Less Is More Reflective: Intervention Intensity and Metacognitive Disruption in Visual Analytics

Mengyu Chen, Yixin Bai, Zheyuan Lin and Emily Wall

Abstract—Exploratory data visualization tools often prioritize efficiency over reflective thinking, leaving users vulnerable to cognitive biases and superficial analyses. This capacity for reflective thinking, specifically, the ability to monitor one's own analytical process, is known as metacognitive awareness, yet it remains largely unsupported in visualization design. In this paper, we investigate two potential interventions for improving metacognitive awareness: *Socratic prompting*, which triggers reflection through contextual questioning, and visualizing *interaction traces*, which surface behavioral patterns to support self-monitoring. We conducted a between-subjects study (N=36) comparing contextual Socratic prompts delivered via Wizard-of-Oz and a bias-aware visualization interface that surfaces interaction traces against a control condition. We measured metacognitive awareness before and after completing a visualization exploration task using an adapted Metacognitive Awareness Inventory (MAI). Contrary to our expectations, we observed a significant decrease in MAI scores for participants in the interaction trace group, while the Socratic group and control group showed no significant decline. Our findings reveal how certain design features intended to support awareness may instead lead to cognitive overload, distraction, or superficial engagement, ultimately diminishing it. This work contributes empirical support for the growing body of work on promoting reflection with visualizations, highlighting the importance of evaluating interventions not only for their intended effects but also for their unintended consequences. All Supplemental Materials are available at https://osf.io/bnamf.

Index Terms—Metacognition, Metacognitive Awareness, Visual Analytics, Socratic Questioning, Interaction Trace, Reflective Thinking

1 Introduction

Visual analytics tools are designed to accelerate the process of data exploration and insight discovery [36, 57]. By offering interactive visualizations, these tools enable users to rapidly filter and aggregate to examine large datasets, transforming raw numbers into actionable understanding. Prevailing design philosophies prioritize efficiency by reducing the number of clicks, minimizing cognitive load, and streamlining the path from question to answer [27, 62]. However, this emphasis on speed and ease of use may carry an often-overlooked cost: the erosion of reflective thinking. When users interact with a tool that prioritizes efficiency, they may fall into cognitive traps such as confirmation bias [38], prematurely settling on an initial interpretation without exploring alternatives. They may passively accept whatever patterns emerge rather than actively questioning the data's provenance or limitations, heightening the likelihood of spurious conclusions and false discoveries [74].

Central to combating these tendencies is a cognitive capacity known as **metacognitive awareness** [18, 60], the ability to monitor, evaluate, and regulate one's own thought processes during an analytical task. Much of the effort in visual analytics research has focused on improving the tools themselves via designing better encodings [40, 42], smarter interactions [28], and more powerful algorithms [35]. Yet this tool-centric perspective implicitly treats the human as a passive recipient of insight rather than an active reasoning agent whose own cognitive processes are equally worthy of support. In the context of visual analytics, this capacity is consequential: because insights are constructed interactively through a sequence of user-driven choices (e.g., what to filter, what to encode, what to compare), the quality of the final analysis depends not only on the data or the tool, but on whether the user can critically reflect on those choices as they make them. A user with high metacognitive awareness might pause to ask themselves: Am I interpreting this correlation correctly? What assumptions am I making? Have I considered alternative explanations? Without this self-monitoring, even powerful visual analytics tools may risk amplifying rather than correcting analytical blind spots. This capacity for self-reflection is widely recognized as a cornerstone of effective learning [75], problem-solving [19], and reasoning [20]. Yet, despite its importance, metacognitive awareness remains largely unsupported by the design of contemporary visual analytics interfaces.

To address this gap, we consider ways of prompting metacognitive awareness. In this work, we consider simple visual cues known as **interaction traces**, visual representations of a user's analytic behavior presented as real-time feedback within the interface, which have been shown to support self-monitoring in visual analytic systems [43, 70]. Lumos, for example, is a tool that captures and displays interaction history using both *in-situ* techniques (at the place of interaction) and *ex-situ* techniques (in a separate, external view), promoting reflection upon analytic intentions and influencing subsequent interactions [43]. Beyond the visualization community, we can also consider lessons from education. For centuries, the **Socratic method** – a dialectical approach to inquiry in which an instructor guides learners toward deeper understanding through systematic questioning rather than direct instruction – has served as a powerful pedagogical approach to foster critical thinking and self-reflection [13]. Rather than lecturing or providing direct answers, a Socratic instructor poses probing questions, e.g., how do you know? What evidence supports your claim? These questions compel learners to examine the foundations of their own reasoning. This process of guided self-interrogation has been shown to cultivate deeper understanding [61], expose hidden assumptions [52], and promote enduring intellectual habits [55].

Recent work suggests that these frameworks may be promising for visualization systems [8]; however, to our knowledge, no work has yet attempted to measure the metacognitive benefits of a visual analytic intervention. **In this paper, we explore whether these two interventions, Interaction Traces and Socratic Prompting, can promote metacognitive awareness during visual data exploration.** We conducted a preregistered between-subjects study[1] with N=36 participants across three conditions: (1) a Socratic prompting condition (Socratic Group), where participants received contextual, Wizard-of-Oz questions designed to encourage reflection; (2) a bias-aware visualization interface condition (Interaction Trace Group), which incorporated visualized interaction history intended to make users conscious of potential biases; and (3) a control condition (Control Group), where participants used a standard visualization tool without any additional

• Mengyu Chen, Zheyuan Lin, and Emily Wall are with Emory University. E-mail: {mengyu.chen | zheyuan.lin | emily.wall}@emory.edu
• Yixin Bai is with The University of Maryland, College Park. E-mail: yixbai01@umd.edu

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

[1] https://aspredicted.org/688y-m289.pdf

interventions. We measured metacognitive awareness using an adapted version of the Metacognitive Awareness Inventory (MAI) [60], a validated instrument commonly used in educational research. Our findings ran counter to our initial expectations. Participants in the Interaction Trace Group showed a significant decrease in MAI scores, suggesting that the intervention intended to heighten awareness may have inadvertently diminished it. Meanwhile, participants in the Socratic and Control groups showed no significant change in MAI scores.

Through analysis of interaction logs, participant-generated insights, and follow-up interview responses, we found that certain design features intended to support awareness may have instead led to unintended consequences such as cognitive overload, distraction, or superficial engagement, paradoxically undermining the very processes they aimed to cultivate. These results contribute empirical support for the growing body of work on promoting reflection with visualizations and introduce challenging confounds for increasing metacognitive awareness.

2 Related Work

This section reviews literature surrounding metacognitive support during data exploration, covering metacognitive awareness in learning science, awareness support in visualization, and the efficacy of Socratic questioning.

2.1 Metacognitive Awareness

Metacognition refers to the awareness and regulation of one's own cognitive processes [18, 60]. It encompasses two broad dimensions: knowledge about cognition (i.e., understanding one's own thinking patterns and strategies) and regulation of cognition (i.e., the ability to plan, monitor, and evaluate one's approach during a task) [60]. In education and learning science, metacognitive awareness has been consistently identified as a predictor of learning quality and problem-solving effectiveness [64, 71], as learners who actively monitor their reasoning tend to detect errors earlier and achieve stronger outcomes [46].

Chen et al. likewise conceptualized visualization as a learning process, and reviewed work on metacognition in the visualization community [8]. Their recent systematic survey revealed that metacognition is rarely a focal construct in the visualization community. Most relevant work addresses metacognitive concepts implicitly through related constructs such as bias awareness [16, 43, 70], sensemaking [56], or reasoning [7]. One notable exception is work by Schlieder et al., who presented a method for measuring metacognitive understanding in visualization learning with deceptive bar charts, finding that such understanding positively influences learning over time and subsequently proposed a model of metacognitive processing for visualization design [59].

2.2 Awareness in Visualization

In open-ended visual data exploration, analysts must go beyond identifying patterns to assess whether their approach is sufficient, whether they have considered alternative perspectives, and whether their conclusions are supported by evidence [23]. Prior research has sought to support this reflective process through several distinct ways. First, bias detection and mitigation research focuses on employing computational metrics to monitor user behavior against statistical benchmarks [3,45,66–68], flagging potentially biased interaction patterns before they lead to flawed conclusions [43, 70] or utilizing system-level design interventions such as highlighting superior alternatives to counteract decision-making biases like the attraction effect [14, 15]. Additionally, systems often make awareness visible by externalizing users' analytical behavior through embedded visual cues. For example, "scented widgets" enhance standard UI controls (e.g., sliders) with usage-based information scent to guide navigation [72]. Interaction traces extend this idea by recording the sequence of a user's actions, such as the data points they hover, click, or filter, and comparing these patterns against the underlying data distribution to reveal where attention has been unevenly allocated [43, 70]. Lumos, for instance, logs and replays interaction histories through in-situ and ex-situ cues so that analysts can retrospectively inspect which subsets of data they over- or under-explored [43], while Wall et al. quantify deviations between observed and expected interaction coverage to surface latent attentional biases, such as toward particular genders or political leanings [70]. Building on these traces, subsequent work compares user-modeling techniques for predicting interaction and detecting exploration bias [25], and pairs visual feedback with additional modalities such as haptics [49]. Yet these traces are typically framed around bias and coverage. How rendering one's own analytic process inspectable shapes a user's internal metacognitive state remains unexamined, motivating the Interaction Trace condition in our study.

