Supplemental Figures

Unmasking Dunning-Kruger Effect in Visual Reasoning and Visual Data Analysis

In this document, we include results of additional analyses of the data.

Study 1

H1: Success Measured by Actual Time Spent

We also consider that participants who performed the fewest movements might not spend the least time, as they are likely to spend more time strategizing before moving. Thus, we conducted additional analyses based on ***time spent*** to see if there are visible differences in results. Depicted as Fig. 1, when success is measured by time spent, similar trends can also be observed. Participants who ranked in the bottom quartile overestimated the efficiency of their time spent (yellow dotted line, left) to be around in the 60th (t = -3.72, p < 0.002), and reasoning ability (green dotted line, left) compared to their peers to be around in the 65th percentiles (t = -5.15, p < 0.01). Similarly, participants in the top quartile underestimated all their time spent and ability (yellow and green dotted line respectively, right) to be around 40th (t = 6.65, p < 0.001) and 50th (t = 4.54, p < 0.01), respectively. We found significant differences between the perceived and actual time spent among both the top and bottom quartiles. This measure also supports **H1.**

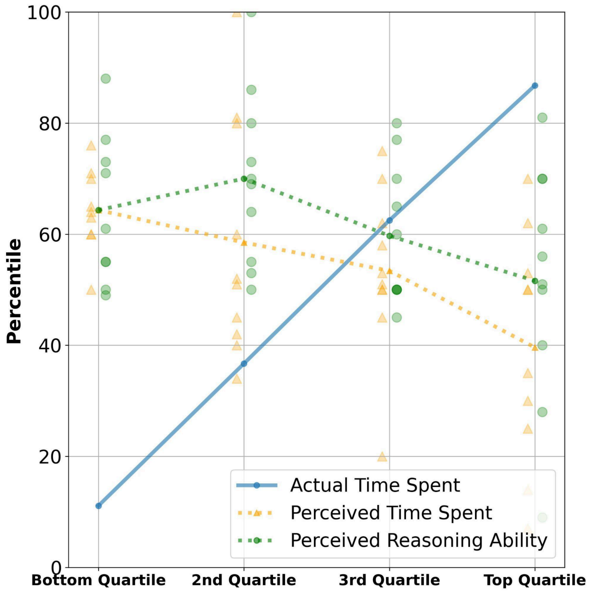


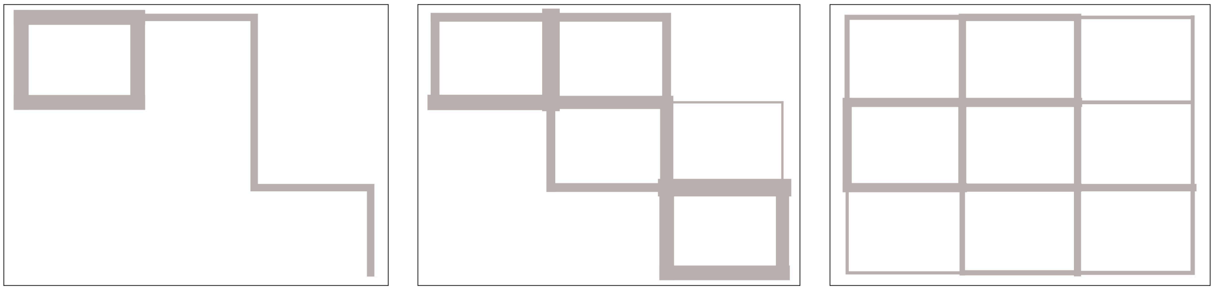
Fig. 1: DK results measured by time spent. X-axis depicts actual performance and y-axis depicts perceived performance. The blue line shows a baseline if participants perceived the performance accurately. The yellow line

depicts participants' perceived performance in a sliding puzzle game. The green colors depict perceived reasoning ability.

H2: Interaction of Individuals Who Did Not Achieve an Optimal Solution

We conducted additional exploratory analysis on the subset of 16 participants who did not achieve an optimal solution to identify if more nuanced differences in strategies could become apparent. We grouped them based on move counts: exactly 12 (n = 6), between 13 - 40 (n = 5), and more than 40 (n = 5).

As depicted by Figure 2, people with greater move counts showed interaction patterns that were relatively evenly distributed across the whole puzzle board (c); while those with lower move counts tended not to move tiles among the top right and bottom left segments of the board. This could be due to strategies that involved placement of some fixed tiles in those areas of the board, or realization by participants that movement in these segments would not lead to an efficiently completed puzzle board without subsequently undoing those interactions.



(a) 12 Moves (b) 13 - 40 Moves (c) Above 40 Moves

Fig. 2: Interaction strategies employed by the 16 participants who did not achieve the optimal solution.

