Beyond Introspection: Inferring Decision Weights from Others' Multi-Attribute Choices

[1]Trent N. Cash (ORCID: 0000-0001-6624-9616) and

[2]Daniel M. Oppenheimer (ORCID: 0000-0002-2363-4220)

[1]School of Psychology, University of Waterloo, Canada

[2]Department of Social and Decision Sciences, Carnegie Mellon University, USA

[2]Department of Psychology, Carnegie Mellon University, USA

Author Note

Correspondence concerning this article should be addressed to Trent N. Cash, School of Psychology, University of Waterloo, 200 University Ave. West, PAS 4042, Waterloo, ON, N2L 3G1. E-Mail: tcash@uwaterloo.ca, Phone: 330-451-9972. This work was supported by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (Award #2333553), and an Academic Research Grant from Sawtooth Software, Inc. that provided us with free access to their Lighthouse Studio CBC software. Trent N. Cash was also supported by the Lupina Foundation. No grantor had any role other than financial support. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the grantors. Supplemental Materials are available at: https://osf.io/t8sbd/?view_only=87691388e51a4291ac1feb65f8f97fb6

©American Psychological Association, 2026. This paper is not the copy of record and may not exactly replicate the authoritative document published in *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Abstract

Extant theory suggests that metacognitive judgments are informed by both privileged metacognitive cues and observable, external cues that can be evaluated via non-privileged mechanisms, such as statistical regularity monitoring and top-down, domain-specific beliefs. Across three studies implementing a variant of the learner-observer-judge paradigm, we explore the extent to which decision makers and observers rely on each of these mechanisms to infer the processes underlying a decision maker's multi-attribute choice behavior. In Study 1, we demonstrate that decision makers (n = 221) more accurately estimate the weights they applied to each attribute in a multi-attribute choice task than observers (n = 220) who studied their choices. This contrast highlights the value of privileged metacognitive cues. However, observers' estimates were still weakly correlated with decision makers' true weights, suggesting that reasonably accurate inferences can be made, even without access to privileged metacognitive cues. In Study 2, we demonstrate that the relative (but not absolute) accuracy of observers' (n = 218) inferences decreases when the context of the decision is obscured, suggesting that domain-specific beliefs slightly improve metacognitive inferences above and beyond what can be inferred via statistical regularity monitoring alone. In Study 3, we replicate these findings and investigate differences in subjective experiences (e.g., motivation, effort) across conditions. Together, our findings suggest that privileged metacognitive cues do improve the accuracy of decision makers' metacognitive knowledge, but that observers can make reasonably accurate inferences about decision makers' choice behavior, particularly when they can combine statistical regularity monitoring with domain-specific beliefs.

Keywords: *Metacognition; Metacognitive Knowledge; Metareasoning; Decision Making; Multi-Attribute Choice*

Introduction

Imagine that you are on your lunch break, trying to figure out what you want to eat. You walk down the street, passing dozens of food trucks until you finally stop at one and buy a sandwich. You return to the office and your coworker asks: "Why did you stop at the sandwich truck instead of the taco truck?" To answer this question, you think back to what was going through your head when you were walking down the street. You recall that the sandwich shop was cheaper than the taco truck – which felt like an important distinction at the time. But you also recall that there was a long line at the taco truck, so even if you wanted a taco, you wouldn't have had time to get one because your lunch break was only 30 minutes. Finally, you recall that the sandwich truck just seemed right at the time, in a way that you couldn't quite explain.

In the metacognition literature, these small bits of information about what you thought about, what you noticed, and how you felt while making your decision are referred to as metacognitive cues (Ackerman & Thompson, 2017; Koriat, 1997). These metacognitive cues can be used to make sense of your cognitive experiences – which can be helpful when nosy coworkers ask you to justify your lunchtime decisions (or when your boss asks you to justify important business decisions). But these cues are not all the same. The first two cues – the price and the lines – were based on observable factors in the world. If another person was watching you, they could have guessed that you wanted a cheaper, faster lunch. In contrast, the feeling that you had when you thought about the sandwich truck, assuming you weren't visibly jumping for joy, was an internal experience that was private to you. As such, you had access to more information about your decision-making process than the person observing you.

How important is this privileged metacognitive insight for developing accurate metacognitive knowledge of how you make decisions? Could an observer generate

metacognitive knowledge that is just as accurate as yours? These questions have been investigated for other cognitive tasks, such as learning (Vesonder & Voss, 1985), but little is known about our ability to make inferences about others' complex decision-making processes. To fill this gap in the literature, we present three studies in which we investigate the ability of observers to accurately infer decision makers' decision weights in multi-attribute choice contexts and isolate the mechanisms underlying this inferential process.

Introspection

People's ability to derive insights about their own cognitive processes is a topic that has interested scientists for centuries (Boring, 1953). This debate was historically framed as a discussion of the existence, accuracy, and value of human introspection – one's direct and intentional access to the inner workings of their mind (Byrne, 2005; Engelbert & Carruthers, 2010; Vermersch, 2009). Classic philosophers often framed introspection as an integral component of the human experience (Descartes, 1641; Kant, 1781) and many modern philosophers still view introspection as an important method by which humans navigate a complex and cognitively demanding world (Schwitzgebel, 2024; Smithies & Stoljar, 2012; Vermersch, 2009).

However, psychological research has called into question the extent to which humans have direct, introspective access to their own cognitive processes. For example, a seminal study by Nisbett and Wilson (1977) found that participants' verbal reports of their decision-making processes were significantly misaligned with their actual choice behavior. Nisbett and Wilson (1977) argued that the reason for this was that participants lack direct, introspective access to how they make decisions and instead form beliefs about their decision processes based on a priori causal theories. Since then, numerous psychological studies in a variety of cognitive

domains have supported the conclusion that humans' introspective capacities are limited or, at the very least, fallible (Bainbridge, 1979; Cash & Oppenheimer, 2024, 2025a; Morris et al., 2025; Nisbett & Bellows, 1977; Pronin, 2009; Wilson & Nisbett, 1978).

Cue-Based Metacognition

One leading theoretical perspective that provides an alternative to the direct-access account of introspection is cue-utilization theory (Koriat, 1997). This theory suggests that, when we engage in cognitive tasks, we implicitly observe cues that we can use to make inferences about the inner workings of our mind (Ackerman, 2019; Ackerman & Thompson, 2017; Koriat, 1997). Cue-utilization theory was initially proposed, and has been widely accepted, in the context of memory and learning (Hoch et al., 2023; Koriat, 1997; Roebers et al., 2019; Undorf et al., 2018; Undorf & Bröder, 2020). However, more recent research and theory has suggested that metacognition can also be cue-based in domains of reasoning, including judgment, decision making, and problem solving (Ackerman & Thompson, 2017; Fiedler et al., 2019).

Koriat (1997) proposed three kinds of metacognitive cues that individuals utilize to inform their metacognitive beliefs. *Intrinsic* metacognitive cues are derived from features of the cognitive tasks themselves. For example, participants use the pronounceability of anagrams (Topolinski et al., 2016) or the familiarity of words (Lauterman & Ackerman, 2024) to inform metacognitive judgments about the solvability of puzzles. *Extrinsic* metacognitive cues capture information about the conditions under which the cognitive task was completed. For instance, participants report greater confidence when they are given more time to solve tasks (Shynkaruk & Thompson, 2006), when they are more familiar with a task (Markovits et al., 2015), or when others generate solutions to the task more quickly (Topolinski & Reber, 2010). Finally, *mnemonic* cues capture the individual's internal experience while completing the task, such as

how easy or fluent the reasoning task felt (Alter & Oppenheimer, 2009; Hertzog et al., 2003; Oppenheimer, 2008; Schwarz et al., 2021). When tasks are more fluent, decision makers report greater confidence (Ackerman & Zalmanov, 2012; Sanchez & Dunning, 2020) and feel less need to engage in further reasoning (Alter et al., 2007, 2013; Thompson et al., 2013).