2.3 Socratic Questions

Socratic questioning is a method of guided inquiry with roots in philosophical pedagogy [53]. Rather than providing direct answers or corrective feedback, it prompts individuals to examine their own assumptions, articulate their reasoning, and evaluate the adequacy of their conclusions [48, 50].

Empirical evidence supports the effectiveness of Socratic questioning in clinical and educational settings. In cognitive behavioral therapy, it is considered a cornerstone technique for guided discovery [10, 50], and Vittorio et al. [65] found that it mediates symptom reduction through cognitive change, with stronger effects among individuals who began with lower skill levels. In educational settings, Socratic questioning has been associated with improvements in critical thinking and metacognitive awareness [29, 73].

Despite this established body of work in education and clinical settings, Socratic questioning has received little empirical investigation in the context of visualization. Our study addresses this gap by adapting Socratic prompts to support metacognitive awareness during interactive visualization tasks.

3 Apparatus

This section presents the materials that formed the basis of our experimental manipulation and measurement. We begin by introducing Lumos, the visualization platform that served as the foundation for all three conditions. We then describe the development of our Socratic questions and conclude with the instruments used to assess metacognitive awareness and visualization literacy.

3.1 Interface: Lumos

In this study, we built upon Lumos [43], a visual data analysis tool originally designed to increase awareness of analytic behaviors through interaction traces, as a platform for comparing three experimental conditions. We chose Lumos because its interaction trace features directly operationalize the bias-aware feedback we sought to investigate; additionally, its base interface served as a neutral platform for the Socratic and Control conditions (detailed in Section 4.1), ensuring that any observed differences could be attributed to the interventions rather than interface variability. We introduce the original design of Lumos here and specify how the interface features differed across conditions in Section 4.1.

The Lumos system employs two complementary visualization techniques to provide real-time feedback about user analytic behavior: (1) in-situ interaction traces that appear directly at the place of interaction, and (2) ex-situ interaction traces presented in external views for comparative analysis. All conditions share identical core visualization functionality including interactive scatterplot manipulation, categorical and numerical filtering capabilities, encoding panel controls, and detail-on-demand information display, while varying exclusively in their awareness feature configurations:

Core Functionality: The interface provides standard visual data analysis components including a Data Panel (Fig. 1, A) showing the loaded dataset, an Attribute Panel B displaying attribute lists with data types, an Encoding Panel C with dropdown controls for chart creation (supporting scatter plot, strip plot, bar chart, line chart), a Filter Panel D with range sliders and multiselect dropdowns, a Visualization Canvas F for rendering charts, and a Details View G showing additional information upon interaction with visualization elements. This system served as the interface for the Control group.

In-Situ Interaction Traces: The system tracks user interactions with visual representations and provides immediate feedback through color encoding J. Visualization elements (points, bars, strips) in the original interface are colored on a white → blue scale based on relative frequency of interactions, with dark blue representing high interaction frequency and white indicating no interactions. Similarly, attributes in the Attribute Panel receive color coding based on usage frequency, and the Details View updates table row background colors when users interact with aggregate visualization elements.

Ex-Situ Interaction Traces: Ex-situ traces are aggregated in a separate Distribution Panel I, which presents comparative visualizations of user interaction patterns against underlying data distributions through attributes colored on a red → green scale. Red backgrounds indicate significant deviation from target distributions while green backgrounds indicate similarity. Each attribute card can be expanded to reveal detailed visualizations that overlay user behavior (blue area) on target distributions (black line), enabling users to assess whether their analytic focus aligns with their intended analysis goals. The core system plus both in-situ and ex-situ interaction traces served as the interface for the INTERACTION TRACE condition.

3.2 Dataset

We generated a dataset of 1,000 fictitious U.S. voters that simulates real-world electoral dynamics with layered statistical fidelity. Each data item represents one voter, characterized by demographics (age, gender, race, income, location), political attributes (party affiliation and policy positions on abortion, gun control, and immigration scaled from -3 [strongly oppose] to +3 [strongly support]) [70], and behavioral traits (turnout history). We targeted 58% Democratic urban voters and 63% Republican rural voters (per 2022 exit polls [11]); however, the realized distributions were 52.6% and 59.1% respectively after enforcing higher-priority demographic constraints, e.g., 86.7% Democratic affiliation among Black voters (target: 86%). Policy positions strictly adhere to 2022 exit polls data: e.g., 90% of Republicans oppose legal abortion (avg. -2.2), while 76% of Democrats support stricter gun control (avg. +2.1). Neutral positions are capped at 10% with ±0.5 noise on political views to prevent oversimplification, representing the growing political polarization in the U.S., where fewer individuals maintain neutral or counter-partisan stances [6]. Names are sampled from the U.S. Census[2] distributions to ensure realism.

3.3 Socratic Question Development

Our Socratic questions are grounded in the taxonomy of Socratic questioning proposed by Paul and Elder [53], which categorizes questions into six dimensions of critical inquiry: *clarity*, *precision*, *accuracy*, *relevance*, *depth*, and *breadth*. These question types systematically probe reasoning by encouraging deeper reflection, from clarifying meaning to examining assumptions and alternative perspectives. To tailor this framework to visualization, we adapted generic prompts into domain-specific questions. For example, the original clarity question ***"Can you provide more elaboration on the topic?"*** was refined to ***"Can you provide more elaboration on your choice of visualization?"*** to better guide participants in reflecting on their design decisions. The full set of questions (and their behavioral triggers) is shown in Table 1 and their development is detailed in Section 4.6.

3.4 Measuring Metacognitive Awareness

We selected the Metacognitive Awareness Inventory (MAI) [60] as our primary effectiveness measure. The MAI is a well-validated instrument in learning science comprised of 52 self-report items that assess two core dimensions: knowledge about cognition and regulation of cognition. For this study, we used a shortened 18-item version [63] to reduce participant burden and minimize fatigue during the study session. This version has been validated as a reliable instrument for capturing the essential constructs of metacognitive awareness among student teachers [33, 34]. The 18 items were evenly split, with nine addressing knowledge about cognition (e.g., 'I am a good judge of how well I understand something') and nine addressing regulation of cognition (e.g., 'I set specific goals before I begin a task'). Responses use a 5-point Likert scale, numerically scored from 1 ("I never do this") to 5 ("I always do this").

We further adapted the MAI to better fit the visualization context by systematically modifying terminology related to data analysis, e.g., **MAI-1**: "goals" → "analysis goals" (*I ask myself periodically if I am meeting my* ***analysis*** *goals*); **MAI-15**: "learn" → "analyze data" (*I* ~~*learn*~~ ***analyze data*** *best when I know something about the topic*). The full adapted survey instrument can be found in Supplemental Materials.

3.5 Measuring Visualization Literacy

Capturing visualization literacy enables us to ensure a baseline competency for participants in our study. We used a subset of the Mini-VLAT [51], a validated 12-item short form of the Visualization Literacy Assessment Test (VLAT) [37]. Since our study focuses on three core visualization types supported by our tool: line charts, bar charts, and scatterplots, we selected one representative question per type (3 questions total), maintaining the original 25-second time limit per question to balance assessment rigor with experimental efficiency. To minimize guessing behavior, we modified the response options by adding an explicit "Omit" choice alongside the original answer alternatives [37].

4 STUDY DESIGN

4.1 Experimental Conditions

To evaluate the effectiveness of our interventions for enhancing metacognitive awareness in the context of visualization, we employ a between-subjects design (Fig. 2) with three conditions: (1) CONTROL group, (2) INTERACTION TRACE group, and (3) SOCRATIC group. All participants completed the same visual data exploration task: analyzing a dataset to generate insights using their assigned version of the visualization tool. All conditions share identical core functionality of the visualization interface including the same dataset, visualization capabilities, and an insight input box (Fig. 1, E) for participants to submit their findings:

CONTROL GROUP: Participants utilized the core version of the Lumos visualization system with *all interaction trace features disabled*. This baseline condition provides only core visualization functionalities described in Section 3.1. The control group establishes a baseline against which to compare metacognitive changes in the INTERACTION TRACE and SOCRATIC groups.

INTERACTION TRACE GROUP: Participants accessed the original Lumos interface with full interaction trace features enabled. This condition incorporates real-time interaction feedback through in-situ (Fig. 1, J) and ex-situ (Fig. 1, I) visualizations (Section 3.1). This condition tests whether existing bias-mitigation approaches adequately support metacognitive engagement.

