and c, green)) and again found comparable results (car: bottom quartile ($t = -6.11, p < 0.01$), top quartile ($t = 5.45, p < 0.01$); credit: bottom quartile ($t = -4.00, p < 0.01$), top quartile ($t = 4.55, p < 0.01$)).

We note that accuracy in the credit task was chance, suggesting higher task difficulty. This could be due in part to the greater number of attributes increasing task complexity. It is unlikely participants categorically misunderstood the task given (1) the higher accuracy (47%) in the car task, and (2) the higher average self-reported credit familiarity level ($\mu_{car} = 2.57$ vs. $\mu_{credit} = 2.98$) with 22 participants rating 3 or above for cars and 32 participants rating 3 or above for credit. Detailed distributions can be found in supplemental materials.

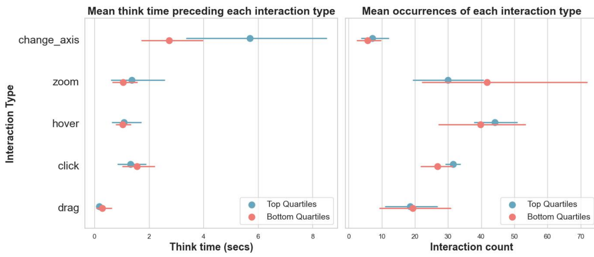
5.2.2 H2: Performance and Interactions

To test **H2**, our analysis was divided into three components of interactive behavior, including (1) rate of interaction with data labeling, (2) think time preceding an interaction, and (3) interaction sequences, which can reflect potential strategies employed by participants.

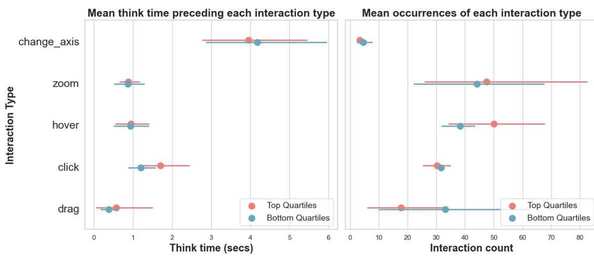
Interaction Rate. Previous research has identified distinct patterns in user behavior where some individuals tend to engage in a meticulous contemplation of each action, resulting in slow, deliberative manipulation of input devices, whereas others opt for more rapid execution [26]. In this study, top and bottom quartile participants in the car task registered 8.75 and 7.55 interactions per second respectively ($t = -0.566, p = 0.58$), while in the credit task, the corresponding rates were 6.02 and 8.19 interactions per second ($t = -0.541, p = 0.31$). In both scenarios, no significant differences in interaction rate were detected among top and bottom quartile.

Think Time. The *think time* metric is operationalized as the temporal interval from the end of one interaction to the initiation of the subsequent interaction. We found a significant difference in think time between two consecutive interactions for the top and bottom quartiles only in the credit task ($u = 1826896, p < 0.01$) using Mann-Whitney U test [44]. We further deconstructed this analysis by interaction type (e.g., considering clicks and hovers separately). As depicted in Figure 7, the left segment of each figure illustrates the mean think time preceding each type of interaction [1], and the right segment enumerates the average counts of the corresponding interaction types [1] for the car (a) and credit (b) tasks. Given that multiple pairwise comparisons were made, the Bonferroni correction [17] was applied to control the family-wise error rate, adjusting the significance level from 0.05 to 0.01. For think time by interaction type, no significant differences were observed between the two quartiles. Similarly, no significant differences in interaction counts were found between the two groups.

Interaction Sequences. Interaction *sequences* could reveal underlying strategies employed by users during the categorization task. To assess this, we computed a transition matrix, aimed at uncovering common sequential patterns of interactions. Figure 8 demonstrates the frequency of one interaction type (x-axis) succeeded by another interaction type (y-axis), stratified by both high-skilled participants and low-skilled participants in car (a) and credit (b) tasks. Each element within the transition matrix has been normalized relative to the cumulative sum of its corresponding row. This normalization means that each element now represents a proportion of the total for that row, ensuring a fair comparison of transition probabilities between different types of interactions (i.e., significantly more 'hover' interactions