Cue-Based Inferences About Others' Cognition

Unlike introspection, many of the cues that are utilized to inform metacognitive judgments are observable; they reflect statistical relationships that exist, or are perceived to exist, in the world (Castel, 2008; Koriat, 1997; Susser & Mulligan, 2019). As such, we can use the same cues that inform our metacognitive judgments to make inferences about others' cognitive processes. Under the right circumstances, all three types of metacognitive cues can be either private to the individual or observable by an external entity. For example, the *intrinsic* cue of task difficulty has both observable components (e.g., the pronounceability of an anagram; Topolinski et al., 2016) and private components (e.g., knowing that anagram problems are harder for me, personally, than for most people). Similarly, the *extrinsi*c cue of time constraints (Shynkaruk & Thompson, 2006) may be visible to an observer (e.g., a large public sign indicating that the café in which you are working closes at 5pm) or private to the individual completing the task (e.g., internal knowledge that you have to pick up your child from daycare at 5pm). Lastly, *mnemonic* cues, like fluency (Oppenheimer, 2008), are often private to the individual (e.g., an internal feeling of struggle) but may be expressed outwardly (e.g., looking strained while completing a task). In experimental settings, the observability of these cues can be manipulated to investigate how metacognitive accuracy varies depending on the availability of each cue.

In the metamemory literature, this kind of experimental manipulation has often been conducted using the learner-observer-judge paradigm introduced by Vesonder and Voss (1985). In this paradigm, one participant, the learner, completes a learning task and predicts which items they will correctly recall, a metacognitive judgment known as a Judgment of Learning, or JoL (Matvey et al., 2001; Undorf & Erdfelder, 2011). Another participant, the observer, is yoked to a learner, observes the learner while they complete some portion of the task, then predicts which items the learner will correctly recall. Lastly, the judge is yoked to a learner and predicts which items the learner will recall based only on knowledge of the item set, without any knowledge about the specific learner. In a sentence-recall task, Vesonder and Voss (1985) found that learners and observers were similarly accurate in their predictions about which sentences the learner would recall. In contrast, judges were less accurate, suggesting that Observers were able to use observable cues to make inferences about the learner's cognitive processes, above and beyond the cues that judges could glean from the item set alone. In a similar study, Matvey et al. (2001) found that observers who watched a learner study words were able to use the learner's latency during the study phase to meaningfully inform JoLs.

However, numerous studies have also demonstrated that learners and observers often use different cues or integrate the same cues in different ways (Wei et al., 2026). For example, Serra and Ariel (2014) found that learners were more sensitive to their past performance than observers were when generating JoLs. Both Koriat and Ackerman (2010) and Undorf and Erdfelder (2011) demonstrated that observers used study time as a cue for JoLs, but only when they experienced the task as a learner first, suggesting that the relationship had to be gleaned through experience. Yang et al. (2018) found that learners used response time as a cue for JoLs, but that Observers did not, positing that they may have instead relied on other, more salient cues, like font size.

Memory and metacognition also have idiosyncratic components that are accessible to learners, but not observers. For example, Tullis and Fraundorf (2017) demonstrated that observers' JoLs are less accurate than learners' JoLs for cue-target word pairs because observers lack access to idiosyncratic associations the learner may have between words (e.g., the learner may associate golf and hobby, while the observer has a stronger association for golf and boring). Similarly, Undorf et al. (2022) found that increasing the personal significance of an item increased both memory performance and metamemory judgments, thus introducing another cue to which observers are not privy.

In all, the results from the metamemory literature suggest that observers can use a variety of cues to make reasonably accurate inferences about a learner's cognitive processes, but that they also lack critical cues that limit the validity of their inferences (Tullis & Fraundorf, 2017; Wei et al., 2026). However, research investigating this capacity outside of the metamemory domain is rather sparse. In one interesting study, Mata and Almeida (2014) demonstrated that participants could accurately predict whether other participants gave intuitive or deliberative responses on the Cognitive Reflection Test (Frederick, 2005) based on how long they spent on the task or how much cognitive load they were under. In another study, Allwood (1994) demonstrated that individuals give better-calibrated confidence judgments for themselves than they do for others, but that resolution was similar. These limited findings highlight the need for research on Observers' inferences in other domains. Given the growing literature in the domain of meta-reasoning (Ackerman & Thompson, 2017), it would be of interest to apply paradigms similar to the learner-observer-judge paradigm to reasoning tasks.

Observers' Inferences in Multi-Attribute Choice

Here, we explore the degree to which decision makers' ability to accurately self-report the weight they place on various attributes during multi-attribute choice is dependent on privileged metacognitive cues. We do so by comparing the accuracy of decision makers' internal estimates of cue-weighting to inferences generated by external observers who lack access to these privileged metacognitive cues. Following recent research, we refer to decision makers' knowledge of their own and others' decision processes as a form of metacognitive knowledge (Cash & Oppenheimer, 2025a). Though it has come to have many meanings across various segments of the literature, we use the term metacognitive knowledge in the broad sense of knowledge that one has about cognition, be it their own or someone else's. This definition aligns with classic definitions by Flavell (1979, p. 907) – "Metacognitive knowledge consists primarily of knowledge or beliefs about what factors or variables act and interact in what ways to affect the course and outcome of cognitive enterprises" – and Dunlosky and Metcalfe (2009, p. 2) – "Metacognitive knowledge pertains to people's declarative knowledge about cognition."

Our question is rooted in an expansive psychological literature that has investigated participants' metacognitive knowledge of attribute weights by asking them to make a series of choices then to self-report the weight they applied to each attribute. The alignment between their self-reported weights and their actual choices serves as a measure of metacognitive knowledge (Cash & Oppenheimer, 2024, 2025a, 2025b; Hoepfl & Huber, 1970; Morris et al., 2025; Slovic, 1969; Slovic et al., 1972; Slovic & Lichtenstein, 1971). However, the extant literature has largely focused on quantifying the accuracy of decision makers' metacognitive knowledge of their own attribute weights, without exploring the extent to which that accuracy is a function of privileged metacognitive cues or, conversely, observable by an outside actor. The one exception is Morris et al. (2025), who recently found that observers generate less accurate estimates of decision makers'

weights than decision makers themselves, but did not explore the impact of different cues that may or may not be available to observers.

To investigate these cues, we must first consider the information that is available to an external observer in a multi-attribute choice context. If an observer knows what alternatives are available to the decision maker and the attributes on which the alternatives vary, the observer could track what choices are made and potentially extract statistical regularities in the decision maker's behavior (Perkovic & Orquin, 2018; Ruffman et al., 2012; Sherman et al., 2020). This could allow the observer to estimate the weight that the decision maker placed on each attribute. When decision makers do this for themselves it would be considered a form of metacognitive monitoring (Cash & Oppenheimer, 2025a), so we will refer to the capacity of observers to do so as *statistical regularity monitoring*. Statistical regularity monitoring is highly related to other socio-cognitive functions that rely on identifying and learning patterns in the environment, such as statistical learning (Perkovic & Orquin, 2018; Sherman et al., 2020; Thiessen, 2017), social learning (Gweon, 2021; Ruffman et al., 2012; Zheng & Wang, 2023), and strategic reasoning (Ioannou & Romero, 2014; Jordan, 1995; Nachbar, 1997).