SOCRATIC GROUP: Participants in this condition used the basic Lumos version (identical to the Control condition) enhanced with contextual Socratic questioning delivered through a pop-up chat interface in the bottom left of the interface (Fig. 1, H). To validate preliminary findings about Socratic prompting, we implemented a Wizard-of-Oz protocol [12, 69] where a researcher ("wizard") monitored participant interactions via screen sharing and delivered real-time Socratic prompts through the chat system based on established behavioral triggers. Example behavioral triggers include when participants appeared stuck, made a significant insight, or exhibited surface-level engagement with the data. This condition isolates the effect of metacognitive questioning by maintaining the same basic visualization interface while adding only the conversational prompting mechanism. To reduce variability in wizard behavior, a single researcher served as the wizard for all study sessions, with prompt triggering determined through the pilot study described in Section 4.6.

4.2 Hypotheses

Based on our experimental design, we formulate the following three hypotheses regarding participant outcomes:

[2] https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

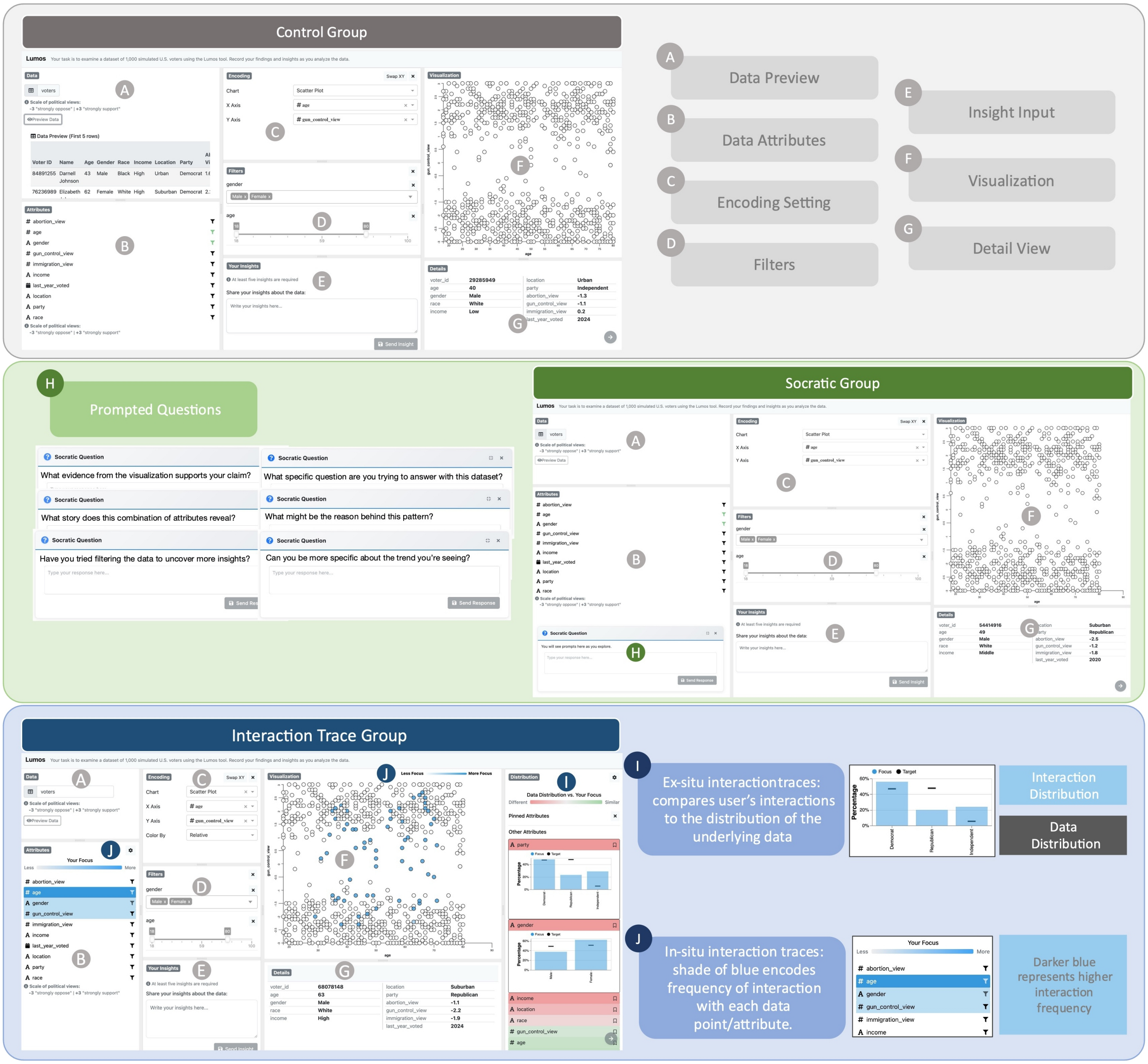

Figure 1: The three interface conditions varied in features. The control group (top) used a basic version of Lumos with all bias-awareness features disabled. The Socratic group (middle) used an interface identical to the control condition but enhanced with Socratic questioning. The Interaction Trace group (bottom) accessed the original Lumos tool with in-situ and ex-situ interaction trace features.

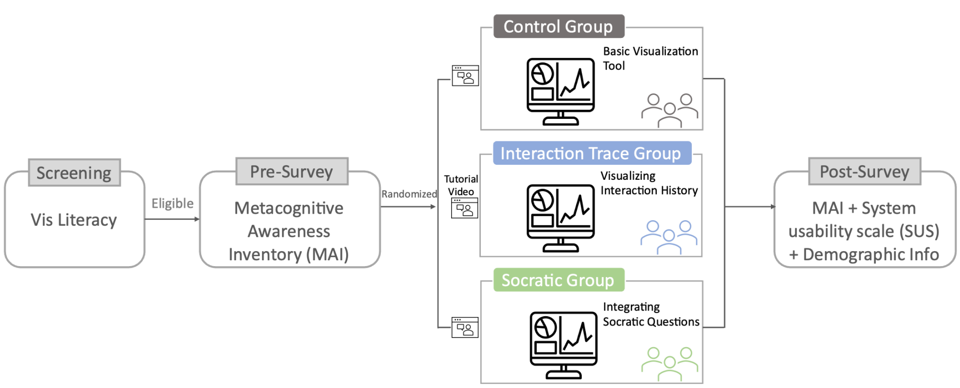

Figure 2: Study procedure: Participants completed a visualization literacy assessment first. Eligible participants were then randomized into one of three groups for an exploratory visual analytic task and completed pre- and post-surveys.

- **H1**: Participants in the Socratic group and the Interaction Trace group will show a greater MAI improvement compared to those in the Control group.
- **H2**: Participants in the Socratic group and the Interaction Trace group will generate, on average, higher-quality insights than those in the Control group.
- **H3**: Participants in the Socratic group and the Interaction Trace group will have different interaction patterns, compared to those in the Control group, specifically in terms of interaction sequences of analytic behaviors.

4.3 Task

Participants across all conditions completed the same task – analyzing a standardized dataset (Section 3.2) to generate insights using their assigned version of the visualization tool. Particularly, participants were required to submit at least five insights via the insight input box in the interface (Fig. 1, E) before proceeding to the post-survey. This threshold was set to ensure sufficient participant engagement with the dataset and to provide an adequate volume of insights for meaningful analysis across participants. Following prior work in visual analytics advocating for open-ended insight elicitation to avoid constraining analytic behavior [44], we intentionally did not define 'insight' in the task instructions to avoid anchoring participants' exploration.

4.4 Participants

We ultimately recruited $N = 36$ participants (12 per condition), determined by: (1) common practice in HCI for detecting medium effect sizes in controlled comparisons [5], and (2) feasibility constraints of conducting 1-on-1 sessions. Participants were recruited via Prolific and received $12/hour. They ranged in age from 19-55 years old ($M = 35.75$, $SD = 9.09$). Fifteen participants identified as female and 21 identified as male. The majority of participants (27) hold a college degree (Bachelor's, Master's, or Doctorate), while 5 have some college or Associate's degree, and 4 have a High School diploma. Their professional backgrounds were diverse, spanning fields such as computer science, psychology, construction, sales, and social work, among others.

4.5 Procedure

Our experiment was approved by our university's IRB (No. 00007476). Informed consent was obtained from all participants at the start of the session. All participants across the three groups followed the same procedure (Fig. 2). The study began with a visualization literacy assessment consisting of 3 timed questions (25 seconds per question) adapted from the Mini-VLAT [51]. Each correct answer earned 1 point (no penalty for incorrect responses), and participants scoring ≤ 1 were excluded from further participation. Eligible participants then completed the Metacognitive Awareness Inventory (MAI) based on their *general experience* to establish baseline metacognitive competence, along with two attention check questions. Participants who failed both attention checks (indicating inattentive responses [54]) were disqualified from further participation. Following MAI completion, participants were randomly assigned to one of the three groups to perform an exploratory visual analytic task while documenting their insights. Upon task completion, they filled out a post-survey featuring: (1) a repeating MAI targeting *task-specific* metacognitive awareness with two new attention check questions (failure on both excluded participants from analysis), (2) the 10-item System Usability Scale (SUS) [4] to quantify tool usability (scored 0–100 using the standard weighting method), and (3) demographic questions covering age, gender, education level, and occupation.