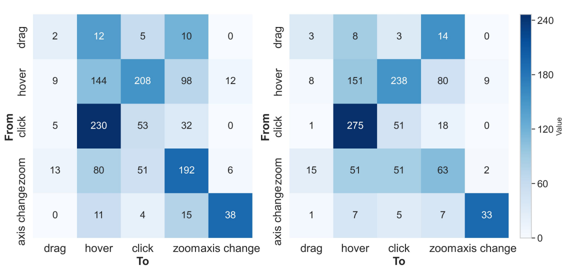
(a) Car



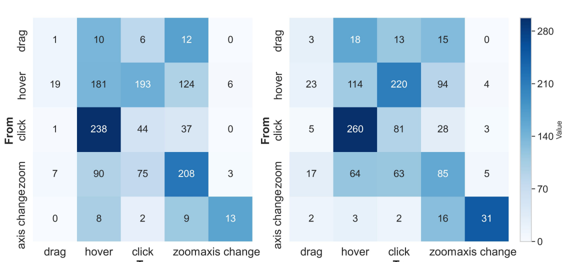
(b) Credit

Fig. 7: Average think time (left) preceding interaction types with 95% confidence interval and corresponding average interaction counts (right) with 95% confidence interval for the car (a) and credit (b) tasks.

occurred compared to 'change axis' interactions). Cells within the matrix are color-coded, with darker shades signifying sequences of interaction types that occur with greater frequency. We then flattened each matrix into one-dimensional vectors and calculated the Pearson Correlation Coefficient [51] between the two vectors, yielding a value of $r(22) = 0.93, p < 0.01$ for the car task, $r(22) = 0.91, p < 0.01$ for the credit task. The significant correlation between the two vectors suggests there were no real differences in interaction patterns across bottom and top quartiles in both tasks.



(a) Car Task: bottom (left) and top (right) quartiles



(b) Credit Task: bottom (left) and top (right) quartiles

Fig. 8: Transition matrix representing interaction sequences.

Discussion of Results. While discernible differences appeared in interaction patterns in the 15-puzzle game in Study 1, we observe relatively little distinction between interactive strategies of top and bottom quartile participants in the categorization tasks. We observed no differences in interaction rate or interaction sequences, and marginal differences in think time for some specific interaction types. Collectively, these findings provide little support for **H2**, that there are detectable differences in interactive strategies used by individuals who are more and less competent. Inconsistent differences observed across some measures of interactive strategy may be attributable to the heightened complexity of the task and breadth of available interactions leading to individual differences in task approach.

5.2.3 H3: Interaction and Personality

To test **H3**, we performed similar analysis as **H2**, focusing on three key dimensions of participants' decision-making behavior: interaction rate, think time, and interaction sequences; but in this analysis stratified by low and high score categories for five personality traits. Similar to Study 1, participants' personality trait scores were standardized.

Interaction Rate. We computed the interaction rate for high and low scores on five personality traits. The results are shown in Table 1. We applied Bonferroni correction [17] to control the family-wise error rate, adjusting the significance level from 0.05 to 0.01. However, no significant differences in interaction rates were observed between high and low scores for the five personality traits in either task.

	O		C		E		A		N	
Score/Task	Car	Credit	Car	Credit	Car	Credit	Car	Credit	Car	Credit
High	9.20	7.22	10.19	7.23	8.54	7.38	8.25	6.68	9.07	9.25
Low	9.43	6.14	8.61	8.88	10.21	6.83	10.67	8.70	8.81	7.37

Table 1: Interaction rate (interactions per second) for individuals with high/low scores on five personality traits.

Think Time. In an effort to understand the determinants of think time preceding each type of interaction, we fit linear mixed-effects models, with interaction type, personality traits and their interaction terms as fixed effects, and participants as a random effect. Five categories of interaction types ('click', 'drag', 'hover', 'zoom', and 'change axis') were included in the model as a categorical variable and the interaction type 'click' was used as the reference category in the coding scheme.

We can see the results stratified by high and low score in the five personality traits for the car task (Table 2) and credit task (Table 3). A lower score in Extraversion exhibits a significantly positive impact on the think time preceding the 'change axis' interaction in both tasks. Other traits also have significant effects on think time for one of the tasks. For example, in the car task, both higher and lower scores in the Conscientiousness trait are significantly positively correlated with think time preceding change axis interaction. This could suggest that the effect of Conscientious on think time is significantly modulated by the type of interaction. Specifically, the 'change axis' interaction might require more deliberation, making typically less conscientious individuals spend more time thinking, which highly conscientious individuals are inclined towards. However, these effects vary across the two tasks.

	O		C		E		A		N	
Task/Score	high	low	high	low	high	low	high	low	high	low
Change axis	-0.31	-1.40	**1.78***	**2.04****	1.08	**2.45***	-0.66	0.77	1.59	-0.43
Drag	0.81	0.52	-0.10	0.3	0.33	0.53	-0.39	-0.30	-0.43	0.38
hover	0.21	**0.92***	-0.17	0.56	0.37	**1.09***	0.19	-0.15	**−0.97***	0.03
Zoom	-0.32	0.12	-0.11	0.83	-0.19	0.66	0.72	0.25	-0.92	0.20

Table 2: Car: determinants of think time.