Observers often also have access to explicit, top-down, domain-specific beliefs about how decision makers generally behave in a certain context (Aydogan, 2021; Heiphetz et al., 2013; Lee et al., 2018). For example, an observer could use cultural knowledge to infer that a typical decision maker places greater weight on the cost of a car than on the shape of the door handle. Similarly, observers may hold stereotypes about specific types of decision makers (Chaxel, 2015; Craig et al., 2022; Fabre et al., 2015). For example, an observer may correctly believe that politically liberal buyers tend to place greater weight on environmental concerns than politically conservative buyers (Cruz, 2017). By combining inferences made from observed

statistical regularities with domain-specific beliefs, it is plausible that an observer could generate a reasonably accurate estimate of the weight that a decision maker placed on various attributes, even without access to privileged metacognitive cues (Ackerman & Thompson, 2017; Alter & Oppenheimer, 2009; Koriat, 1997). Similarly, a decision maker could use these observable cues to make inferences about their own cognition, without considering any privileged metacognitive cues (as posited by self-perception theory, accounting for a seeming lack of privileged self-insight when people attempt to understand their own behaviors; Bem, 1972).

Absolute vs. Relative Accuracy

To fully measure the accuracy of decision makers' and observers' metacognitive knowledge, we must consider two types of accuracy that are commonly assessed in the metacognition literature. Relative accuracy, or resolution, is the degree to which item-level metacognitive judgments discriminate item-level performance (Fleming & Lau, 2014, Koriat, 2007). For example, a participant's confidence judgments would demonstrate relative accuracy if the participant was highly confident on questions they answered correctly and not-so-confident on questions they answered incorrectly. Relative accuracy is typically measured using gamma correlations or receiver-operating-characteristic (ROC) curves (Fleming & Lau, 2014). To apply this to the context of multi-attribute choice, we will define relative accuracy as the correlation between participants' self-reported decision weights for individual attributes and their (or someone else's) revealed decision weights for those attributes (Cash & Oppenheimer, 2025a). Because these weights are scale (not ordinal) variables, we use Pearson correlations.

On the other hand, absolute accuracy, or calibration, is the absolute discrepancy between a global metacognitive judgment (e.g., confidence) and a measure of overall performance (Fleming & Lau, 2014; Koriat, 2007). For example, absolute accuracy is often assessed by

asking participants to judge how confident they are in their performance on a task and then comparing this confidence judgment to their actual mean performance (Moore & Schatz, 2017). A participant whose confidence judgment is numerically close to their true mean performance is said to demonstrate good absolute accuracy. In the context of multi-attribute choice, we conceptualize absolute accuracy as the absolute, numerical difference between participants' self-reported decision weights and their (or someone else's) revealed decision weights. Following Cash and Oppenheimer (2025a), we operationalize absolute accuracy in two ways: 1) the Euclidean Distance between the two sets of weights; and 2) how often the two sets of weights make similar choice predictions (see Methods).

Current Studies

In the studies presented here, we use an adapted version of the learner-observer-judge paradigm (Undorf & Erdfelder, 2011; Vesonder & Voss, 1985) to investigate the extent to which privileged metacognitive cues, domain-specific beliefs, and statistical regularity monitoring improve the absolute and relative accuracy of metacognitive knowledge of attribute weights in a multi-attribute choice task. We isolate these three sources of metacognitive knowledge across three studies. In Study 1, we randomly assign participants to either make choices themselves (i.e., Buyers) or to observe another participant making choices (i.e., Observers). Both sets of participants then estimate the weight that the decision maker placed on each attribute. In this study, Observers have full information about the task itself and the choices the Buyers made, but lack access to several privileged metacognitive cues, such as how long Buyers studied each choice or how difficult each choice felt. As such, the accuracy of Observers' self-reported weights captures the degree to which Buyers' decision weights can be inferred via domain-specific beliefs and statistical regularity monitoring. On the other hand, differences in accuracy

between the Buyers and Observers would highlight the value of privileged metacognitive cues for generating accurate metacognitive knowledge and retaining it in memory.

In Study 2, we replicate the Observer condition from Study 1 but remove all contextual information about the domain in which the decision maker was making choices (i.e., No-Context Observers). This manipulation eliminates the observer's access to domain-specific beliefs (Aydogan, 2021; Heiphetz et al., 2013; Lee et al., 2018), thus forcing the observer to rely entirely on statistical regularity monitoring. If No-Context Observers' metacognitive knowledge is less accurate in Study 2 than Observers' metacognitive knowledge in Study 1, it would suggest that domain-specific beliefs can supplement statistical regularity monitoring when making inferences about the weights applied by other decision makers in a multi-attribute choice context. If there is no difference between the conditions, it would suggest that statistical regularity monitoring is equally effective with or without access to domain-specific beliefs. Lastly, in Study 3 we run a simultaneous replication of Studies 1 and 2 using a randomized experimental design that improves upon the internal validity of the initial studies and better controls for sources of error variance that could limit conclusions.

Transparency and Openness

For each study, we report how we determined our sample size, all data exclusions, all manipulations, and all measures. All studies follows JARS (Appelbaum et al., 2018). All data, analysis code, and research materials are available at https://osf.io/t8sbd/?view_only=87691388e51a4291ac1feb65f8f97fb6. Data were analyzed using R, version 4.1.2 (R Core Team, 2025) and the package *ggplot*, version 3.5.2 (Wickham, 2016). Study 1 (https://aspredicted.org/hwvj-mtdg.pdf), Study 2 (https://aspredicted.org/4s86-vrj5.pdf) and Study 3 (https://aspredicted.org/un8hi6.pdf) were all pre-registered on AsPredicted.

Study 1

Methods

Overview

This study occurred in two phases. In the first phase – the Buyer Phase – participants were invited to complete a variant of the *Knowledge of Weights* (*KoW*) paradigm, a validated and reliable experimental method for assessing metacognitive knowledge in multi-attribute choice (Cash & Oppenheimer, 2025a). Participants were asked to make hypothetical choices about which homes they would be most interested in buying based on a given attribute set, then to self-report the weight they placed on each attribute. In Phase 2 – the Observer Phase – new participants were yoked to a participant from the Buyer Phase and shown screenshots of the choices they had made. Observers then estimated the weight that they believed the Buyers had placed on each attribute. We then assessed the alignment between revealed weights estimated from the Buyers' choices to the Buyers' and Observers' stated weights, which captured the accuracy of participants' metacognitive knowledge. We then compared the accuracy of metacognitive knowledge for Buyers and Observers. This study was pre-registered: https://aspredicted.org/hwvj-mtdg.pdf. All analyses were pre-registered, except where noted. This study complies with APA ethical standards for the treatment of human participants and was approved by the Carnegie Mellon University IRB (STUDY2023_00000133).

Participants

Buyer Phase. To determine the required sample size, we conducted a simulation in which random agents completed the choice survey. We then measured the standard error of the utility estimates for these responses. Following conventional guidelines (Halverson, 2020), we sought

sufficient power to reduce the standard error of our utility estimates for the random agents to less than .05. The minimum number of participants required to achieve this goal was about 150, but we increased the sample to 220 to enhance precision. Doing so reduced the standard error of the utility estimates to .04 on average. Prolific recruited one more participant than requested, resulting in a total of 221 participants completing the study. Of these 221 participants, 151 (68.3%) identified as White, 33 (14.9%) as Black, 12 (5.4%) as Asian, six (2.7%) as Hispanic/Latino, one (0.5%) as Native American, 17 (7.7%) as more than one race or ethnicity, and one participant (0.5%) did not report their race or ethnicity. The sample included 150 (67.9%) females, 68 (30.8%) males, and three (1.4%) non-binary participants. The average age of our sample was 39.2 years (sd = 11.3) and 122 (55.2%) of the participants had a college degree. All participants passed at least three out of five attention checks (81.9% passed all five), so, following our pre-registration, no participants were excluded from our analysis. Participants were paid $2.00. Additional demographics are reported in the Supplemental Materials.