4.6 Pilot Study: Establishing a Wizard-of-Oz Protocol

Prior to the full-scale study, we conducted two iterative pilot rounds (a total of N=8) to optimize study parameters. The pilots simultaneously developed and evaluated a Socratic prompting protocol (Table 1) to provide reproducible triggering logic specifically for the Socratic group. The protocol operationalizes when and what Socratic prompts to present based on user interactions.

During initial pilot studies, we employed a think-aloud protocol where participants verbalized their thought processes while analyzing the dataset without predefined prompts, to capture as many natural prompt triggers as possible. A primary behavior we observed was "aimless exploration," where participants repeatedly cycled through different data attributes without making clear analytical progress. Our think-aloud protocol revealed two distinct reasons for this single behavior. Some participants admitted they lacked a clear analytical goal, which inspired a **Clarity Prompt** to help them establish one: "*What specific question are you trying to answer with this dataset?*" However, other participants exhibiting the same behavior had a clear goal but were stuck looking for the right data attribute or visualization to address it. This led to a **Breadth Prompt** focused on alternative approaches: "*Have you considered alternative data attributes or visualizations that can help you solve the problem?*" Consequently, we designed a two-stage protocol. When "aimless exploration" is detected, we first ask the Clarity Prompt. If the user confirms they have a goal, we then follow up with the Breadth Prompt to help them find a new approach. For full pilot study details, see Supplemental Materials.

5 Findings

Participants took an average of about 40 minutes (SD = 10.23) to complete the full task, which included two surveys, with the visualization task itself averaging 23.24 minutes (SD = 5.75). On the visualization literacy measure, eligible participants scored an average of 2.56 out of 3 (SD = 0.50). Next, we describe the three primary measures used in our analysis: metacognitive awareness, insight quality, and interactive behaviors.

5.1 MAI Score

For each participant, we calculated a total metacognitive awareness score by summing the scores from all 18 items. Each item was rated on a 5-point Likert scale from 1 ("I never do this") to 5 ("I always do this"), giving possible total scores ranging from 18 to 90. Consequently, a higher total score indicates a greater level of self-reported metacognitive awareness. Participants completed the same questionnaire before (in reference to their general analytic habits) and after the intervention (in reference to the specific task), and the change in pre- and post-task served as our primary measure of metacognitive effectiveness across conditions. Across all conditions, the average general MAI score was $M = 68.86$ ($SD = 9.87$) pre-task and $M = 66.94$ ($SD = 9.08$) post-task.

The results revealed distinct patterns between conditions, as shown in Fig. 3. As expected, the Control condition exhibited no change in metacognitive awareness scores from pre- to post-task ($M = 65.5$, $SD = 7.8$ pre; $M = 65.2$, $SD = 7.5$ post; $p = .911$), with a negligible effect size ($d = -0.03$). In contrast, the Interaction Trace condition showed a significant decline from pre- ($M = 73.2$, $SD = 7.6$) to post-intervention ($M = 68.4$, $SD = 6.4$; $p = .041$), representing a medium effect size ($d = -0.67$). The Socratic condition also declined, though non-significantly, from 70.2 ($SD = 10.7$) to 67.2 ($SD = 11.7$; $p = .237$), with a small effect size ($d = -0.36$). Contrary to our hypothesis, participants in the intervention conditions did not show a greater improvement in metacognitive awareness compared to the Control condition. Thus, we find **no support for H1**. We explore possible explanations for these unexpected findings and their implications in Section 6.

5.2 Insights

We analyzed insights generated by participants during the task, allowing us to compare participants' self-reported changes in metacognitive awareness with a more objective measure of what they actually produced. In this analysis, we defined an insight as "an individual observation about the data by a participant, a unit of discovery" [1,58]. Because the task instructions asked participants to submit at least five insights, we treated each submission as one insight unit. In total, we collected 236 insights from the study. After removing three insights that were duplicates due to copy-paste errors, we retained 233 insights ($M = 6.47$ per participant) for subsequent analysis.

Table 1: Behavior-Triggered Socratic Prompting Protocol

Behavioral Trigger	Question Type	Objective	Example Prompt
Vague insights	Precision	These questions prompt participants to be specific and detailed in their analysis.	"Can you be more specific about the trend you're seeing?"
Exploration impasses	Breadth	These questions prompt participants to consider alternative perspectives, contexts, or broader implications of their findings.	"Have you considered alternative data attributes/visualizations that can help you to solve the problem?"
Mechanical attribute combinations	Relevance	These questions help participants focus on the most pertinent aspects of the dataset and avoid tangential or irrelevant information.	"What story does this combination of attributes reveal?"
Aimless exploration	Clarity	These questions help participants ensure their understanding of the dataset, visualizations, and insights is clear and well-defined.	"What specific question are you trying to answer with this dataset?"
Shallow insights	Depth	These questions encourage participants to explore the underlying complexities and implications of the dataset and their insights.	"What might be the reason behind this pattern?"
Factual error in insights	Accuracy	These questions encourage participants to verify the correctness of their interpretations and conclusions.	"Is there any piece of evidence that might contradict or complicate that conclusion?"
Underutilization of tool features	Breadth	These questions prompt participants to consider alternative perspectives, contexts, or broader implications of their findings.	"Have you tried filtering the data to uncover more insights?" or "You haven't used a [bar chart] yet; have you considered other visualization types?"

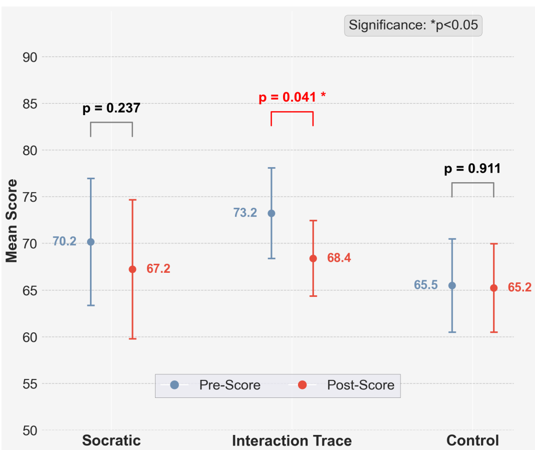

Figure 3: Mean pre- vs post-MAI scores Distribution

Previous work has proposed various ways to categorize insights [1, 9, 22, 39]. For example, Guo et al. [24] classified insights into facts, hypotheses, and generalizations, which we initially preregistered. However, when applied, we found that most insights fell into the generalization category, making the schema uninformative for our analysis. Instead, we adopted the framework proposed by Saraiya et al. [58]. This framework goes beyond simply categorizing insights into fixed types – it analyzes insights from multiple dimensional perspectives. Specifically, Saraiya et al. characterized insights along four dimensions: **directness vs. unexpectedness, correctness, breadth vs. depth, and domain value**.

Two authors of this paper first collaboratively developed detailed grading criteria, adapted from He et al.'s work [26], and then graded each insight on the four dimensions using a 1-to-5 scale. The full grading rubric can be found in the Supplemental Materials. For *directness vs. unexpectedness*, *breadth vs. depth*, and *domain value*, both authors graded independently to obtain an unbiased measure of inter-coder reliability. For *correctness*, the two authors split the insights and graded them independently to improve efficiency, as this dimension relied on objective verification through reproducing visualizations in the tool or reviewing screen recordings rather than subjective interpretation.

Directness vs. unexpectedness represents a spectrum where a score of 1 indicates a direct and expected insight that confirms well-known patterns, while a score of 5 indicates an unexpected insight that reveals non-obvious patterns. **Correctness** reflects whether the insight can be verified against the actual data in the visualization, with scores ranging from 1 (contradicts the data) to 5 (accurately reflects numerical patterns). **Breadth vs. depth** captures the scope of the insight: broad insights (lower scores) refer to high-level categorical patterns across attributes, whereas deep insights (higher scores) involve nuanced relationships such as interactions between specific subgroups. **Domain value** captures the complexity and significance of an insight. Low-value insights include simple observations like basic counts or single-variable patterns. High-value insights involve pattern recognition across multiple variables, connections to political context, or generation of testable hypotheses. Detailed examples for each dimension are provided in the Supplemental Materials.