* represents p-value < 0.05; ** represents p-value < 0.01

Interaction Sequence. Utilizing a method analogous to that delineated in Section 5.2.2, we examined the interaction sequence and interaction attention among participants scoring high/low on five personality traits. We computed the transition matrix to identify underlying sequential patterns exhibited by the high- and low-scoring individuals. Significant differences were not detected in either of the two tasks.

	O		C		E		A		N	
Task/Score	high	low	high	low	high	low	high	low	high	low
Change axis	0.56	**3.13****	-0.91	-0.08	-0.98	**3.18****	-1.55	**−3.32****	-0.34	-0.17
Drag	-0.08	-0.01	-0.11	1.60	-0.28	0.32	-0.13	0.16	**−1.90****	0.09
hover	0.54	0.44	**0.74***	0.36	-0.62	-0.34	-0.59	-0.19	-0.44	-0.25
Zoom	0.09	0.15	0.46	**−1.56****	**−0.78***	-0.17	0.03	-0.47	-1.01	0.22

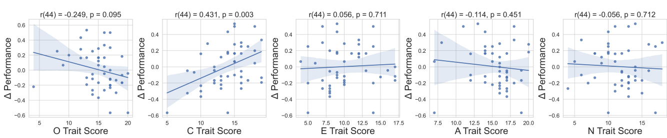
Table 3: Credit: determinants of think time.

* represents p-value < 0.05; ** represents p-value < 0.01

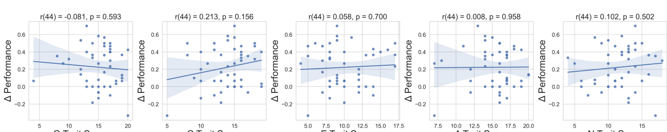
Discussion of Results. Although significant effects were observed for think time for specific personality traits combined with certain interaction types, these effects varied between the two tasks. Despite sharing commonalities in nature and setting, the unique domain-specific attributes and possible differences in task complexity likely contributed to these divergent outcomes. Thus, it is crucial to recognize that while some findings are statistically significant, they are specific to the contexts we studied. Overall, these findings provide mixed support for **H3**.

5.2.4 H4: Personality and Δ Performance

To test **H4**, we conducted a Pearson correlation analysis to discern if there was a correlation between individual personality traits and the susceptibility to DKE (Δ Performance = Estimated Percentile - Actual Percentile). As depicted in Figure 9, only the Conscientiousness (C) trait showed a significant effect in the car task ($r(44) = 0.431, p < 0.01$), While we observe some nonzero trend lines for other traits, the differences are not statistically significant. These findings provide little support for **H4** , that there are some correlations between personality traits and task performance.



(a) Car



(b) Credit

Fig. 9: Correlation between personality traits and ΔPerformance = Estimated Percentile – Actual Percentile) for (a) the car task, and (b) the credit task.

Discussion of Results. While we observed a notable relationship between Conscientiousness and miscalibration of perceived performance, this pattern was not evident in the credit task, which may suggest a context-dependent relationship, e.g., confounded by task difficulty.

5.2.5 H5: Δ Performance and Domain

To test **H5**, we consolidated the analysis across both the car and credit tasks to probe for a correlation between self-reported domain familiarity and the manifestation of DKE.

The Pearson correlation analysis revealed a significant positive correlation between the difference in performance and domain familiarity ($r(90) = 0.448, p < 0.01$) (see Supplemental Materials). This finding suggests that individuals who perceive themselves as more familiar with a specific domain may be likely to overestimate their abilities within that domain, whereas those who report less familiarity may conversely underestimate their capabilities, supporting **H5**, that people's overestimation of their performance is positively associated with their familiarity of the domain.

Discussion of Results. Individuals with higher self-reported familiarity in a specific domain were found to overestimate their abilities, which, in a broad sense, aligns with the general trend of DKE that people with lower ability at a task (using domain familiarity as a proxy for ability) tend to overestimate their ability [20]. On the other hand, the lack of significant correlation in Study 1 suggests that this trend might not be universally applicable across all domains or that other factors may influence the relationship between domain familiarity and accuracy of self-assessment.