Observer Phase. To equate the two samples, we recruited 221 Prolific participants, all of whom completed the study. One participant failed three out of five attention checks and, following our pre-registration, was excluded from analysis. Of the remaining 220 participants, 138 (62.7%) identified as White, 52 (23.6%) as Black, 10 (4.5%) as Asian, 10 (4.5%) as Hispanic/Latino, and 10 (4.5%) as more than one race or ethnicity. The sample included 134 (60.9%) females, 83 (37.7%) males, two (0.9%) non-binary participants, and one (0.5%) participant who did not report their gender. The average age of our sample was 42.4 years (sd = 13.0) and 122 (55.5%) of the participants had a college degree. All included participants passed at least three out of five attention checks (83.6% passed all five). Participants were paid $2.00. Additional demographics are reported in the Supplemental Materials.

Procedures.

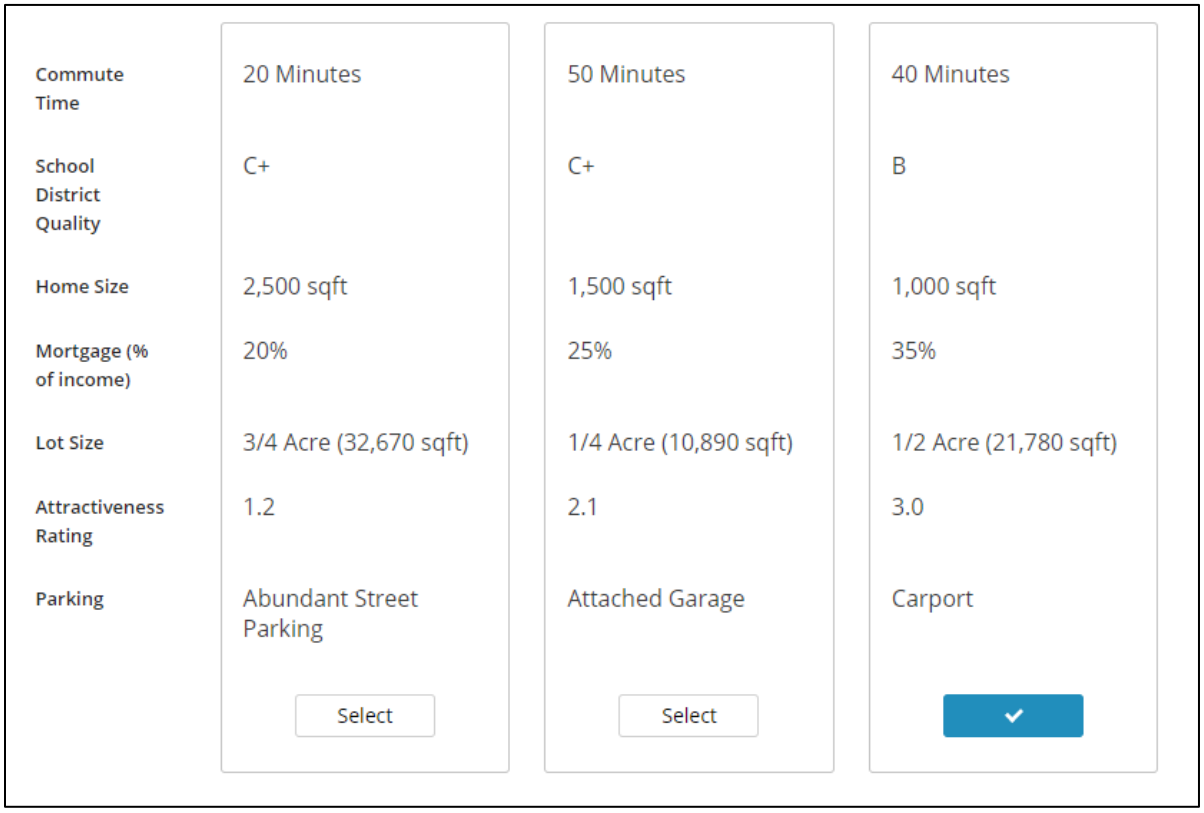
Buyer Phase. In line with the *Knowledge of Weights* paradigm (Cash & Oppenheimer, 2025a), participants first completed a Choice Based Conjoint (CBC) survey in which they were asked to imagine that they were on the real estate market, looking to purchase a new home. At the start of the survey, participants were given brief descriptions of the seven attributes on which the homes would vary (e.g. home size, commute time, etc.). These descriptions highlighted the five possible levels of each attribute, explained the scale on which each attribute was measured, and provided a reference point (e.g., mean) for typical real-world values for each attribute. Participants were then shown 14 sets of three hypothetical homes and asked to pick which one they would be most interested in buying from each set (see Supplemental Materials for exact language).

After making their choices, participants were asked to self-report the weight that they believed they put on each attribute when making their hypothetical choices. Importantly, these self-reports were phrased as retrospective reflections about their choice behavior, thus making them a metacognitive task, rather than a preference elicitation task (Cash & Oppenheimer, 2025a). Participants self-reported their weights in two formats, the order of which was randomized. In one format, participants were tasked with self-reporting the *percentage* of their decision-making process that was based on each attribute (i.e., their attribute weights). Their responses were required to sum to 100% across attributes. We refer to these percentages as Stated Attribute Weights or **SAW**s. In the other format, participants rated the importance of each attribute to their decision-making process on a Likert-style scale (1 = Not At All Important; 9 = Extremely Important). We will refer to these ratings as Attribute Importance Ratings, or **AIRs**.

To incentivize accurate self-reports, participants were told that the 5% of participants whose AIRs most accurately reflected their choice behavior would earn a $0.35 bonus and that the 5% of participants whose SAWs most accurately reflected their choice behavior would earn a $0.35 bonus, for a maximum potential bonus of $0.70. After each self-report measure, participants also self-reported their confidence in the accuracy of their self-reports. Finally, participants completed a standard demographic survey.

Observer Phase. The procedure for the Observer Phase was largely the same as the Buyer Phase, except that participants were tasked with studying the choices made by a Buyer instead of making choices for themselves. When Observers entered the study, they were informed that they would be paired with a participant from a previous study, then shown the same attribute descriptions as the previous participant. They were then told that they would see the choices made by the previous participant and that their goal was to "study the previous participant's choices and figure out which characteristics [attributes] influenced their decisions most." Participants were then yoked to a Buyer and shown screenshots of the 14 choices the Buyer made. An example screenshot is provided in Figure 1.

Figure 1: Sample Screenshot for Observers



Note. The blue checkmark indicates the home that was selected by the buyer when they made this choice.

Observers were required to spend at least 10 seconds studying each choice and were automatically moved onto the next choice after 60 seconds. The lower bound was designed to reflect the 10th percentile of choice time for Buyers (10 seconds) and the upper bound was designed to allow ample time, well above the 95th percentile of choice time for Buyers (42 seconds), without providing sufficient time for participants to completely disengage from the task. After Observers had studied all 14 choices made by the Buyer to whom they were yoked, they completed the SAW and AIR items, rating how heavily they thought the Buyer weighted each attribute. To promote incentive compatibility, participants were told that the 5% of

participants whose AIRs most accurately reflected the Buyer's choice behavior would receive a $0.35 bonus and the 5% of participants whose SAWs most accurately reflected the Buyer's choice behavior would receive a $0.35 bonus, for a total possible bonus of $0.70. After each measure, Observers also self-reported their confidence in the accuracy of their SAWs and AIRs. Once again, the survey ended with a brief demographic questionnaire. The Observer Phase was conducted about a month and a half after the Buyer Phase.

Materials.