Following the approach of He et al., we assessed inter-rater reliability by calculating Kendall's tau-b correlation coefficients for the three dimensions that were independently graded by both coders. All correlations were statistically significant ($p < 0.01$). The τ_b values were 0.49 for directness vs. unexpectedness, 0.70 for breadth vs. depth, and 0.65 for domain value, indicating substantial agreement between two coders for the three dimensions. When the two coders assigned scores that differed by more than one point, they met to discuss the discrepancy and decided whether to revise the score or keep the original ratings. The final score for each of these three dimensions was obtained by averaging the two coders' ratings. For correctness, because the two coders divided the insights between them to grade separately, we did not compute inter-rater reliability for this dimension. Instead, each coder's assigned scores were used directly. Fig. 4 presents the distribution of final scores across all four dimensions by condition.

The Interaction Trace condition generated the highest average number of insights per user ($M = 7.1, SD = 3.9$), followed by the Control condition ($M = 6.3, SD = 2.4$) then Socratic ($M = 6.0, SD = 1.7$). Mann-Whitney U tests [41] with Bonferroni correction [17] for multiple comparisons (adjusted $\alpha = 0.0042$) revealed one robust significant finding: the Socratic condition produced significantly higher Breadth vs. Depth scores ($M = 2.38, SD = 0.56$) compared to the Interaction Trace condition ($M = 2.01, SD = 0.40$) ($U = 3789.50, p = 0.0037$). This indicates that insights generated through the Socratic condition achieved greater depth, specificity, and nuanced analysis (e.g., identifying interactions between demographics), whereas the Interaction Trace condition produced more broad, categorical insights (e.g., simple group comparisons). The difference between Control and Socratic conditions approached significance ($p = 0.0405$) but did not remain significant after correction for multiple comparisons. No significant differences were found for Domain Value, Directness vs. Unexpectedness, or Correctness.

In addition to the insight-level analysis, we also examined the distribution of insight scores by computing the average insight score per participant per condition. No significant differences between conditions

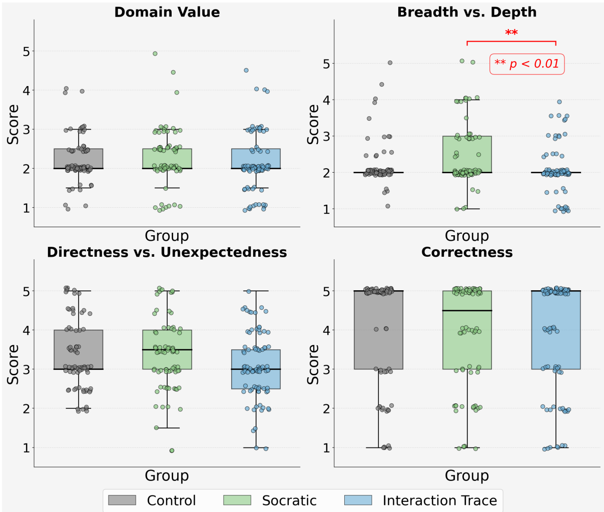

Figure 4: Insight Quality Score Distribution by Condition

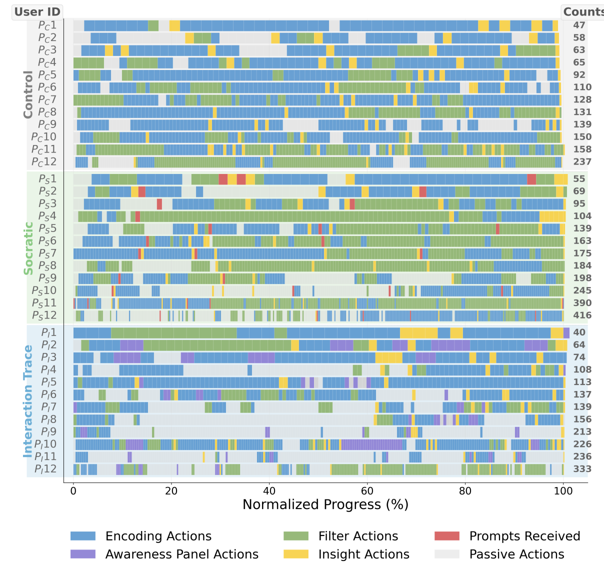

Figure 5: Interaction Sequence. Normalized behavioral timelines (0-100% session progress) with action categories color-coded. Prompts received appear only in the SOCRATIC group; awareness panel actions appear only in the INTERACTION TRACE group.

were observed across the four dimensions. Full results can be found in the Supplemental Materials. Thus, we find **mixed support for H2**.

5.3 Interactions

To characterize exploratory behaviors across conditions, we examined the sequence of participants' interactions.

5.3.1 Descriptive Statistics

We classified all logged events into three types: *configure actions* (e.g., encoding changes, filter operations, and insight submissions), *passive actions* (e.g., hovering, clicking data points, and toggling data previews), and *condition-specific actions* (e.g., awareness panel toggles for INTERACTION TRACE; prompt responses for SOCRATIC). All three groups saved a similar number of insights (range: $6.0-7.1$). The INTERACTION TRACE group spent the most time on task ($M = 26.7, SD = 12.2$ min), followed by the SOCRATIC group ($M = 23.9, SD = 9.2$ min) and the CONTROL group ($M = 19.1, SD = 8.8$ min). In terms of total actions, the SOCRATIC group was highest ($M = 186.1, SD = 115.6$), followed by INTERACTION TRACE ($M = 153.3, SD = 84.9$) and CONTROL ($M = 114.8, SD = 54.4$). The SOCRATIC group performed the most configure actions ($M = 144.4, SD = 90.5$), while the INTERACTION TRACE group performed the fewest ($M = 78.8, SD = 50.9$), suggesting a larger proportion of their activity consisted of passive and condition-specific actions. Full action breakdown provided in Supplemental Materials.

5.3.2 Interaction Sequences

To provide a complementary view of how participants engaged with the system across conditions, for each participant, we visualized their chronologically-ordered interaction logs.

Action Categorization. All interaction events were grouped into six high-level categories: *Encoding Actions* (e.g., axis changes, chart type changes, aggregations), *Awareness Panel Actions* (Interaction Trace group only), *Filter Actions* (adding, changing, or removing filters), *Insight Actions* (saving or editing insights), *Prompts Received* (answering prompted questions only accessible to the SOCRATIC group), and *Passive Actions* (data points hovers/clicks, data preview toggles). These categories were assigned distinct color families to support visual pattern recognition.

Normalization. To enable meaningful comparison across users with different numbers of total interactions, we normalize each user's interaction sequence to a standardized 0-100% session progress scale. The first interaction in a user's session is mapped to 0% (session start), the final interaction to 100% (session end), and all intermediate interactions are linearly scaled according to their position within the sequence. This normalization ensures that all users share the same horizontal space regardless of whether they performed 50 or 500 interactions, allowing direct visual comparison of behavioral patterns relative to session progress.

Observations. The interaction sequences revealed different behavioral patterns across the three conditions. Most notably, participants in the INTERACTION TRACE condition exhibited longer periods of passive actions (hover and click events, shown in gray) compared to the other groups. These extended passive sequences suggest that participants in this condition spent more time simply observing the visualization by clicking/hovering over data points or previewing the dataset, without actively manipulating encodings or applying filters. This pattern is consistent with interview reports (see Section 6) in which participants described feeling distracted and spending time on chasing feedback provided by the awareness panel.

A key structural difference between conditions is that the awareness panel (shown in purple) was only accessible to participants in the INTERACTION TRACE condition, as this intervention was designed to provide real-time visual feedback on interaction history. Accordingly, purple squares appear exclusively in the INTERACTION TRACE rows, serving as a visual marker of when participants engaged with this feature. The pattern of frequent purple squares in some participants' sequences (e.g., row of P_I10) suggests active toggling of the awareness panel, while extended gray sequences in others (e.g., row of P_I9) indicate longer periods of passive review. Interestingly, these two behaviors were not typically interwoven within the same participant's sequence; rather, individuals tended to adopt one dominant pattern or the other. In some cases, participants engaged in extended passive review immediately after using the awareness panel (e.g., row of P_I11), suggesting they may have consulted the feedback and then stepped back to reflect on what they had seen before proceeding.

Similarly, the SOCRATIC condition was the only group that received periodic prompts (shown in red), which were delivered by a Wizard-of-Oz protocol to stimulate reflective thinking. These prompt events appear exclusively in the Socratic group rows ($M = 2.91$ prompts per participant, $SD = 1.3$ prompts) and are often followed by increased exploratory activity such as filter actions (green; e.g., row of P_S4) and encoding changes (blue; e.g., row of P_S2), likely suggesting that the prompts encouraged broader exploration.

In contrast, the CONTROL group, which received no intervention, showed more balanced use of encoding (blue), filter (green), and insight actions (yellow), with shorter and less frequent passive sequences (e.g., row of P_C7). The reduced prevalence of extended gray sequences, in both duration and frequency, may suggest that CONTROL participants spent less time in passive observation and instead maintained a more consistent rhythm of active interaction. Some participants' shorter passive pauses appear interspersed naturally between periods of active work (e.g., row of P_C4), rather than accumulating into lengthy review phases. This pattern of balanced and varied activity, coupled with brief but regular pauses, may suggest more self-directed and purposeful interaction, consistent with its role as a baseline for healthy metacognitive functioning. Thus, we find **observational support for H3**.