5.2.6 Exploratory Analysis

To understand the extent to which our findings are attributable to the use of visualization specifically, rather than simply an artifact of the labeling task, we conducted additional exploratory analyses to analyze how users engaged with the axes of the scatterplot by selecting different attribute pairs. Some pairs are inherently more informative for the tasks, i.e., would produce more clear clustering of correctly labeled points. To quantify this, we calculated the ratio of inter-class to intra-class distances for each attribute pair. Here, inter-class distance is the average distance between category centroids, and intra-class distance measures the distance of data points from their category's centroid [45]. A higher ratio signifies better category separability, which we will refer to as a more informative attribute pair. Analysis details are in the Supplemental Materials. We found that top performers tended to more often choose more informative attribute combinations, suggesting that the interactions with the axes and the resulting visual relationship between attributes likely contributed to the disparity in performance.

For instance, in the car task, top performers most often chose the combination Weight × Engine Size with a ratio of 2.6 compared to the bottom performers' most chosen combination of Wheel Base × Engine Size with a less informative ratio of 1.35. Likewise, in the credit task, top performers most often chose the combination Number of Loans × Outstanding Debt with a ratio of 2.54 compared to the bottom performers' most chosen combination of Monthly Balance × Number of Delayed Payments with a less informative ratio of 1.02. These disparities in ratios could signify the successful use of visual clustering strategies for the labeling task.

Our findings further showed that in both tasks, 30 out of 46 participants interacted with the axes more than the default requirement (twice, to set the initial configuration), indicating a more meaningful engagement with the interactive axes. Analysis of these participants revealed varied strategies: some participants employed explicit spatial techniques, as evidenced by feedback like, "I tried to organize them into sensible groups using the axis to sort them," highlighting a deliberate manipulation of visual components. In contrast, others followed more ambiguous methods not directly tied to visualization, e.g., one participant said "I decided to categorize credit scores based on their monthly balance and number of delayed payments, I believe. I think these were the most important among all other factors." While we cannot assert with absolute certainty that mentioning specific attribute pairs equates to direct interaction with visual elements, such references suggest an inclination towards visual analysis rather than simply labeling in the absence of the visual analysis setting. Moreover, the widespread use of visual components among participants underscores the critical role of visualization in their decision-making process. By manipulating axes, identifying patterns, and grouping data visually, users are engaging in a form of visual reasoning that leverages spatial relationships and graphical representations to draw conclusions.

We also found that DKE persisted even when we focused our exploration on the subset of 30 participants who interacted more heavily with the axes, doing so at least twice. Particularly, we observed a significant overestimation of performance by the lower-performing group in both the car task ($t = -2.15, p = 0.049$) and the credit task ($t = -3.09, p < 0.01$). Conversely, the higher-performing group significantly underestimated their abilities in both the car ($t = 4.55, p < 0.01$) and credit ($t = 3.07, p < 0.01$) tasks. Collectively, these exploratory findings increase our confidence that the observed phenomenon is attributable to DKE in judgments facilitated by visualization.

6 DISCUSSION

Implications of DKE in Visualization. Across two experimental contexts we observed DKE, the systematic miscalibration of perceived ability by top and bottom performers. Because DKE has been observed in many diverse domains [18, 20, 56], it is somewhat unsurprising that DKE appears to affect the visualization context as well. In fact, the categorization task in Study 2 taken outside the interactive interface is not unlike some academic contexts where DKE has been previously observed, e.g., in written exams [42]. Nonetheless, there are critical differences that the contexts of the present studies afford, namely through analyses of DKE with respect to interactive strategies and personality traits. The observable differences in interaction patterns across proficiency levels, such as movement paths in the 15-puzzle game (Figure 3) and counts for interaction types (Figure 7) in the categorization task suggest ways that DKE may uniquely influence interactive behaviors or, conversely, that interactive behaviors may be indicative of susceptibility to DKE.

Similarly, some personality traits, specifically people who exhibit high Conscientiousness may be more prone to a miscalibration between ability and perception, which was particularly noted in the contexts of the puzzle game and the car task. These findings enhance our understanding of visual reasoning and judgment and suggest potential applications such as tailored guidance and bias mitigation, specifically by adapting to a user's proficiency level and personality traits.

For instance, consider a financial investment platform designed to empower users to make informed decisions about their investment strategies that dynamically adapts to an individual's susceptibility to biases like DKE. Individuals with a tendency to over-estimate their abilities (e.g., novices, individuals with high conscientiousness, etc.) could benefit from techniques like subjective probability correction, wherein the visuals shown to users are adjusted to counteract biased interpretations of uncertainty in the data [73]. High-skilled individuals, on the other hand, could benefit from systems that implement techniques to boost confidence in their visualization designs, interpretations of data, etc. using psychological skill training techniques such as goal-setting or positive self-talk [32]. In doing so, this provides a tailored experience that compensates for an individual's tendency to over- or underestimate their own abilities. Moreover, incorporating personality trait assessments into the platform design could enable interfaces that are more responsive to individual differences, offering personalized feedback to help users accurately evaluate their investment strategies. This approach enables designers and developers to create more intuitive and effective interfaces that cater to a diverse range of skill levels and cognitive styles. However, systems that utilize these personalized strategies would need to take precautions to protect the privacy of users.