H4: Correlation Between Personality Traits and Difference Performance

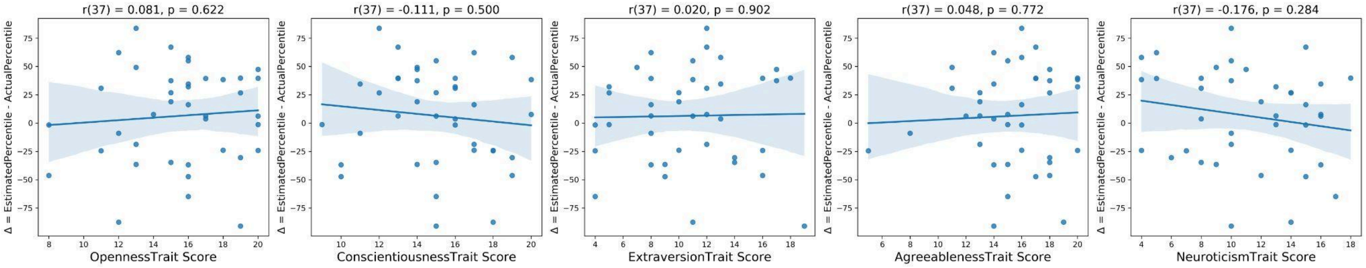


Fig. 3: Correlation between personality traits and difference performance.

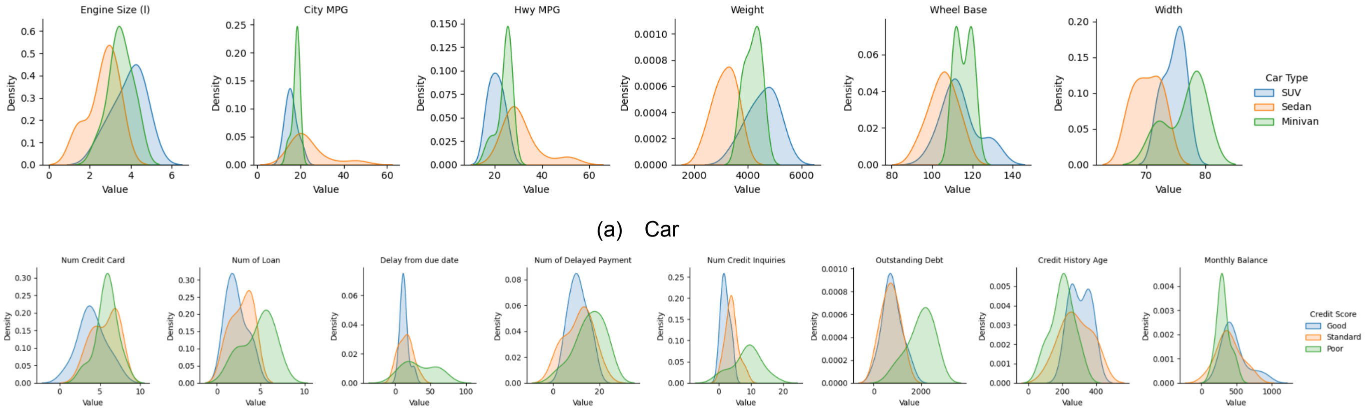
As depicted by Fig. 3, altering our selection method to a random choice of 9 from the 23 participants who attained the optimal move count as the top quartile—instead of further differentiating by task completion time—eliminated the statistical significance for all personality traits. This suggests that the significant association with conscientiousness might be dependent on the performance metrics we have adopted, or it could be an artifact of a spurious correlation. This finding underscores the complexity and nuanced nature of task performance assessment. It suggests that the criteria chosen to define and measure performance are not just methodological details but can fundamentally influence the interpretation and understanding of the relationship between personality traits and task outcomes.

Study 2

Dataset

	Car	Credit
1	Engine Size	Num of Credit Cards
2	City MPG	Num of Loan
3	Hwy MPG	Num of Delayed Payment
4	Weight	Num of Credit Inquiries
5	Wheel Base	Outstanding Debt
6	Width	Monthly Balance
7		Delay from Due Date
8		Credit History Age

Table 1. Dataset attributes.



(a) Car

(b) Credit

Fig. 4: Density distributions for each attribute separated by category

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, as detailed in Table.1. The data points and attributes were selected by inspection such that there was varying separability of the classes based on attributes of the data, as depicted by Fig. 4.

H2: Interaction Rate

Inspired by Feng et al. [1], we also include additional analysis about transforming a time-series interaction showing participants' activities over time into frequencies and calculating an average of the users' high-frequency powers.

[1] Feng, Peck, and Harrison, “Patterns and Pace.”