6 FOLLOW-UP INTERVIEWS

To complement our quantitative findings and better understand underlying cognitive and experiential factors affecting MAI scores, we conducted follow-up interviews with a subset of participants. We aimed to invite all participants to take part in the follow-up interview, but ultimately three from each intervention condition and two from the CONTROL condition agreed to participate (a total of N=8). Follow-up interviews were conducted three weeks after the formal study session. Each interview lasted approximately 30 minutes and participants were compensated at $12/hour. Of the 8 interviewees, 3 were female and 5 were male, with ages ranging from 24 to 48 ($M = 31.88, SD = 7.70$). Educational backgrounds included one participant with a high school diploma, five with bachelor's degrees, and two with master's degrees. Participants came from diverse professional backgrounds including engineering, political science, IT, business, data analysis, software architecture, and information security.

6.1 Method

Cross-Condition Questions. All participants were asked to reflect on their general tendencies versus task-specific performance. Specifically, they were asked to reflect on what experiences came to mind when completing the pre-task questionnaire (e.g., reading charts at work, creating graphs in school, using spreadsheets) and whether they noticed any discrepancies between their self-rated reflective thinking and their actual performance during the task. We also probed participants' prior experience with creating visualizations (none, a little, moderate, extensive) and whether they drew on strategies that had worked for them in the past. Finally, participants were asked about perceived task difficulty, and those who found the task difficult were asked to reflect on where they believed that difficulty primarily originated.

Condition-specific Questions. Participants in the CONTROL group were asked whether there were moments when they wished the tool had provided more feedback or guidance, and if so, what kind of support they would have found helpful. For the INTERACTION TRACE group, participants were asked about their experience with the awareness panel. Participants in the SOCRATIC group were asked how the periodic prompts affected their experience, whether the questions felt helpful, intrusive, neutral, or something else and why. For both intervention groups, we queried participants about their observed change in MAI scores, asking what they believed contributed to the increase or decrease, and invited feedback on features to keep or improve with the tool.

6.2 Observations

All interviews were transcribed and analyzed using thematic analysis to identify recurring patterns within conditions.

6.2.1 Control Group: The Stability of Self-Directed Exploration.

The CONTROL group's stable MAI scores establish an important baseline: the visual data exploration task itself, when performed without intervention, neither enhances nor diminishes metacognitive awareness. Both CONTROL participants we interviewed drew extensively on their professional backgrounds with data, and entered the task with clear goals. $P1_{Control}$ ($\Delta_{MAI} = +4$) had a hypothesis to test ("I have a target in mind"), while $P2_{Control}$ ($\Delta_{MAI} = -1$) sought "meaningful data." They both gravitated toward familiar chart types (e.g., bar charts, line charts), with $P2_{Control}$ explicitly stating he "tried not to deviate from what [he] knew and what worked well for [him] in the past."

When they encountered uncertainty, both adapted through trial and error without losing direction. $P1_{Control}$ described "going back and forth between attributes," while $P2_{Control}$ noted "just trying stuff... and when it wasn't working, [he]'d try something different." This pattern of goal-directed, self-regulated exploration is consistent with their stable MAI scores.

6.2.2 Socratic Group: The Double-Edged Sword of Prompts.

The SOCRATIC group's non-significant MAI decline reflects a tradeoff: the prompts successfully stimulated broader exploration but at a cost to efficiency and metacognitive confidence for some users.

All three participants reported that the prompts stimulated new thinking and broadened their exploration. $P2_{Socratic}$ ($\Delta_{MAI} = +1$) explicitly stated the prompts made him "think broader," and supported reflection: "The questions were also awesome... I also re-summarized everything I just thought about... and it made everything smoother." $P3_{Socratic}$ ($\Delta_{MAI} = -18$) noted they "gave me ideas on new things to think about that I didn't think about before." $P1_{Socratic}$ ($\Delta_{MAI} = +6$) found specific prompts "really helpful" and described "I was checking on the questions" throughout the task.

However, the prompts also introduced costs and elicited mixed emotional responses. $P2_{Socratic}$ described them as "intrusively neutral" and admitted they "made me a little panicky" by triggering self-doubt. $P3_{Socratic}$ noted they caused "momentary distraction" when they appeared during analysis, though the disruption was brief. In contrast, $P1_{Socratic}$ reported no disruption, suggesting individual differences perhaps in cognitive style or tolerance for interruption, may moderate the intervention's effectiveness.

The non-significant decline in the SOCRATIC group's collective MAI scores is thus interpreted as a tradeoff. The prompts broadened their thinking and sustained high activity, yet introduced cognitive costs, e.g., anxiety, distraction, and inefficiency, that likely contributed to the non-significant decline in metacognitive awareness.

6.2.3 Interaction Traces: Goal Displacement.

The most prominent theme across all three INTERACTION TRACE participants was a shift in focus from their own analytical goals to satisfying the tool's feedback. One participant described this displacement directly: "to get the red to go to green, rather than focusing on what I was trying to get out of the information… I had a clear question in my head… and then it wasn't the question I was looking to answer" ($P2_{Trace}$, $\Delta_{MAI} = -7$), noting that the feedback felt like a correctness signal that eroded their confidence: "I went from going into it feeling quite confident… to saying, well, actually, this isn't in my skill set." $P1_{Trace}$ ($\Delta_{MAI} = -1$) experienced a milder version, noting the tool "tries to prompt you in the direction that it wants you to go."

All three participants also described the awareness panel as adding mental burden. $P3_{Trace}$ ($\Delta_{MAI} = +2$) found features "very distracting," and $P2_{Trace}$ distinguished between visual and mental distraction: "not visually... but mentally it was," describing it as "another step that I wasn't sure was necessary."

In sum, the INTERACTION TRACE condition shifts focus from internal reflection to external validation, creating goal displacement and mental burden, and thus a significant decline in metacognitive awareness.

7 DISCUSSION

7.1 When Feedback Undermines Reflection

Our findings reveal a potential tension in the design of metacognitive supports: feedback intended to support reflection can paradoxically undermine it. We observed *goal displacement* as a key mechanism through which persistent feedback may disrupt metacognition. In the INTERACTION TRACE condition, users abandoned their own analytical questions to chase the tool's feedback, such as prioritizing the red/green indicator showing the data vs. interaction distribution (Fig. 1, 1). As one participant noted: "It's red, so I must be doing something wrong." We conducted an exploratory MAI subscale analysis (knowledge vs. regulation) to further examine this: the INTERACTION TRACE group's decline was driven primarily by the knowledge subscale ($t = -3.176, p = .009$) rather than regulation ($t = -1.172, p = .266$), suggesting that the feedback potentially overwhelmed rather than supported users' metacognitive knowledge – they lost their sense of understanding of their own thinking, rather than their ability to regulate it. This finding aligns with Jivet et al.'s observation that dashboards can produce unintended consequences when learners lack the metacognitive skills to interpret the feedback they receive [32].

7.2 When Persistence Paralyzes Focus

Beyond the nature of feedback, our findings likewise highlight the importance of intervention *persistence* as a design dimension. Unlike the episodic Socratic prompts, which allowed users to recover focus between interruptions [31], the always-on interaction traces may have trapped users in a continuous cycle of self-monitoring. By permanently embedding the feedback within the visual interface, the interaction traces may have forced users to constantly divide their attention between the primary analytical task and the secondary task of evaluating their own behavior, creating an unsustainable "metacognitive overhead" [76]. This phenomenon can be understood through the lens of *explicit monitoring theory* [2], which posits that heightened self-awareness causes individuals to consciously monitor the step-by-step execution of processes that are otherwise fluid. This tension is empirically supported by prior work investigating similar interaction traces, which found that participants prefer summative visualization feedback shown *after* a task over real-time and in-situ representations [70], further supporting the case for episodic over persistent feedback.

7.3 When Interaction Traces Succeed vs. Backfire

The contrast between our findings and the original Lumos evaluation [43] points to an important boundary condition: *task structure*. In the original Lumos evaluation, the tool was shown to increase users' awareness of their analytic behaviors in a task focused on "recommending the characteristics of movies that a movie production company should make next," a task with a clear and concrete output (a recommendation) [43]. Participants in that study reported that the interaction traces helped them "keep track of their exploration" and "remove the user's internal bias." However, in our study, the same interaction trace feature produced a significant decline in metacognitive awareness during open-ended exploration. We argue that in structured tasks with clear deliverables, real-time feedback may serve as useful guidance toward that goal. However, in open-ended exploration, such feedback may have introduced a competing "performance" goal (e.g., "making it green"), ultimately displacing the user's analytical questions. Notably, the original Lumos study hints at this tension, observing that some participants *"didn't know exactly what to do about the [red-green] cards"* – a signal that the feedback lacked actionable scaffolding even in a more structured context.