Metacognitive Bias or Statistical Artifact? The underlying cause driving the DKE continues to be a matter of intense debate among researchers. Some critics of DKE argue that the self-assessment errors observed by Kruger and Dunning can be largely reduced to statistical artifacts rather than true metacognitive deficits [5, 41]. Specifically, Krueger et al. argue that a combination of a statistical artifact known as "regression toward the mean" and a "better-than-average" heuristic might explain the observed gaps between actual and perceived performance, particularly the larger discrepancies at lower skill levels. This occurs as imperfect correlations between actual and perceived performance inevitably lead the self-assessments of low performers to regress back toward the average, further amplified by the common belief that one is above average, while high performers tend to underestimate theirs due to regression to the mean, somewhat counterbalanced by the same better-than-average belief. Consequently, high performers appear to make more accurate self-assessments than low performers [41]. But our findings from Study 2 differ from the anticipated pattern, with larger gaps for higher skilled participants in one of the tasks. Specifically, in the car task, the discrepancies (Δ = Estimated Percentile − Actual Percentile) were $\Delta_{top} = -53$ and $\Delta_{bottom} = 27$ percentile points, and the credit task displayed discrepancies of $\Delta_{top} = -33$ and $\Delta_{bottom} = 35$ percentile points. While an asymmetry was indeed detected, the gap at the lower end was considerably smaller in the less challenging car task ($u_{accuracy} = 46.7\%$) and slightly larger in the more demanding credit task ($u_{accuracy} = 32.59\%$). This suggests that in the easier task, less skilled individuals exhibited better calibration, marked by a smaller discrepancy, while in the harder task, their calibration was less accurate compared to those with higher performance, evidenced by a slightly greater discrepancy. Contrary to another criticism that highlighted the instrumental role of task difficulty on the asymmetry in DKE [5]—wherein less skilled individuals were thought to have better calibration in moderately difficult tasks compared to higher performers—our results suggest a reversal of this relationship.

In response to criticisms above, supporters argue that even after adjusting for statistical reliability concerns in real-world tasks with ecological validity, the DKE pattern still persists, albeit slightly attenuated, but does not disappear [19, 20].

Confounding Domain Familiarity? Does over- and under-estimation of performance correspond to domain familiarity? To assess whether domain familiarity introduced confounding effects, we analyzed H1 by stratifying the data based on different levels of familiarity across the two studies. Given the limited data available, we divided participants into two groups: those with a familiarity rating of ≥ 3, and those with a rating of ≤ 2 (on a scale of 1-5). Consistent and significant DKE trends were observed in both studies across the divided groups, with the exception of the bottom quartile in the credit task for those who rated familiarity ≤ 2, where the result followed the trend of overestimation in the bottom quartile, but was not statistically significant for this group in this task ($t = -0.97, p = 0.37$). This finding invites further investigation into how task-specific factors and the level of domain familiarity influence self-perception of skills and performance. Details of this analysis are provided in the supplemental materials.

Limitations and Future Work. One limitation of our studies is that we focused primarily on interaction sequences and rate as measures of interactive strategy. However, there are many other facets of interactive behavior, potentially influenced by DKE, that were not captured in this analysis. This could include measures such as participants' error correction frequency (how often participants change their labels), which can reflect their confidence and self-awareness. Moreover, passive interactions, such as gaze patterns, may also be revealing of underlying strategies. The analysis of gaze can serve as a useful indicator of attention and cognitive processing, complementing explicit interaction patterns. To delve into this aspect, we conducted an exploratory eye-tracking analysis, focusing on passive interaction patterns through gaze. Further details can be found in the Supplemental Materials.

Additionally, future work could explore methodologies for integrating gaze data with other interaction metrics in more nuanced ways. This might involve developing new analytic techniques or machine learning models that take into account both the users' active interactions and their passive gaze behaviors together. Another valuable direction for future work is to develop and evaluate interventions to mitigate the effects of DKE. Potential strategies could include designing adaptive interfaces that respond to real-time analysis of DKE-related behaviors or provide feedback on user performance such as in the form of peer percentile rankings, thus aiding in the calibration of their self-assessment.