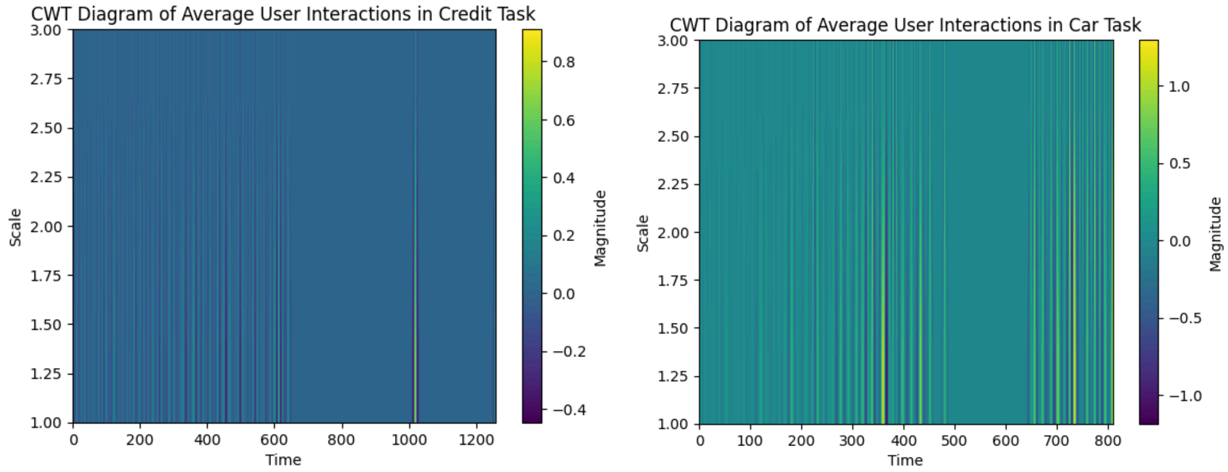


Fig. 5: CWT diagrams across two studies.

The CWT Diagrams (see Fig. 5) demonstrate the interaction pace for users in car and credit tasks on average. The x-axis represents the time differences between the interactions, and the y-axis is a scale to capture frequencies. The color represents the magnitude of the wavelet transform at each point: higher magnitudes suggest significant activity or patterns at that scale and time.

The pacingHF (users' high-frequency powers) for car task users that are in the bottom in task performance is 0.0133, whereas the number goes up to 0.0158 for the top performing users in the car task; the users in the credit task who are low-performing have pacingHF of 0.0171, whereas the for the high-performing users is 0.0244. For both car and credit task, users in the top quartile tend to have a comparatively higher pacingHF. Additionally, pacingHF is higher in the credit task, meaning that the users in the credit tasks have higher frequencies in interactions on average.