This divergence also highlights a key distinction in how visualization tools are evaluated: a tool that users find useful or usable does not necessarily equate to enhanced metacognitive awareness. The original Lumos study evaluated the system primarily through self-reported usefulness scores (e.g., 1-to-5 Likert scales) and qualitative interviews, finding that users generally appreciated the system's novel capabilities. In contrast, our MAI-based pre-post measure revealed metacognitive harm that satisfaction ratings alone would not have detected.

More broadly, these findings suggest that providing bias awareness, e.g., surfacing what users have or have not explored, does not itself constitute metacognitive regulation. Awareness without actionable scaffolding leaves users informed of their behavioral patterns but without the means to interpret or act on them, a distinction critical for future intervention design.

7.4 The Disruption-Correction Gap

Despite perceived disruption from both interventions, no significant differences in insight correctness were found across conditions. This may point to a distinction between *process disruption* and *analytical outcome*: while our interventions introduced moments of disruption to users' cognitive flow, they failed to provide the corrective scaffolding needed to improve reasoning quality. A pause only supports metacognitive regulation if it facilitates targeted mental work, such as reconsidering assumptions or questioning initial hypotheses. This suggests that simply introducing friction is insufficient; effective metacognitive support needs to move beyond disruption to offer guidance that actively addresses the cognitive biases inherent in data exploration [14, 38, 67]. This pattern resonates with recent work by Ojewale et al. [47], who found that combining visual and verbal uncertainty cues in LLM-assisted decision making improved users' subjective sensitivity to accuracy yet simultaneously amplified overreliance on incorrect outputs, revealing a dissociation between subjective and behavioral calibration. Our findings suggest an analogous dissociation: interventions designed to support metacognitive awareness did not translate into improved analytical outcomes, reinforcing the broader principle that adding more supportive signals is insufficient without ensuring they produce behavior-relevant change.

7.5 Design Implications

Taken together, our findings point to two actionable principles for metacognitive support tools in visual analytics. First, designers should **scaffold actionability over status**: rather than displaying a static indicator (e.g., a red/green balance score), systems should pair awareness cues with non-prescriptive prompts that invite exploration (e.g., "You have deeply explored Attribute X; would you like to overlay Attribute Y for comparison?"). This ensures users are offered a pathway forward rather than merely feeling evaluated [32]. Second, systems should **support, rather than replace, users' goals**: feedback should be presented as an adaptable resource rather than an imposed standard. Allowing users to toggle feedback, adjust sensitivity, or declare their analytical intent (e.g., broad exploration vs. deep drill-down) would enable the system to support diverse sensemaking strategies rather than penalizing them.

7.6 Limitations and Future Work

Several limitations should be considered when interpreting our findings. First, the sample size for interviews was small (three participants per intervention condition and two for the CONTROL condition), limiting the generalizability of qualitative insights. While the themes we identified were generally consistent across participants within each group, we cannot rule out that additional interviews could reveal other mechanisms or individual differences.

Second, the Socratic prompts were delivered at moments determined by the experimenter's observational judgment. Participants believed the prompts were system-generated, but they had no control over when prompts appeared. This raises a potential confound: the observed effects may depend on when prompts appear [21, 30], not just what they ask. Future research should systematically compare different timing strategies, for instance, (i) prompts triggered by user actions (e.g., prolonged inactivity, repeated filters), (ii) prompts that appear at predefined intervals, or (iii) prompts that users can request on demand.

Additionally, the prompting strategy was fixed across all participants and sessions. Future work could investigate how an adaptive system could learn from user behaviors over time, tailoring interventions to individual needs and preferences, adapting if users are finding prompts helpful vs. distracting. To design these types of personalized interventions, future work should examine the relationship between personality traits, prior experiences, etc., and intervention efficacy and receptivity.

8 CONCLUSION

We investigated whether contextual Socratic questioning and visualized interaction traces could enhance metacognitive awareness during visual data exploration. Through a between-subjects study, we found that neither intervention supported regulation as expected; instead, we observed a gradient of disruption, suggesting that more intrusive interventions may inadvertently undermine metacognitive awareness. Our findings highlight that fostering metacognitive awareness requires careful consideration of cognitive load, user agency, and how design choices shape users' relationships with both the tool and the task. We hope our unexpected findings and insights gained from our analysis help inform future efforts to support deeper, more reflective data exploration.

ACKNOWLEDGMENTS

We are grateful to our CAV lab mates and to all those who contributed their time and effort to assist with the early pilot studies. This work was supported by NSF award IIS-2340539.

REFERENCES

[1] L. Battle and A. Ottley. What exactly is an insight? a literature review. *2023 IEEE visualization and visual analytics (VIS)*, pp. 91–95, 2023. doi: 10.48550/arXiv.2307.06551 5, 6

[2] S. L. Beilock and T. H. Carr. On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, 130(4):701–725, 2001. doi: 10.1037//0096-3445.130.4.701 9

[3] N. Boonprakong, B. Tag, J. Goncalves, and T. Dingler. How do hci researchers study cognitive biases? a scoping review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, art. no. 473, 20 pp. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3713450 2

[4] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996. doi: 10.1201/9781498710411-35 5

[5] K. Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 981–992, 2016. doi: 10.1145/2858036.2858498 5

[6] Carnegie Endowment for International Peace. Polarization, democracy, and political violence in the united states: What the research says, September 2023. Accessed: April 10, 2025. 3

[7] M. Chen, Y. Liu, and E. Wall. Unmasking Dunning-Kruger Effect in Visual Reasoning & Judgment. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):743–753, Jan. 2025. doi: 10.1109/TVCG.2024.3456326 2

[8] M. Chen, A. Yang, S. Min, K. A. Hamilton, and E. Wall. A novel lens on metacognition in visualization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–18. ACM, 2025. doi: 10.1145/3706598.3714400 1, 2

[9] E. K. Choe, B. Lee, et al. Characterizing visualization insights from quantified selfers' personal data presentations. *IEEE computer graphics and applications*, 35(4):28–37, 2015. doi: 10.1109/MCG.2015.51 6

[10] G. I. Clark and S. J. Egan. The socratic method in cognitive behavioural therapy: A narrative review. *Cognitive Therapy and Research*, 39:863–879, 2015. doi: 10.1007/s10608-015-9707-3 2

[11] CNN. National house exit polls 2022, 2022. Accessed: April 10, 2025. 3

[12] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pp. 193–200, 1993. doi: 10.1016/0950-7051(93)90017-N 3

[13] H. Delic and S. Bećirović. Socratic Method as an Approach to Teaching. *European Researcher*, 111:511–517, Nov. 2016. doi: 10.13187/er.2016.111.511 1

[14] 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, 2018. doi: 10.1109/TVCG.2018.2865233 2, 9

[15] E. Dimara, A. Bezerianos, and P. Dragicevic. The attraction effect in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):471–480, 2017. doi: 10.1109/TVCG.2016.2598594 2

[16] 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 2

[17] O. J. Dunn. Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. doi: 10.2307/2282330 6

[18] J. H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10):906–911, 1979. doi: 10.1037/0003-066X.34.10.906 1, 2

[19] J. H. Flavell. Metacognitive aspects of problem solving. In *The nature of intelligence*, pp. 231–236. Routledge, 2024. 1

[20] L. Fletcher and P. Carruthers. Metacognition and reasoning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1366–1378, 2012. doi: 10.1098/rstb.2011.0413 1

[21] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler et al. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(1):119–146, 2005. doi: 10.1145/1057237.1057243 9

[22] S. R. Gomez, H. Guo, C. Ziemkiewicz, and D. H. Laidlaw. An insight-and task-based methodology for evaluating spatiotemporal visual analytics. In *2014 IEEE conference on visual analytics science and technology (VAST)*, pp. 63–72. IEEE, 2014. doi: 10.1109/VAST.2014.7042482 6

[23] F. L. Greitzer, C. F. Noonan, and L. Franklin. Cognitive foundations for visual analytics. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2011. doi: 10.2172/1013936 2

[24] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics*, 22(1):51–60, 2015. doi: 10.1109/TVCG.2015.2467613 6

[25] S. Ha, S. Monadjemi, R. Garnett, and A. Ottley. A unified comparison of user modeling techniques for predicting data interaction and detecting exploration bias. *IEEE transactions on visualization and computer graphics*, 29(1):483–492, 2023. doi: 10.1109/TVCG.2022.3209476 2

[26] C. He, L. Micallef, L. He, G. Peddinti, T. Aittokallio, and G. Jacucci. Characterizing the quality of insight by interactions: A case study. *IEEE transactions on visualization and computer graphics*, 27(8):3410–3424, 2020. doi: 10.1109/TVCG.2020.2977634 6

[27] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Commun. ACM*, 55(4):45–54, Apr. 2012. doi: 10.1145/2133806.2133821 1

[28] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue*, 10(2):30–55, 2012. doi: 10.1145/2133806.2133821 1