Buyer Phase. To create the CBC survey, we used the six attributes validated by Cash and Oppenheimer (2025a): 1) Commute time (to work); 2) Home size; 3) Mortgage (cost as % of income); 4) School district quality; 5) Attractiveness of the home (as rated by buyers); and 6) Lot size). We then added a seventh attribute that is typically provided in advertisements for homes: 7) Parking. We then split each attribute into five discrete levels, which were used to create the alternatives in the CBC survey (see Supplemental Materials for details). We chose levels that reflected homes typically seen in the American housing market (e.g., sizes ranged from 1,000 – 3,000 square feet). We then generated 300 versions of the CBC survey. Each version of the survey contained a set of 14 choices, each of which included three hypothetical homes that were random permutations of the possible levels of each attribute. Since we had more versions than participants, each participant experienced a unique set of choice tasks. The CBC survey versions were generated using Lighthouse Studio (Sawtooth Software, Inc., 2024).

Observer Phase. We randomly selected 50 Buyers[1] whose choices would be used as stimuli for the Observer Phase. We then created a replica version of the CBC survey that each Buyer completed and marked the choice that was selected by the Buyer for each choice task, indicated by a blue checkmark (see Figure 1). We then took a screenshot of each choice task – which showed both the chosen and unchosen alternatives – and embedded these screenshots (n = 700) into a Qualtrics survey. Observers were then randomly yoked to one of the 50 Buyers and were shown all 14 of the screenshots generated from that Buyer.

Analysis & Results

Estimating Revealed Attribute Weights

To analyze this data, we first had to estimate the weights revealed by Buyers' choices during the CBC task. We refer to these as Revealed Attribute Weights, or **RAWs**. The first step in calculating RAWs was to estimate the utility that each participant placed on each level of each attribute. We did so using Hierarchical Bayes (HB) estimation, which is generally considered to be the gold standard for analyzing data from CBC surveys (Eggers et al., 2022; Hein et al., 2020; Orme, 2002). We conducted this HB analysis using the HB function in Lighthouse Studio (Sawtooth Software, Inc., 2024).

The HB model begins with a null preference model that sets all utilities to zero. The model then generates a new set of utilities for each level of each attribute for each participant and estimates how likely it was that each participant would have made the choices that they did if they had used the old or new utility sets. If their choices are more likely under the new set of utilities than the prior set of utilities, then the new utilities are retained as the current estimate

[1]We used only a random subset of Buyers' screenshots because creating the screenshots was a time-intensive process and we sought to minimize the time between phases. However, this imbalanced design was a threat to internal validity, so we corrected it in Study 3.

and used to inform the next set of utilities. This continuous updating process – known as a Markov Chain Monte Carlo – allows the model to converge on accurate utility estimates. For this study, we allowed the model to generate 20,000 iterations, the last 10,000 of which were averaged to generate point estimates of utilities.

Once this process is complete, it generates point estimates for the utility of each attribute for each participant (7 attributes x 5 levels = 35 utilities for each Buyer). We converted these utility estimates into attribute-level RAWs using the following formula (Cash & Oppenheimer, 2025a; Eggers et al., 2022; Orme, 2002), where U_j is a vector containing the utility values for the five levels of attribute *j,* U_i is a vector containing the utility values for the five levels of attribute *i,* and the set from which attribute *i* is pulled includes all attributes from the CBC, including *j:*

$$RAW_j = 100 * \frac{\max(U_j) - \min(U_j)}{\sum_{i=1}^{N}(\max(U_i) - \min(U_i))}$$

Because this formula calculates RAWs separately for each participant, it can account for Buyers having different preference orders across the levels of each attribute (e.g., one Buyer may prefer a larger lot, while another may prefer a smaller lot). A more thorough description of this method of estimating RAWs, as well as the strengths and weaknesses of the approach, is discussed at length by Cash and Oppenheimer (2025a). A technical overview of the HB function can be found in Sawtooth Software, Inc. (2021)

Descriptives.

Mean RAWs, SAWs, and AIRs for both phases are reported in Table 1.

Table 1: Mean RAWs, SAWs, and AIRs

Buyer Phase			
Attribute	**Mean RAW (sd)**	**Mean SAW (sd)**	**Mean AIR (sd)**
Commute Time	18.29 (8.34)	17.53 (16.04)	6.10 (2.34)
Home Size	11.79 (6.58)	12.38 (10.08)	5.48 (2.25)
Mortgage	14.54 (7.24)	20.91 (17.50)	6.54 (2.24)
School Quality	20.81 (13.54)	16.66 (21.55)	4.78 (3.04)
Attractiveness	12.11 (6.77)	8.93 (10.24)	4.52 (2.39)
Lot Size	9.28 (5.71)	11.17 (12.75)	4.75 (2.47)
Parking	13.17 (6.61)	12.43 (12.87)	5.45 (2.46)
Observer Phase			
Attribute	**Mean RAW (sd)**	**Mean SAW (sd)**	**Mean AIR (sd)**
Commute Time	17.87 (8.34)	18.02 (17.71)	5.77 (2.35)
Home Size	13.22 (6.43)	14.45 (10.38)	5.96 (2.04)
Mortgage	13.70 (7.33)	18.65 (16.76)	6.09 (2.15)
School Quality	20.95 (13.49)	14.54 (17.48)	5.07 (2.66)
Attractiveness	12.10 (6.89)	12.82 (12.17)	5.33 (2.30)
Lot Size	9.13 (6.07)	11.72 (10.41)	5.47 (2.23)
Parking	13.03 (6.73)	9.80 (11.00)	5.00 (2.30)

Note. RAWs were not identical for the Buyer Phase and Observer Phase because only a random subset of 50 Buyers' choices were used in the Observer phase. However, the RAWs were quite similar across the two phases, suggesting that randomization was effective.

Validating RAWs

To ensure that our RAW estimations were accurately capturing Buyers' decision weights, we exploratorily evaluated each choice task completed by each Buyer (n = 221 x 14 choices = 3,094 choices) and estimated the utility that would have been assigned to each alternative if the Buyer had used their RAWs as their true decision weights. We did so by multiplying the level of each attribute (scored as 1-5) by the RAW for that attribute and summing across the seven attributes, resulting in a utility value between 100-500 for each alternative. The attribute level scores were transformed to reflect the Buyer's rank-order preferences for the five levels of each attribute, with their highest-utility level being scored as a 5 and their lowest-utility level being scored as a 1 (see Cash & Oppenheimer, 2025a for validation of this method). We found that Buyers' RAWs correctly predicted (i.e., assigned the highest utility to) the alternative they selected for 94.0% (95% CI = 93.1 – 94.8%) of choice tasks, suggesting that our RAW estimation procedure was effective in capturing Buyers' true decision processes.

Data Transformation.

RAWs and SAWs are both compositional in nature – that is, the values are required to sum to a constant value (in this case, 100) for each participant. As such, each participant's attribute-level RAWs were not independent of one another, nor were their attribute-level SAWs. These statistical dependencies can create spurious relationships between RAWs and SAWs, particularly when directly comparing the two (e.g., via correlations). Following best practices, we mitigated these dependencies by applying a centered log-ratio ratio transformation to RAWs and SAWs before analyzing them (Cash & Oppenheimer, 2025a; Smithson & Broomell, 2024). We refer to the transformed weights as tRAWs and tSAWs.