H5: Performance and Domain Familiarity

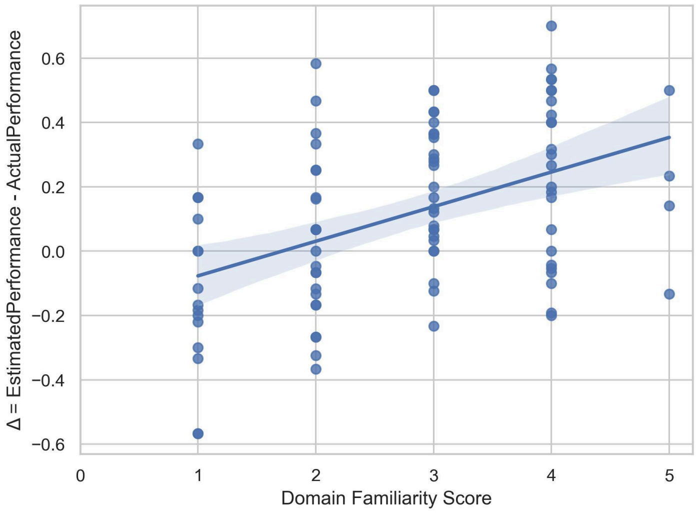


Fig. 6: Correlation between domain familiarity and DKE

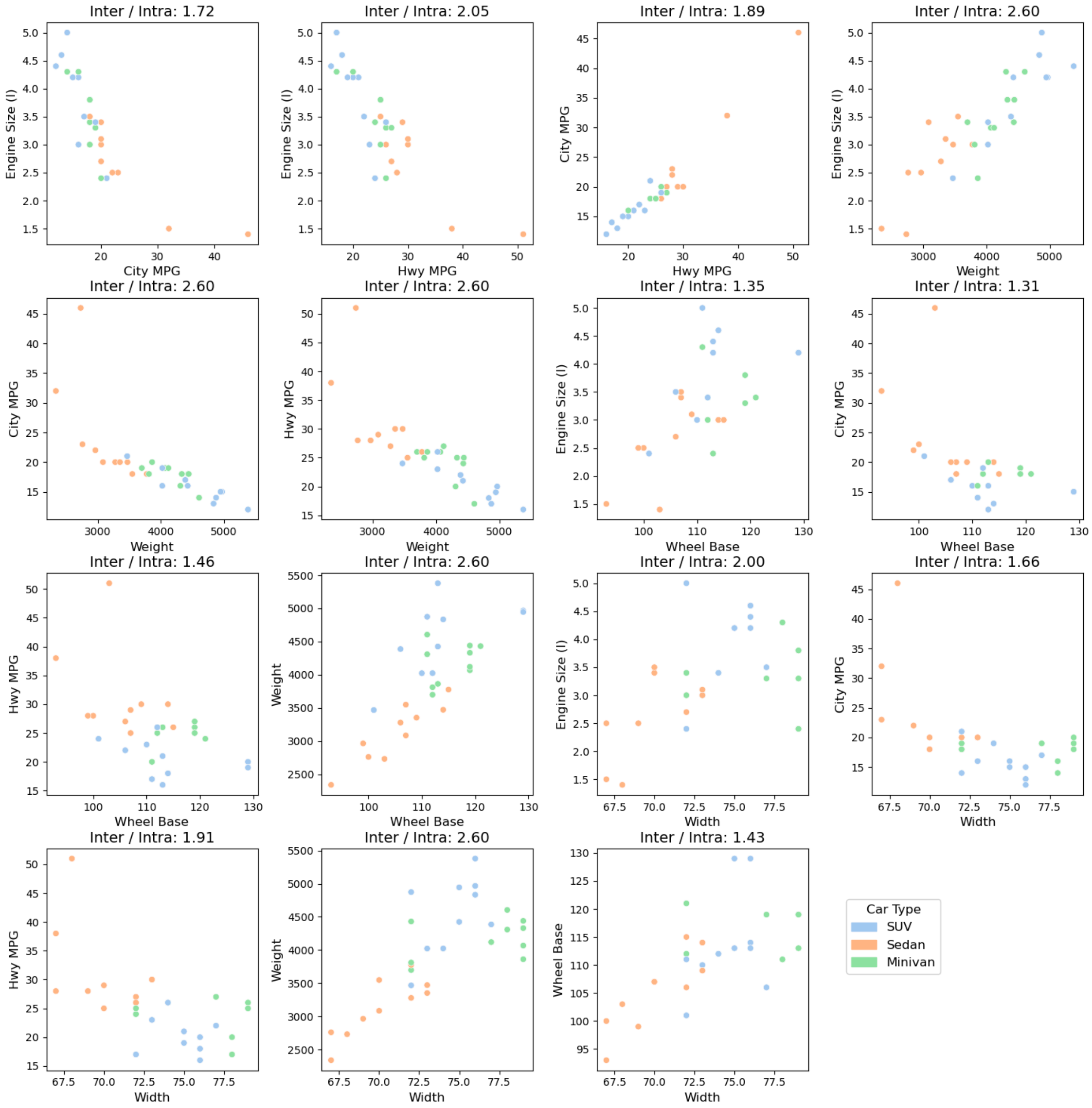
To test H5, we consolidated the analysis across both the car and credit tasks to probe for a correlation between domain familiarity and the manifestation of DKE. As depicted by Fig. 6, the x-axis represents self-reported domain familiarity scores and the y-axis denotes the discrepancy between perceived and actual performance. The Pearson correlation analysis revealed a significant positive correlation between the two variables (r(90)=0.448, p < 0.01). 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. This supports H5, that people's overestimation of their performance is positively associated with their familiarity of the domain.