7 CONCLUSION

Across two online studies focusing on visual reasoning and judgment, we observed the Dunning-Kruger Effect. Specifically, two extreme performance groups misjudged their abilities: the bottom quartile tended to overestimate, while the top quartile tended to underestimate their performance. Our results suggest that there are some observable differences in interactive strategies employed by individuals that corresponds with high and low performance across the two studies. We also discovered certain personality traits such as conscientiousness significantly correlate to the susceptibility to DKE in some of the tasks, which are prone to underestimate or overestimate one's performance. The findings from these two studies contribute to an empirical foundation for future personalized interventions to improve the visual data analysis process rooted in personality traits and interactive strategies.

Acknowledgments

This work was partially supported by NSF award IIS-2340539.

References

[1] L. Battle and J. Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. In *Computer graphics forum*, vol. 38, pp. 145–159. Wiley Online Library, 2019. doi: 10.1111/cgf.13678

[2] E. Blakey and S. Spence. *Developing metacognition*. ERIC Clearinghouse on Information and Technology, 1990.

[3] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics*, 20(12):1663–1672, 2014. doi: 10.1109/TVCG.2014.2346575

[4] A. Brüngger, A. Marzetta, K. Fukuda, and J. Nievergelt. The parallel search bench zram and its applications. *Annals of Operations Research*, 90(0):45–63, 1999. doi: 10.1023/A:1018972901171

[5] K. A. Burson, R. P. Larrick, and J. Klayman. Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of personality and social psychology*, 90(1):60, 2006. doi: 10.1037/0022-3514.90.1.60

[6] Y.-C. Chen, C.-Y. Chao, and H.-T. Hou. Learning pattern recognition skills from games: Design of an online pattern recognition educational mobile game integrating algebraic reasoning scaffolding. *Journal of Educational Computing Research*, 61(6):1232–1251, 2023. doi: 10.1177/07356331231171622

[7] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 116–126. IEEE, Phoenix, Arizona, USA. doi: 10.1109/VAST.2017.8585665

[8] L. H. Christoph et al. *The role of metacognitive skills in learning to solve problems*. SIKS, 2006.

[9] P. Cowley, L. Nowell, and J. Scholtz. Glass box: An instrumented infrastructure for supporting human interaction with information. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 296c–296c. IEEE, Hawaii, USA, 2005. doi: 10.1109/HICSS.2005.286

[10] M. Danesi and M. Danesi. Puzzles and spatial reasoning. *Ahmes' Legacy: Puzzles and the Mathematical Mind*, pp. 105–125, 2018. doi: 10.1007/978-3-319-93254-5_4

[11] D. A. Davis, P. E. Mazmanian, M. Fordis, R. Van Harrison, K. E. Thorpe, and L. Perrier. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *Jama*, 296(9):1094–1102, 2006. doi: 10.1001/jama.296.9.1094

[12] J. Diamond and W. Evans. The correction for guessing. *Review of educational research*, 43(2):181–191, 1973. doi: 10.3102/00346543043002181

[13] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri. Mitigating the attraction effect with visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):850–860, 2019. doi: 10.1109/TVCG.2018.2865233

[14] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic. A task-based taxonomy of cognitive biases for information visualization. *IEEE transactions on visualization and computer graphics*, 26(2):1413–1432, 2020. doi: 10.1109/TVCG.2018.2872577

[15] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192, 2006. doi: 10.1037/1040-3590.18.2.192

[16] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering reasoning processes from user interactions. *IEEE Computer Graphics and Applications*, 29(3):52–61, 2009. doi: 10.1109/MCG.2009.49

[17] O. J. Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090

[18] D. Dunning. The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, vol. 44, pp. 247–296. Elsevier, 2011. doi: 10.1016/B978-0-12-385522-0.00005-6

[19] D. Dunning, K. Johnson, J. Ehrlinger, and J. Kruger. Why people fail to recognize their own incompetence. *Current directions in psychological science*, 12(3):83–87, 2003. doi: 10.1111/1467-8721.01235

[20] J. Ehrlinger, K. Johnson, M. Banner, D. Dunning, and J. Kruger. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes*, 105(1):98–121, 2008. doi: 10.1016/j.obhdp.2007.05.002

[21] M. Feng, E. Peck, and L. Harrison. Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE transactions on visualization and computer graphics*, 25(1):501–511, 2019. doi: 10.1109/TVCG.2018.2865117

[22] J. H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979. doi: 10.1037/0003-066x.34.10.906

[23] R. B. Frary. Formula scoring of multiple-choice tests (correction for guessing). *Educational measurement: Issues and practice*, 7(2):33–38, 1988. doi: 10.1111/j.1745-3992.1988.tb00434.x

[24] G. Gigerenzer and D. G. Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996. doi: 10.1037/0033-295x.103.4.650

[25] G. Gigerenzer and P. M. Todd. *Simple heuristics that make us smart*. Oxford University Press, USA, 1999. doi: 10.1002/acp.843

[26] A. Goguey, C. Gutwin, Z. Chen, P. Suwanaposee, and A. Cockburn. Interaction pace and user preferences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, Virtual Conference. doi: 10.1145/3411764.3445772

[27] L. R. Goldberg. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28, 1999.