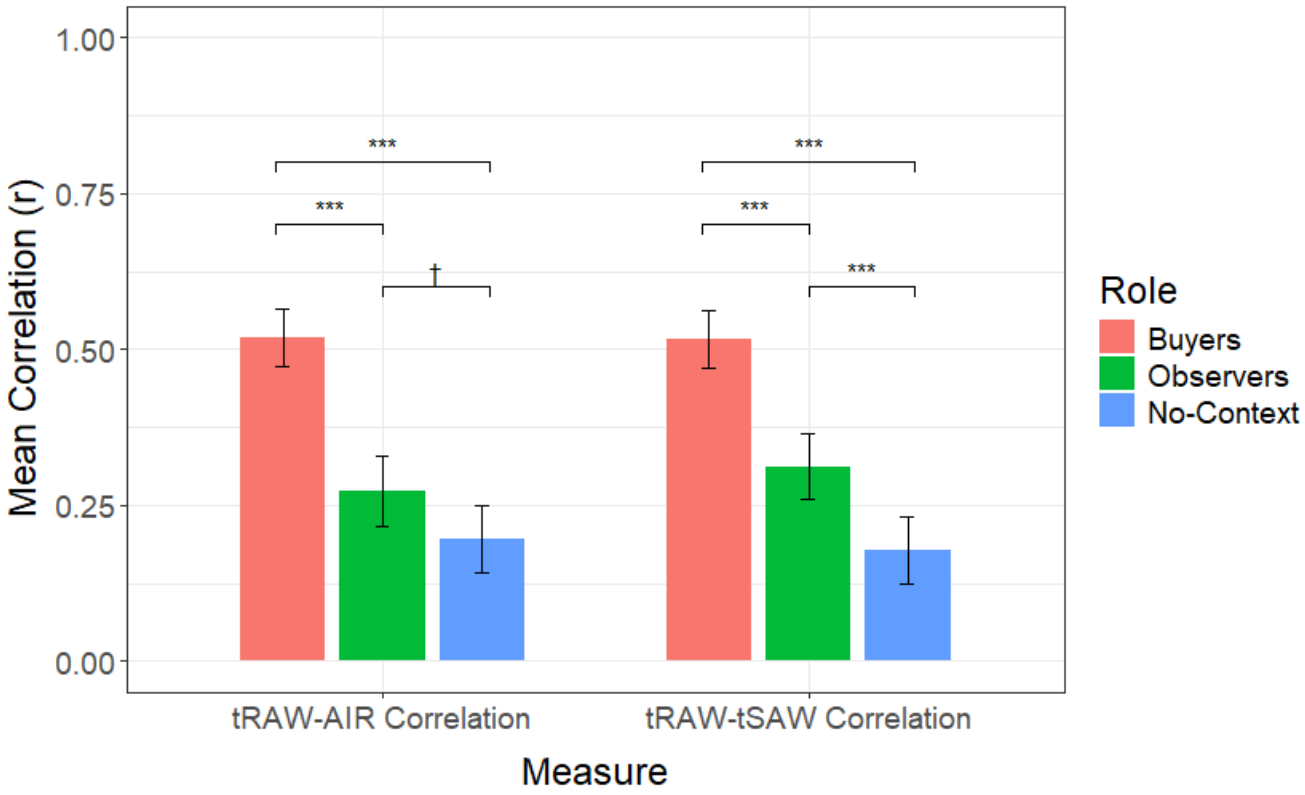
Correlations Between tRAWs, tSAWs, and AIRs

We first calculated the correlations between tRAWs and tSAWs across attributes for each participant. This metric captures a participant's ability to identify how heavily a decision maker weighted individual attributes relative to how heavily they weighted other attributes. The mean tRAW-tSAW correlation for Buyers was $r = .52$ ($sd = 0.36$) and the mean tRAW-tSAW correlation for Observers was $r = .31$ ($sd = 0.40$). The mean for Buyers was significantly greater than the mean for Observers, $t(431.97) = 5.63$, $p < .001$.

We next calculated the correlations between tRAWs and AIRs across attributes for each participant. The mean tRAW-AIR correlation for Buyers was $r = .52$ ($sd = 0.34$). The mean tRAW-AIR correlation for Observers was $r = .27$ ($sd = 0.42$). The mean for Buyers was significantly greater than the mean for Observers, $t(410.75) = 6.68$, $p < .001$. These correlational results, summarized in Figure 2, tell a consistent story: Buyers had more accurate metacognitive knowledge of their own attribute weights than could be inferred by Observers, but Observers were able to achieve correlations that indicated at least a weak ability to infer the Buyers' decision weights.

We originally pre-registered that we would report tRAW-tSAW and tRAW-AIR correlations at the attribute level (across participants). In response to reviewer feedback, and in line with conventions in the metacognition literature (Fleming & Lau, 2014), we chose to report participant-level correlations (across attributes). For transparency, attribute-level correlations are reported in the Supplemental Materials. This same deviation holds for Study 2 but was corrected for Study 3 in which both analyses were pre-registered.

Figure 2. tRAW-tSAW and tRAW-AIR Correlations, by Condition



*†p < .10; ***p < .001; No-Context Observer Data from Study 2.*

Euclidean Distance between tRAWs and tSAWs

As another metric of the accuracy of participants' metacognitive knowledge, we evaluated the Euclidean Distance between tRAWs and tSAWs for each participant. This measure captures the absolute distance between two sets of values – in this case, tRAWs and tSAWs. Lower Euclidean Distances suggest that two sets of values are more similar on an absolute scale, indicating more accurate metacognitive knowledge. We calculated Euclidean Distances according to the following formula (Cash & Oppenheimer, 2025a; Smithson & Broomell, 2024), where *D* denotes an individual's Euclidean Distance, *i* refers to the different attributes in the set, and *K* refers to the total number of attributes:

$$D = \frac{1}{K}\sqrt{\sum_{i=1}^{K}(tRAW_i - tSAW_i)^2}$$

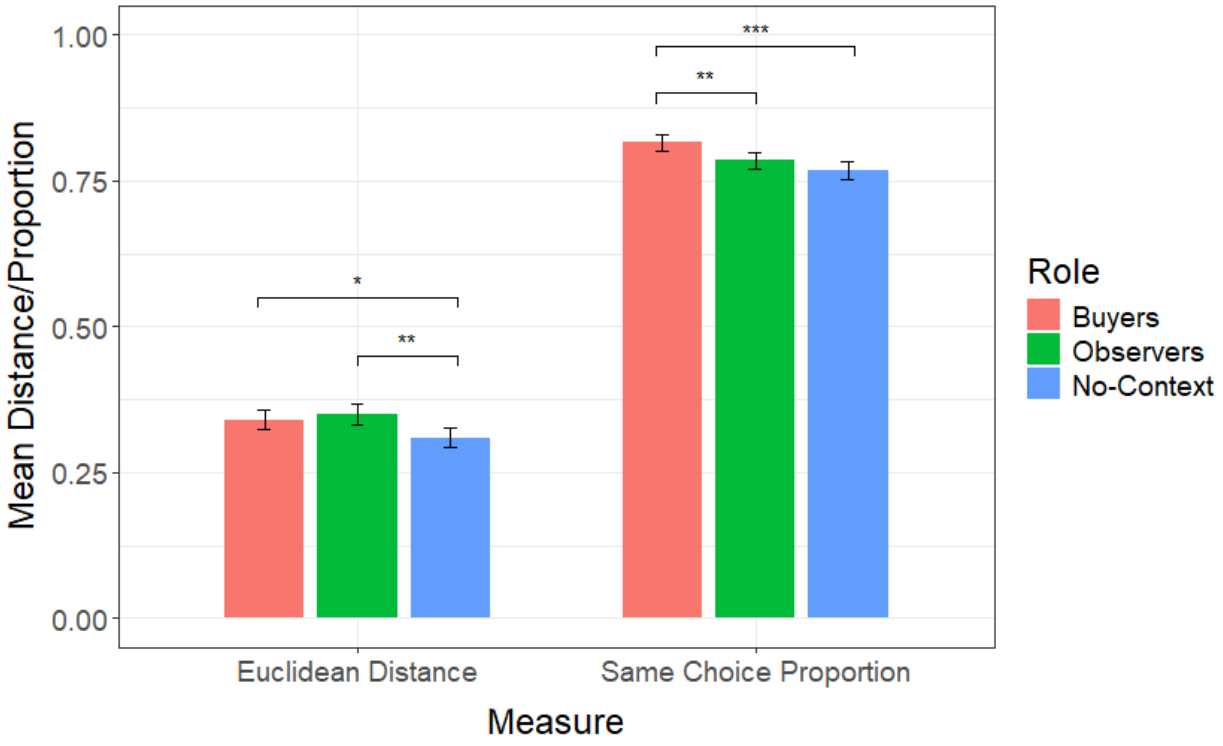
The mean Euclidean Distance for Buyers was 0.34 (*sd* = 0.13) and the mean Euclidean Distance for Observers was 0.35 (*sd* = 0.14). While the mean Euclidean Distance was lower for Buyers than Observers (see Figure 3), the difference was not significant, $t(435.72) = 0.57$, $p = 0.57$. The discrepancy between this result and the results from the correlational analyses suggests that Observers' SAWs may not be less accurate than Buyers' SAWs on an absolute scale, but that Observers' SAWs less accurately capture the *relative* importance of various attributes. We will return to this discrepancy in the general discussion. Results were qualitatively similar when we replicated these results with the sum of absolute differences between each set of weights (i.e., the City Block Distance) instead of Euclidean Distance (see Supplemental Materials).