DKE Stratified by Domain Familiarity Levels

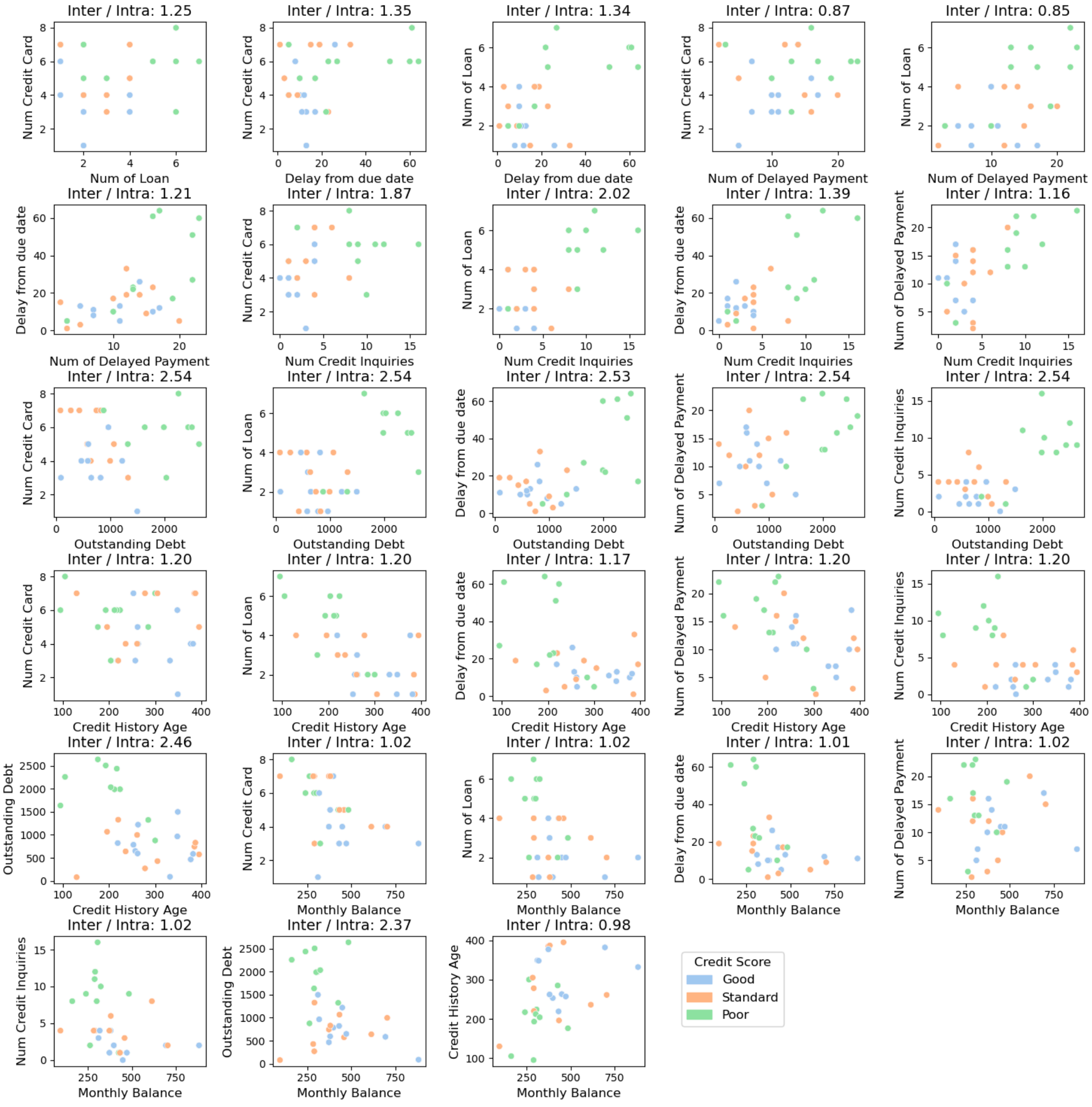
To assess whether domain familiarity introduced confounding effects, we analyzed H1 by stratifying the data based on different levels of familiarity. Given the limited data available, we divided participants into two groups: those with a familiarity rating of 3 or above, and those with a rating of 2 or below. In Study 1, we found a significant and consistent DKE trend (bottom quartile: $t = -2.58, p = 0.028$, top quartile: $t = 2.69, p = 0.023$) for familiarity level rating 2 or below and (bottom quartile: $t = -5.53, p < 0.01$, top quartile: $t = 6.11, p < 0.01$) for familiarity level rating 3 or above. We also observed a significant and consistent trend in Study 2 car task, with bottom quartile: $t = -2.55, p = 0.029$, top quartile: $t = 6.63, p < 0.01$ for familiarity level rating 2 or below and bottom quartile: $t = -6.21, p < 0.01$, top quartile: $t = 3.74, p < 0.01$ for familiarity level rating 3 or above. For Study 2 credit task, We also observed a consistent trend with bottom quartile: $t = -6.20, p < 0.01$, top quartile: $t = 3.45, p < 0.01$ for familiarity level rating 3 or above. For familiarity level ratings of 2 or below, a significant and consistent trend was observed only among the top quartile. The bottom quartile also demonstrated a trend of overestimation, but this was not statistically significant (bottom quartile: $t = -0.97, p = 0.37$, top quartile: $t = 15.78, p < 0.01$).

Exploratory Analysis

We analyzed how participants engaged with the axes of the scatterplot, which enabled them to view the 30 data points from various angles 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 assessed the clarity of clustering for each attribute pair by calculating the ratio of inter-class to intra-class distances. The inter-class distance measures the average distance between the centroids of the three distinct categories, while the intra-class distance assesses how far each data point is from its category's centroid. A higher ratio indicates better separability among the three categories for a given pair of attributes, as depicted in Fig.7. In the car task (a), the attribute pairs that yielded the highest ratio of 2.6 were Weight x Engine Size, Weight x City MPG, Weight x Hwy MPG, Weight x Wheel Base, and Weight x Width. Notably, 33 out of 46 participants engaged with at least one of these most informative attribute pairs. In the credit task (b), the highest ratio is 2.54 and corresponding pairs were Num of Credit Card x Outstanding Debt, Num of Load x Outstanding Debt, Num of Delayed Payment x Outstanding Debt, Num of Credit Inquiries x Outstanding Debt. Here, 16 out of 46 participants interacted with at least one of the specific attribute pairs.