[28] S. Gomez and D. Laidlaw. Modeling task performance for a crowd of users from interaction histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2465–2468. ACM, Austin, Texas, USA, 2012. doi: 10.1145/2207676.2208412

[29] D. Gotz, S. Sun, and N. Cao. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 85–95. ACM, Sonoma, California, USA, 2016. doi: 10.1145/2856767.2856779

[30] T. M. Green and B. Fisher. Towards the Personal Equation of Interaction: The impact of personality factors on visual analytics interface interaction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 203–210. IEEE, Salt Lake City, Utah, USA. doi: 10.1109/VAST.2010.5653587

[31] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi: 10.1109/TSSC.1968.300136

[32] A. Heydari, H. Soltani, and M. Mohammadi-Nezhad. The effect of psychological skills training (goal setting, positive selftalk and imagery) on self-confidence of adolescent volleyball players. *Pedagogics, psychology, medical-biological problems of physical training and sports*, 22(4):189–194, 2018. doi: 10.15561/18189172.2018.0404

[33] S. Huang. When peers are not peers and don't know it: The dunning-kruger effect and self-fulfilling prophecy in peer-review. *Bioessays*, 35(5):414–416, 2013. doi: 10.1002/bies.201200182

[34] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics*, 25(1):903–913, 2019. doi: 10.1109/TVCG.2018.2864889

[35] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011. doi: 10.1037/h0099210

[36] D. Kahneman and S. Frederick. A model of heuristic judgment., 2005.

[37] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Visual analytics: Definition, process, and challenges*. Springer, 2008. doi: 10.1007/978-3-540-70956-5_7

[38] H. Kim, J. Choo, H. Park, and A. Endert. Interaxis: Steering scatterplot axes via observation-level interaction. *IEEE transactions on visualization and computer graphics*, 22(1):131–140, 2016. doi: 10.1109/TVCG.2015.2467615

[39] Y.-S. Kim, K. Reinecke, and J. Hullman. Data through others' eyes: The impact of visualizing others' expectations on visualization interpretation. *IEEE transactions on visualization and computer graphics*, 24(1):760–769, 2018. doi: 10.1109/TVCG.2017.2745240

[40] Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman. A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM,

Glasgow, Scotland, UK. doi: 10.1145/3290605.3300912

[41] J. Krueger and R. A. Mueller. Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of personality and social psychology*, 82(2):180, 2002. doi: 10.1037/0022-3514.82.2.180

[42] J. Kruger and D. Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999. doi: 10.1037/0022-3514.77.6.1121

[43] S. Lee, S.-H. Kim, and B. C. Kwon. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920

[44] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947. doi: 10.1214/aoms/1177730491

[45] M. Michael and W.-C. Lin. Experimental study of information measure and inter-intra class distance ratios on feature selection and orderings. *IEEE Transactions on Systems, Man, and Cybernetics*, (2):172–181, 1973. doi: 10.1109/TSMC.1973.5408500

[46] A. V. Moere, M. Tomitsch, C. Wimmer, B. Christoph, and T. Grechenig. Evaluating the effect of style in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2739–2748, 2012. doi: 10.1109/TVCG.2012.221

[47] D. A. Moore and P. J. Healy. The trouble with overconfidence. *Psychological review*, 115(2):502–517, 2008. doi: 10.1037/0033-295X.115.2.502

[48] 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*, 28(1):1009–1018, 2022. doi: 10.1109/TVCG.2021.3114827

[49] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink. Analytic provenance: process+interaction+insight. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, pp. 33–36. ACM, Vancouver, BC, Canada, 2011. doi: 10.1145/1979742.1979570

[50] S. R. Pavel, M. F. Robertson, and B. T. Harrison. The dunning-kruger effect and siuc university's aviation students. *Journal of Aviation Technology and Engineering*, 2(1):6, 2012. doi: 10.5703/1288284314864

[51] K. Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895. doi: 10.1098/rspl.1895.0041

[52] V. Peña-Araya, A. Bezerianos, and E. Pietriga. A comparison of geographical propagation visualizations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, Honolulu, Hawaii, USA. doi: 10.1145/3313831.3376350