RAW and SAW Choice Predictions

Using the same procedure as we did for validating RAWs, we evaluated each choice task completed by the Buyers (*n* = 221 x 14 choices = 3,094 choices) and each choice task seen by the Observers (*n* = 220 x 14 choices = 3,080 observed choices) and estimated the utility that would have been assigned to each alternative if the Buyer had used the SAWs provided by the participant (Buyer or Observer) as their true decision weights. We then evaluated the percentage of choice tasks for which the SAWs and the corresponding Buyer's RAWs predicted that the Buyer would have selected the same alternative. Greater alignment between RAW and SAW predictions indicates more accurate metacognitive knowledge. Buyers' SAWs predicted the same choices as their RAWs for 81.6% (95% CI = 80.2 – 83.0%) of choice tasks. Observers' SAWs predicted the same choice as their Buyers' RAWs for 78.5% (95% CI = 76.9 – 79.9%) of choice

tasks (see Figure 3). The proportion was significantly higher for Buyers than Observers, $X^2(1, N = 6{,}062) = 9.40$, $p = .002$, suggesting that Buyers' metacognitive knowledge of their attribute weights was more accurate than the weights inferred by Observers. As an exploratory robustness check, we ran a logistic multi-level model (choice tasks nested within individuals) predicting the likelihood of RAWs and SAWs making different predictions. The fixed effect of being an Observer (vs. Buyer) was significant ($B = 0.21$, $SE = .08$, $OR = 1.24$, $p = .01$). The full model is reported in the Supplemental Materials. As an additional robustness check, we demonstrate in the Supplemental Materials that the pattern of results is replicated when comparing SAW predictions to the choices that Buyers actually made, rather than the choices that their RAWs predicted they would make.

Figure 3. Mean Euclidean Distance and Same Choice Proportion, By Condition



**p < .05; **p < .01, ***p < .001; No-Context Observer Data from Study 2.*

Confidence

Finally, we evaluated participants' confidence in the accuracy of their SAWs and AIRs, which was reported on 0-100 scale. Buyers reported a mean confidence of 74.84 (*sd* = 21.41) for SAWs and a mean confidence of 79.00 (*sd* = 19.15) for AIRs. Observers reported a mean confidence of 59.25 (*sd* = 22.33) for SAWs and a mean confidence of 63.11 (*sd* = 21.04) for AIRs. Buyers were significantly more confident than Observers for both SAWs, $t(438.05) = 7.48$, $p < .001$, and AIRs, $t(434.76) = 8.29$, $p < .001$. This suggests that Observers were less confident in the accuracy of their inferences about decision makers' attribute weights than Buyers were in their own metacognitive knowledge. Given that Observers were less accurate, and significantly so on 3/4 measures, this lower confidence was seemingly warranted.

Study 1 Discussion

Across four measures of metacognitive knowledge, we found consistent evidence that Buyers could more accurately report the weight they placed on each attribute than Observers. This effect was statistically significant for three of four measures. These results suggest that Observers' lack of access to privileged metacognitive cues limited the accuracy of their insights about Buyers' decision-making processes, and that Buyers leveraged these privileged metacognitive cues to inform their metacognitive knowledge. However, we also found that Observers were not completely unable to extract weights from Buyers' choices. Indeed, Observers' stated weights were weakly-to-moderately correlated with the Buyers' revealed weights and predicted the same choices as the Buyers' stated weights in nearly 80% of choice tasks. This suggests that, even without access to privileged metacognitive cues, we can make meaningful inferences about others' cognitive processes through non-privileged processes, such as statistical regularity monitoring and domain-specific beliefs. As such, we conclude that these results suggest that metacognitive knowledge can be developed through observable cues, but that, perhaps unsurprisingly, privileged metacognitive cues increase the accuracy of metacognitive knowledge. In Study 2, we will further probe the mechanisms by which Observers generate inferences about Buyers' attribute weights. Specifically, we will seek to isolate the role of statistical regularity monitoring from the role of domain-specific beliefs.

Study 2

Methods

Overview

In Study 2, participants completed a variant of the Observer condition from Study 1 in which they were shown manipulated versions of the screenshots of the choices made by the Buyers. These manipulated screenshots obscured any contextual information about the decision domain (i.e., homebuying) and the attributes on which homes were compared. This was achieved by replacing attribute labels with Greek letters (e.g., Commute Time -> Alpha Score) and removing units of measure (e.g., acres) from the screenshots. This manipulation resulted in the participants acting as No-Context Observers who could make inferences about the Buyers' attribute weights based only on statistical regularity monitoring, without access to domain-specific beliefs. We then compared the accuracy of No-Context Observers' SAWs and AIRs to the accuracy of SAWs and AIRs generated by the Buyers and Observers in Study 1. Given that Buyers were generally more accurate than Observers, it would be surprising if they were not more accurate than No-Context Observers as well. The key comparison, however, is between Observers and No-Context Observers. If the No-Context Observers are less accurate than Observers, this would suggest that domain-specific beliefs play a critical role in the ability of Observers to infer decision makers' attribute weights. If not, it would suggest that statistical regularity monitoring alone is the driving force behind these inferences. Study 2 allows us to distinguish between these hypotheses. This study was pre-registered: https://aspredicted.org/4s86-vrj5.pdf. All analyses were pre-registered, except where noted. This study complies with APA ethical standards for the treatment of human participants and was approved by the Carnegie Mellon University IRB (STUDY2023_00000133).

Participants

To remain consistent with Study 1, we recruited 220 new Prolific participants, 219 of whom completed the study. Following our pre-registration, one participant was excluded for

failing 3/5 attention checks, leaving 218 viable participants. Of these participants, 152 (69.7%) identified as White, 57 (26.1%) as Black, two (0.9%) as Asian, one (0.5%) as Hispanic/Latino, one (0.5%) as Native American, and five (2.3%) as more than one race or ethnicity. The sample included 118 (54.1%) females, 97 (44.5%) males, and three (1.4%) non-binary participants. The average age of our sample was 37.5 years (sd = 12.0) and 184 (84.4%) of the participants had a college degree. All valid participants passed at least three out of five attention checks (77.5% passed all five). Participants were paid $2.51[2]. Additional demographic information is reported in the Supplemental Materials.

Procedure.