Inter / Intra: 1.72
Engine Size (l)
City MPG
Inter / Intra: 2.05
Engine Size (l)
Hwy MPG
Inter / Intra: 1.89
City MPG
Hwy MPG
Inter / Intra: 2.60
Engine Size (l)
Weight
Inter / Intra: 2.60
City MPG
Weight
Inter / Intra: 2.60
Hwy MPG
Weight
Inter / Intra: 1.35
Engine Size (l)
Wheel Base
Inter / Intra: 1.31
City MPG
Wheel Base
Inter / Intra: 1.46
Hwy MPG
Wheel Base
Inter / Intra: 2.60
Weight
Wheel Base
Inter / Intra: 2.00
Engine Size (l)
Width
Inter / Intra: 1.66
City MPG
Width
Inter / Intra: 1.91
Hwy MPG
Width
Inter / Intra: 2.60
Weight
Width
Inter / Intra: 1.43
Wheel Base
Width
Car Type
SUV
Sedan
Minivan

(a) Car

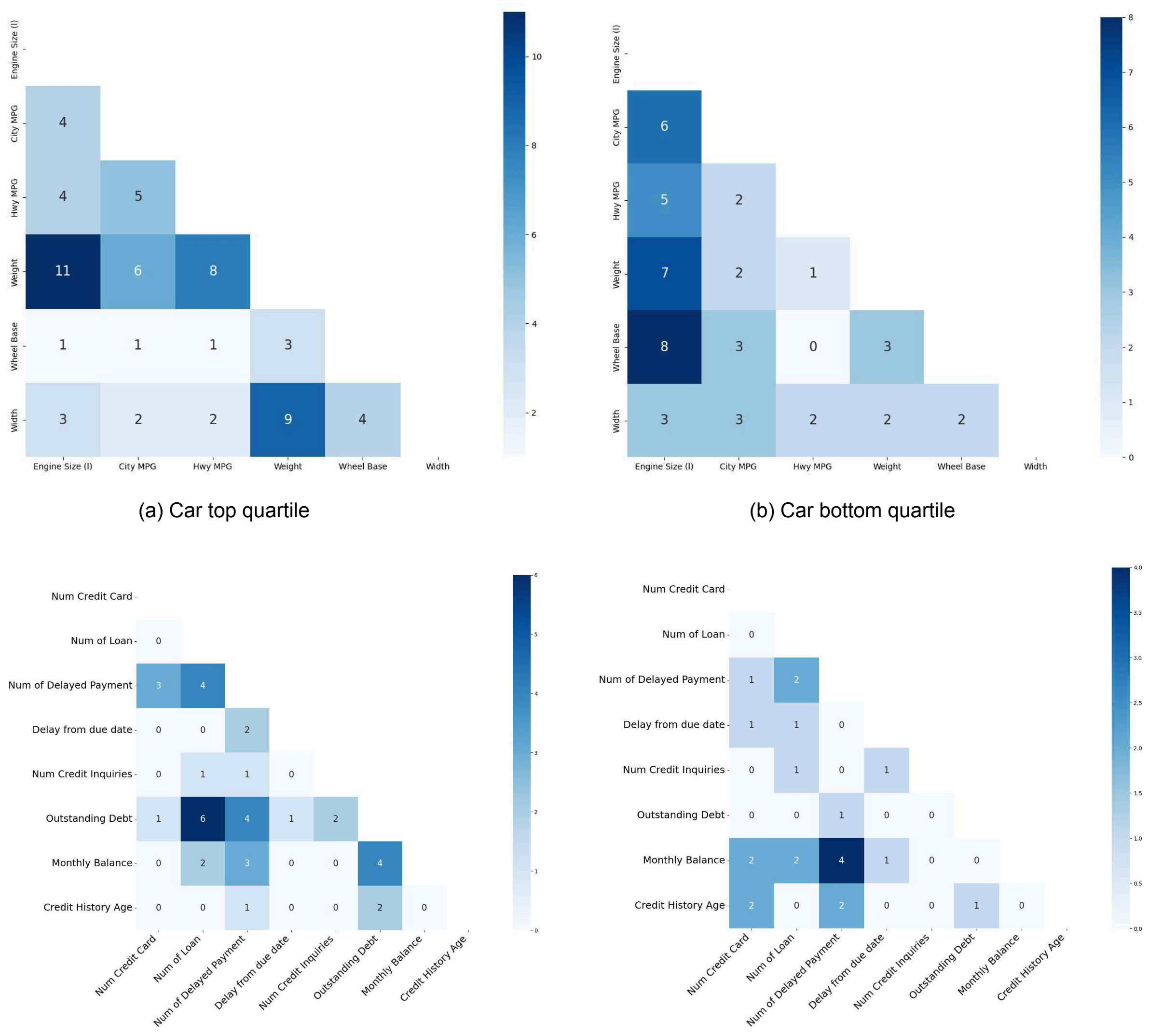


(b) Credit

Fig.7: The distinguishability of each combination of attribute pairs across three categories in both tasks.