[53] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics*, 22(1):31–40, 2016. doi: 10.1109/TVCG.2015.2467551

[54] M. Rahmani. Medical trainees and the dunning–kruger effect: when they don't know what they don't know. *Journal of Graduate Medical Education*, 12(5):532–534, 2020. doi: 10.4300/JGME-D-20-00134.1

[55] B. W. Roberts, C. Lejuez, R. F. Krueger, J. M. Richards, and P. L. Hill. What is conscientiousness and how can it be assessed? *Developmental psychology*, 50(5):1315, 2014. doi: 10.1037/a0031109

[56] C. Sanchez and D. Dunning. Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of personality and Social Psychology*, 114(1):10, 2018. doi: 10.1037/pspa0000102

[57] D. J. Sanchez and A. Speed. The impact of individual traits on domain task performance: Exploring the dunning-kruger effect. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2020. doi: 10.2172/1716582

[58] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. doi: 10.2307/2333709

[59] The Decision Lab. Dunning–kruger effect. `https://thedecisionlab.com/biases/dunning-kruger-effect`, 2021. Retrieved March 7, 2024.

[60] E. R. Thompson. Development and validation of an international english big-five mini-markers. *Personality and individual differences*, 45(6):542–548, 2008. doi: 10.1016/j.paid.2008.06.013

[61] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124

[62] K. Umbleja, M. Ichino, and H. Yaguchi. Improving symbolic data visualization for pattern recognition and knowledge discovery. *Visual Informatics*, 4(1):23–31, 2020. doi: 10.1016/j.visinf.2019.12.003

[63] S. VanderPlas and H. Hofmann. Spatial reasoning and data displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):459–468, 2016. doi: 10.1109/TVCG.2015.2469125

[64] E. Wall, A. Arcalgud, K. Gupta, and A. Jo. A markov model of users' interactive behavior in scatterplots. In *2019 IEEE Visualization Conference (VIS)*, pp. 81–85. IEEE. doi: 10.1109/VISUAL.2019.8933779

[65] E. Wall, L. Blaha, C. Paul, and A. Endert. A formative study of interactive bias metrics in visual analytics using anchoring bias. In *IFIP Conference on Human-Computer Interaction*, pp. 555–575. Springer, Paphos, Cyprus, 2019. doi: 10.1007/978-3-030-29384-0_34

[66] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 ieee conference on visual analytics science and technology (vast)*, pp. 104–115. IEEE, Phoenix, Arizona, USA, 2017. doi: 10.1109/VAST.2017.8585496

[67] E. Wall, L. M. Blaha, C. L. Paul, K. Cook, and A. Endert. Four perspectives on human bias in visual analytics. *Cognitive biases in visualizations*, pp. 29–42, 2018. doi: 10.1007/978-3-319-95831-6_3

[68] E. Wall, A. Narechania, A. Coscia, J. Paden, and A. Endert. Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE transactions on visualization and computer graphics*, 28(1):966–975, 2022. doi: 10.1109/TVCG.2021.3114862

[69] E. Wall, J. Stasko, and A. Endert. Toward a design space for mitigating cognitive bias in vis. In *2019 IEEE Visualization Conference (VIS)*, pp. 111–115. IEEE, Vancouver, BC, Canada. doi: 10.1109/VISUAL.2019.8933611

[70] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, 2nd ed., 2004. doi: 10.1016/B978-155860819-1/50001-7

[71] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE transactions on visualization and computer graphics*, 13(6):1129–1136, 2007. doi: 10.1109/TVCG.2007.70589

[72] C. Xiong, L. Van Weelden, and S. Franconeri. The curse of knowledge in visual data communication. *IEEE transactions on visualization and computer graphics*, 26(10):3051–3062, 2020. doi: 10.1109/TVCG.2019.2917689

[73] F. Yang, M. Hedayati, and M. Kay. Subjective probability correction for uncertainty representations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17. Hamburg, Germany. doi: 10.1145/3544548.3580998

[74] J. S. Yi, Y. ah Kang, J. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007. doi: 10.1109/TVCG.2007.70515

[75] Z. Zhou, X. Wen, Y. Wang, and D. Gotz. Modeling and leveraging analytic focus during exploratory visual analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, article no. 21, pp. 1–15. ACM, Yokohama, Japan. doi: 10.1145/3411764.3445674

[76] C. Ziemkiewicz, A. Ottley, R. J. Crouser, A. R. Yauilla, S. L. Su, W. Ribarsky, and R. Chang. How visualization layout relates to locus of control and other personality factors. *IEEE transactions on visualization and computer graphics*, 19(7):1109–1121, 2013. doi: 10.1109/TVCG.2012.180