[29] Y.-R. Ho, B.-Y. Chen, and C.-M. Li. Thinking more wisely: using the socratic method to develop critical thinking skills amongst healthcare students. *BMC Medical Education*, 23:173, 2023. doi: 10.1186/s12909-023-04134-2 2

[30] S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler et al. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pp. 257–264. Association for Computing Machinery, New York, NY, USA, 2003. doi: 10.1145/642611.642657 9

[31] S. T. Iqbal and E. Horvitz. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pp. 677–686. Association for Computing Machinery, New York, NY, USA, Apr. 2007. doi: 10.1145/1240624.1240730 9

[32] I. Jivet, M. Scheffel, H. Drachsler, and M. Specht. Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. In *European Conference on Technology Enhanced Learning*, pp. 82–96. Springer, 2017. doi: 10.1007/978-3-319-66610-5_7 9

[33] H. Kallio, K. Virta, and M. Kallio. Modelling the components of metacognitive awareness. *International Journal of Educational Psychology*, 7(2):94–122, 2018. doi: 10.17583/ijep.2018.2789 3

[34] H. Kallio, K. Virta, M. Kallio, A. Virta, F. R. Hjardemaal, and J. Sandven. The utility of the metacognitive awareness inventory for teachers among in-service teachers. *Journal of Education and Learning*, 6(4):78–91, 2017. doi: 10.5539/jel.v6n4p78 3

[35] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization: Human-centered issues and perspectives*, pp. 154–175. Springer, 2008. 1

[36] D. A. Keim, F. Mansmann, and J. Thomas. Visual analytics: how much visualization and how much analytics? *SIGKDD Explor. Newsl.*, 11(2):5–8, May 2010. doi: 10.1145/1809400.1809403 1

[37] 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, 2016. doi: 10.1109/TVCG.2016.2598920 3

[38] S. Li, T. J. Davidson, C. Xiong Bearfield, and E. Wall. Confirmation bias: The double-edged sword of data facts in visual data communication. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, art. no. 1175, 16 pp. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3713831 1, 9

[39] Z. Liu and J. Heer. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics*, 20(12):2122–2131, 2014. doi: 10.1109/TVCG.2014.2346452 6

[40] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, Apr. 1986. doi: 10.

1145/22949.22950 1

[41] 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*, 18(1):50–60, Mar. 1947. Publisher: Institute of Mathematical Statistics. doi: 10.1214/aoms/1177730491 6

[42] T. Munzner. Visualization analysis and design. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Courses*, SIGGRAPH Courses '25, art. no. 10, 2 pp. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3721241.3733989 1

[43] A. Narechania, A. Coscia, E. Wall, and A. Endert. Lumos: Increasing Awareness of Analytic Behavior during Visual Data Analysis. *IEEE TVCG*, 2022. doi: 10.1109/TVCG.2021.3114827 1, 2, 9

[44] C. North. Toward measuring visualization insight. *IEEE computer graphics and applications*, 26(3):6–9, 2006. 5

[45] A. Nussbaumer, K. Verbert, E.-C. Hillemann, M. A. Bedek, and D. Albert. A framework for cognitive bias detection and feedback in a visual analytics environment. In *2016 European Intelligence and Security Informatics Conference (EISIC)*, pp. 148–151, 2016. doi: 10.1109/EISIC.2016.038 2

[46] K. Ohtani and T. Hisasaka. Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, 13(2):179–212, 2018. doi: 10.1007/s11409-018-9183-8 2

[47] V. Ojewale, J. Ryan, S. Venkatasubramanian, and C. M. Boykin. More is not better: Visual uncertainty cues and the fragility of trust calibration in llm-assisted decision making. *Computers in Human Behavior: Artificial Humans*, p. 100307, 2026. doi: 10.1016/j.chbah.2026.100307. 9

[48] J. C. Overholser. Elements of the socratic method: I. systematic questioning. *Psychotherapy: Theory, Research, Practice, Training*, 30(1):67–74, 1993. doi: 10.1037/0033-3204.30.1.67 2

[49] J. R. Paden, A. Narechania, and A. Endert. Biasbuzz: Combining visual guidance with haptic feedback to increase awareness of analytic behavior during visual data analysis. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–8. ACM, 2024. doi: 10.1145/3613905.3651064 2

[50] C. A. Padesky. Socratic questioning: Changing minds or guiding discovery? In *Invited keynote address delivered at the European Congress of Behavioural and Cognitive Therapies, London*, 1993. Retrieved from http://www.padesky.com/clinicalcorner/ on Feb 20, 2026. 2

[51] S. Pandey and A. Ottley. Mini-vlat: A short and effective measure of visualization literacy. In *Computer Graphics Forum*, vol. 42, pp. 1–11. Wiley Online Library, 2023. doi: 10.1111/cgf.14809 3, 5

[52] R. Paul and L. Elder. The Art of Socratic Questioning, Part III. 1

[53] R. Paul and L. Elder. Critical thinking: The art of socratic questioning. *Journal of developmental education*, 31(1):36, 2007. 2, 3

[54] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology*, 70:153–163, 2017. doi: 10.1016/j.jesp.2017.01.006 5

[55] A. S. Pihlgren. Socrates in the classroom: Rationales and effects of philosophizing with children. Jan. 2008. 1

[56] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, vol. 5, p. 2, 2005. 2

[57] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613, 2014. doi: 10.1109/TVCG.2014.2346481 1

[58] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics*, 11(4):443–456, 2005. doi: 10.1109/TVCG.2005.53 5, 6

[59] A. Schlieder, J. Rummel, P. Albers, and F. Sadlo. The role of metacognition in understanding deceptive bar charts. In *2024 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 51–59, 2024. doi: 10.1109/BELIV64461.2024.00011 2

[60] G. Schraw and R. S. Dennison. Assessing metacognitive awareness. *Contemporary educational psychology*, 19(4):460–475, 1994. doi: 10.1006/ceps.1994.1033 1, 2, 3

[61] M. Shah. The Socratic Teaching Method: A Therapeutic Approach to Learning. *Teaching philosophy*, 31:267–275, Sept. 2008. doi: 10.5840/teachphil200831328 1

[62] S. Sivamurugan and S. Susila Sakthy. Framework for quick visuals insight discovery in the data workflows. In *2025 International Conference on Computing and Communication Technologies (ICCCT)*, pp. 1–7, 2025. doi: 10.1109/ICCCT63501.2025.11019696 1

[63] T. Tuononen, H. Hyytinen, M. Räisänen, T. Hailikari, and A. Parpala. Metacognitive awareness in relation to university students' learning profiles. *Metacognition and Learning*, 18(1):37–54, 2023. doi: 10.1007/s11409-022-09314-x 3

[64] M. V. J. Veenman, B. H. A. M. Van Hout-Wolters, and P. Afflerbach. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, 1(1):3–14, 2006. doi: 10.1007/s11409-006-6893-0 2

[65] L. N. Vittorio, S. T. Murphy, J. D. Braun, and D. R. Strunk. Using socratic questioning to promote cognitive change and achieve depressive symptom reduction: Evidence of cognitive change as a mediator. *Behaviour Research and Therapy*, 150:104035, 2022. doi: 10.1016/j.brat.2022.104035 2

[66] 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, 2019. doi: 10.1109/VISUAL.2019.8933779 2

[67] E. Wall, L. Blaha, C. Paul, and A. Endert. A formative study of interactive bias metrics in visual analytics using anchoring bias. In *Proceedings of the 17th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT)*, vol. 11747 of *LNCS*, pp. 555–575. Springer, 2019. doi: 10.1007/978-3-030-29384-0_34 2, 9

[68] 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 *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 104–115. IEEE, 2017. doi: 10.1109/VAST.2017.8585669 2

[69] E. Wall, S. Ghorashi, and G. Ramos. Using expert patterns in assisted interactive machine learning: A study in machine teaching. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part III 17*, pp. 578–599. Springer, 2019. doi: 10.1007/978-3-030-29387-1_34 3

[70] 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, 2021. doi: 10.1109/TVCG.2021.3114862 1, 2, 3, 9

[71] M. C. Wang, G. D. Haertel, and H. J. Walberg. What helps students learn. *Educational Leadership*, 51(4):74–79, 1993. 2

[72] 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 2

[73] Y.-T. C. Yang, T. J. Newby, and R. L. Bill. Using socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments. *The American Journal of Distance Education*, 19(3):163–181, 2005. doi: 10.1207/s15389286ajde1903_4 2

[74] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174053 1

[75] N. Zhao, J. G. Wardeska, S. Y. McGuire, and E. Cook. Metacognition: An effective tool to promote success in college science learning. *Journal of College Science Teaching*, 43(4):48–54, 2014. doi: 10.2505/4/jcst14_043_04_48 1

[76] J. Zumbach. Cognitive overhead in hypertext learning re-examined: Overcoming the myths. *Journal of Educational Multimedia and Hypermedia*, 15:411–432, Jan. 2006. 9