Subsequently, we explored the frequency of attribute pair utilization by participants while performing tasks. As depicted in Fig. 8, it emerged that in the car task, the top quartile performers (a) commonly adopted the strategy of using Weight x Engine Size (with a ratio of 2.6) eleven times, whereas the bottom quartile performers (b) most frequently turned to Wheel Base x Engine Size (with a lower ratio of 1.35) eight times. A similar pattern was observed in the credit task: the top quartile performers (c) predominantly used Number of Loans x Outstanding Debt (the pair with the highest ratio of 2.54) six times, while the bottom quartile performers (d) used Monthly Balance x Number of Delayed Payments (which had a ratio of 1.02) four times.

The tendency of the top quartile performers in both tasks to frequently choose the most distinguishable attribute pairs, as opposed to their bottom quartile counterparts, potentially explains the disparity in performance levels between the groups.



(a) Car top quartile

(b) Car bottom quartile

(c) Credit top quartile

(d) Credit bottom quartile

Fig. 8: Frequency of axes changes heat map

Eye Tracking

In our study, we utilized *WebGazer*, a real-time eye-tracking library that harnesses the capabilities of webcams on laptops[2]. This technology enabled us to record the eye gaze locations on screen of participants as they engaged in the categorization tasks in the credit and car domains. However, we chose to consider this approach as exploratory for two reasons. First, compared to professional research-grade eye trackers such as Tobii X2-30 which has an

[2] Papoutsaki et al., "WebGazer: Scalable Webcam Eye Tracking Using User Interactions."

accuracy of 2.46 degrees[3]. Our calibration analysis of *WebGazer* indicated poor accuracy with an average shift of 199.09 pixels in the x-direction and 152.97 pixels in the y-direction (details on calibration analysis are in below). Second, we experienced data loss, with missing eye tracking data for 8 out of 46 participants in the car task and 7 out of 46 participants in the credit task.

Calibration Analysis of Eye Tracker

Eye Tracking Data Loss. We integrated the *WebGazer* library into our interface, requiring participants to maintain a specific position, signified by a green box in a video shown on the left. This alignment is crucial for the proper functioning of the eye-tracking technology. Aware that the video might divert the participants' attention and hinder their concentration on the tasks, we introduced a check-in page. This page guided participants to sustain the designated position throughout the study. Any deviation from this position meant that *WebGazer* could not capture the participant, resulting in a loss of eye-tracking data for a total of 8 participants for the car task and 7 for the credit task.

Calibrating WebGazer Accuracy. Our study included an in-depth analysis of eye-tracking data and the accuracies related to it. We gathered the coordinates of users' clicks, utilizing them as a baseline to deduce the users' gaze direction. Our methodology comprised the following key procedures with the key findings:

1. Eye-tracking Data Processing: For each click event, we isolated the pertinent eye-tracking data that occurred within a 500-millisecond window surrounding the click timestamp.
2. Calculating Averages: We determined the average x and y coordinates of the eye-tracking data inside the designated time frame, leading to the computation of the average shifts in both horizontal and vertical planes. Papoutsaki et al. reported that *Webgazer's* eye tracking data points are at most 72 pixels away within 500ms[4]. Nevertheless, we found that the average shift of 199.09 pixels in the x-direction and 152.97 pixels in the y-direction, which is 2.92 times the reported accuracy in the x-direction and 2.31 times in the y-direction. Fig. 9 demonstrates the direction of shifts for all available eye-tracking data across the two studies.

[3] Clemotte et al., “Accuracy and Precision of the Tobii X2-30 Eye-Tracking under Non Ideal Conditions.”
[4] Papoutsaki et al., “WebGazer: Scalable Webcam Eye Tracking Using User Interactions.”

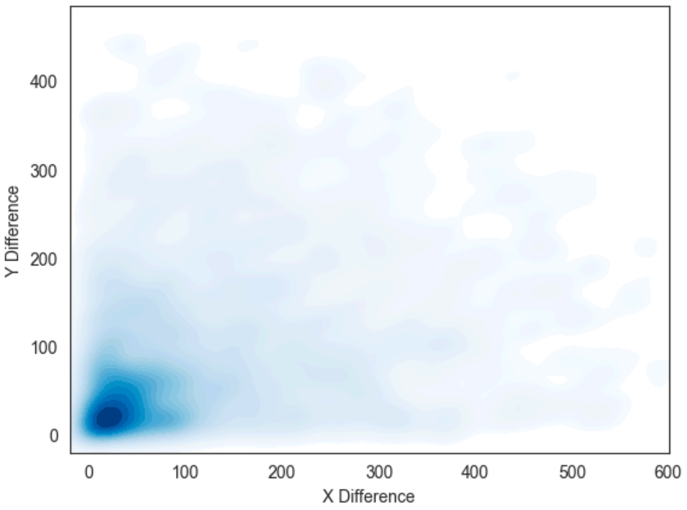
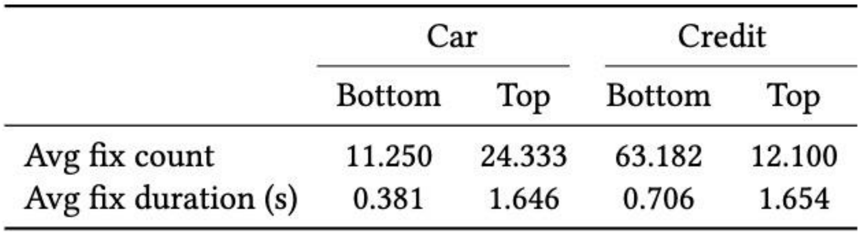


Fig. 9 Absolute Difference between Actual Clicking and WebGazer Prediction Points

Fixation Identification

In this analysis, we sought to understand how individuals' fixations, that aggregated by a group of gaze points[5], relate to their decision-making process. We employed the I-Velocity Threshold (IVT) algorithm[6] to identify fixations from the raw gaze data. We chose to use IVT because it identifies fixations based on the velocity of gaze transitions (compared to alternative methods that consider raw spatial coordinates), and is hence more resilient to the inaccuracy of gaze location. This quality is pivotal, especially when contrasted with dispersion-based algorithms, which can be disproportionately influenced by disparities in screen resolution, thereby skewing the spatial pattern of gaze points[7]. This algorithm thus utilizes a specific velocity threshold to differentiate these eye movement types. We conducted iterative tests varying the threshold between 20 and 80 pixels/second for the velocity threshold to differentiate fixations from saccades. We selected a threshold of 35 pixels/second because it resulted in average fixation durations—across both bottom and top quartiles and in both tasks— that exceeded a benchmark of 330 milliseconds (ms). This benchmark was informed by prior work which suggests an average fixation duration of 330 ms for scene perception tasks[8].



	Car		Credit	
	Bottom	Top	Bottom	Top
Avg fix count	11.250	24.333	63.182	12.100
Avg fix duration (s)	0.381	1.646	0.706	1.654

Table 2. Fixation metrics across two quartiles.

[5] Blascheck et al., "Visualization of Eye Tracking Data."
[6] Erkelens and Vogels, "The Initial Direction and Landing Position of Saccades."
[7] Erkelens and Vogels.
[8] Rayner, "Eye Movements in Reading and Information Processing: 20 Years of Research."

Table 2. presents the average duration per fixation and the mean count of fixations per participant, segmented by top and bottom quartiles across the two tasks. In the car task, the top quartile exhibited a greater number of fixations coupled with a longer average fixation duration. Conversely, in the credit task, those in the bottom quartile exhibited a higher number of fixations, while the top quartile sustained a longer average fixation duration. However, we did not detect significant differences in the two fixation metrics across bottom and top quartiles in the two tasks.

Discussion of Results

In spite of challenges like data loss and inaccuracies of the web-based eye tracker, these findings lead to questions for further exploration on the potential for eye-tracking.

For instance, the increased fixation count and fixation duration in the credit task may correspond to the intrinsic complexity of the task. The variation in fixations between the top and bottom quartiles also illuminates the differences in strategies participants may employ when navigating tasks of differing complexities. However, as the analysis is exploratory, future work is needed to confirm these observations.

References

Blascheck, T., K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl. “Visualization of Eye Tracking Data: A Taxonomy and Survey.” *Computer Graphics Forum* 36, no. 8 (2017): 260–84. https://doi.org/10.1111/cgf.13079.

Clemotte, A., M. Velasco, D. Torricelli, R. Raya, and R. Ceres. “Accuracy and Precision of the Tobii X2-30 Eye-Tracking under Non Ideal Conditions:” In *Proceedings of the 2nd International Congress on Neurotechnology, Electronics and Informatics*, 111–16. Rome, Italy: SCITEPRESS - Science and and Technology Publications, 2014. https://doi.org/10.5220/0005094201110116.

Erkelens, Casper J., and Ingrid M.L.C. Vogels. “The Initial Direction and Landing Position of Saccades.” In *Studies in Visual Information Processing*, 6:133–44. Elsevier, 1995. https://doi.org/10.1016/S0926-907X(05)80012-1.

Feng, Mi, Evan Peck, and Lane Harrison. “Patterns and Pace: Quantifying Diverse Exploration Behavior with Visualizations on the Web.” *IEEE Transactions on Visualization and Computer Graphics* 25, no. 1 (January 2019): 501–11. https://doi.org/10.1109/TVCG.2018.2865117.

Papoutsaki, Alexandra, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. “WebGazer: Scalable Webcam Eye Tracking Using User Interactions,” n.d.

Rayner, Keith. “Eye Movements in Reading and Information Processing: 20 Years of Research.” *EYE MOVEMENTS IN READING*, n.d.