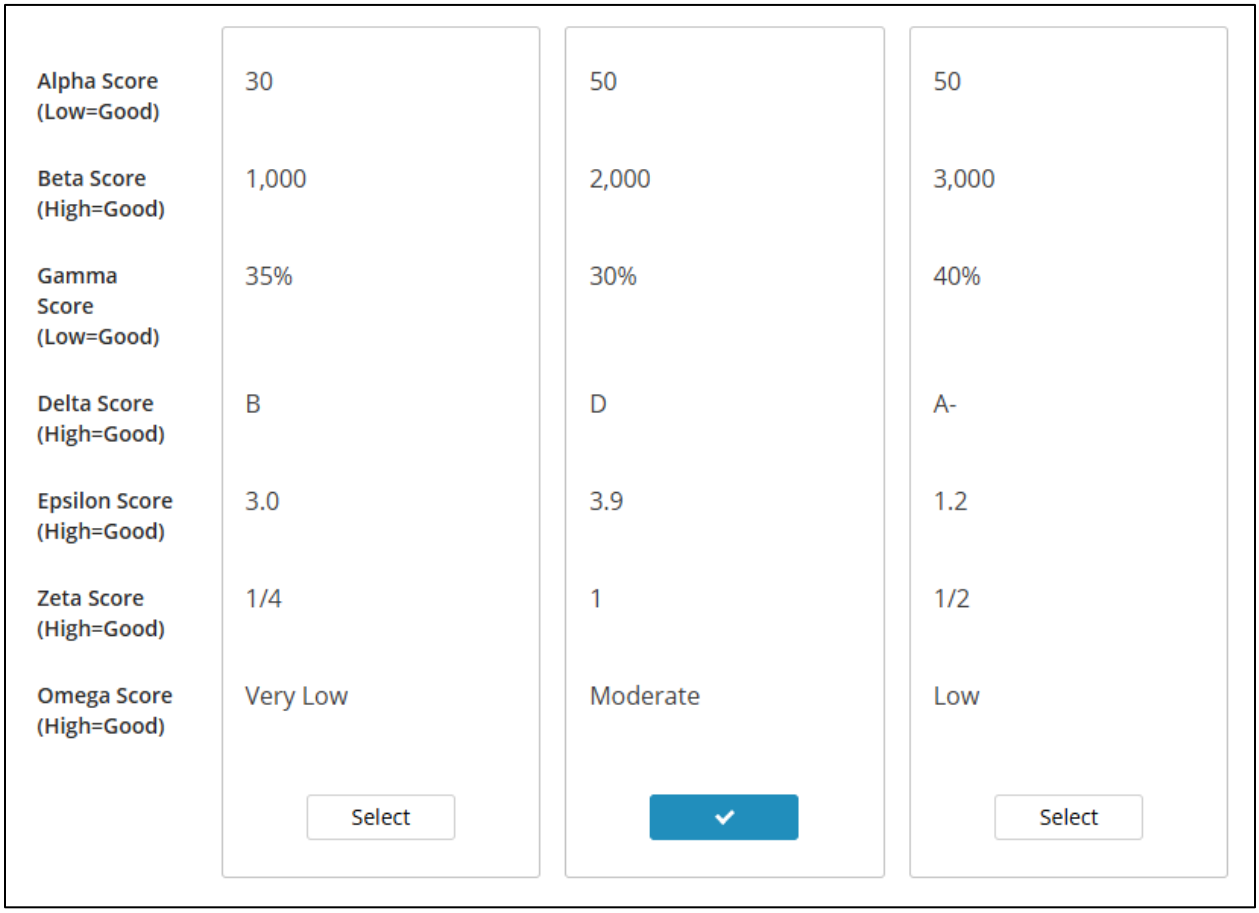
The procedure for the No-Context Observers was largely the same as the Observers from Study 1, with manipulations to eliminate contextual information. When No-Context Observers entered the study, they were informed that they would be paired with a Prolific participant from a previous study and shown the choices that they made between "hypothetical objects." They were instructed that they would not be told what those objects were. They were then given descriptions of the attributes on which the objects varied and told that the names of the attributes had been replaced with Greek letters.

These descriptions were largely identical to the description given in Study 1 but differed in two key ways. First, the high-level definitions of the attributes (e.g., "This characteristic indicated how large each home was") were removed and replaced with generic indicators of the typical range for the attribute (e.g., "Beta Score is a numeric measure that typically ranges from about 1,000 - 3,000"). Second, participants were informed about the typical directional

[2] Participants were paid more in Study 2 because the task took longer than Study 1. They were initially informed that they would be paid $2.00, but their rate was increased to maintain a fair compensation rate.

preference for each attribute (e.g., “Buyers typically prefer objects with lower Gamma Scores”). These directional preferences were determined based on the Study 1 data and aligned with reasonable intuitions (e.g., Buyers prefer lower mortgages). The exact language for each attribute description is provided in the Supplemental Materials. After reading the attribute descriptions, participants were instructed that their goal was to “study the previous participant’s choices and figure out which characteristics [attributes] influenced their decisions most.” Participants were then yoked to a Buyer and shown the manipulated screenshots of the 14 choices the Buyer made. An example manipulated screenshot is provided in Figure 4.

Figure 4: Sample Screenshot for No-Context Observers



Note. The blue checkmark indicates the home that was selected by the buyer when they made this choice.

No-Context Observers were required to spend at least 10 seconds studying each choice and were automatically moved onto the next choice after 60 seconds. After No-Context Observers had studied all 14 choices made by the Buyer to whom they were yoked, they completed the SAW and AIR questions, rating how heavily they thought the Buyer weighted each attribute. The same incentive scheme was used as in Study 1. After each measure, No-Context Observers also self-reported their confidence in the accuracy of their SAWs and AIRs. The survey concluded with a brief demographic questionnaire. In the demographic questionnaire,

participants were asked to guess what kinds of objects the previous participants were picking between and not a single participant guessed that the objects were homes. This suggests that we successfully obscured the context of the decision.

Materials

Using the same 50 randomly-selected Buyers as in the Observer Phase of Study 1, we created a replica version of the CBC survey that each Buyer completed. However, we made a series of manipulations to remove any contextual information. First, we replaced the names of each attribute with a Greek letter, followed by the word score (e.g., Commute Time -> Alpha Score; School Quality -> Delta Score). To avoid confusion, we also added language in each screenshot reminding them whether Buyers typically preferred higher or lower values for each attribute (e.g., "High = Good"). Second, we removed any units of measure that would indicate the choice domain, including square footage (Home Size) and acreage (Lot Size). The values themselves were unchanged (e.g., "3,000 sqft" -> "3,000"). Third, we converted the levels of the Parking attribute from a descriptive qualitative scale (ranging from Limited Street Parking to Attached Garage) to a generic Likert-style scale (Very Low to Very High). We then marked the choice that was selected by the buyer for each choice task, indicated by a blue checkmark (see Figure 4). Finally, we took a screenshot of each choice task – which showed both the chosen and unchosen alternatives – and embedded these screenshots (n = 700) into a Qualtrics survey. Context-Free Observers were then randomly yoked to one of the 50 Buyers and shown all 14 of the screenshots generated from that Buyer.

Analysis & Results

To assess No-Context Observers' accuracy, we compared their SAWs and AIRs to the RAWs estimated from the choices made by the Buyer (from Study 1) to whom they were yoked, using the same metrics as in Study 1.

Descriptives.

Mean RAWs, SAWs, and AIRs for No-Context Observers are reported in Table 2.

Table 2: Mean RAWs, SAWs, and AIRs

Attribute	Mean RAW (sd)	Mean SAW (sd)	Mean AIR (sd)
Alpha Score (Commute Time)	18.28 (8.54)	19.95 (12.54)	6.55 (2.09)
Beta Score (Home Size)	12.84 (6.29)	17.01 (10.35)	6.38 (1.84)
Gamma Score (Mortgage)	13.65 (7.17)	12.74 (7.58)	5.98 (1.83)
Delta Score (School Quality)	20.21 (13.30)	13.04 (9.42)	5.73 (2.14)
Epsilon Score (Attractiveness)	12.75 (7.07)	12.36 (9.40)	5.67 (2.00)
Zeta Score (Lot Size)	9.11 (6.08)	9.97 (7.31)	5.38 (2.05)
Omega Score (Parking)	13.16 (6.66)	14.93 (11.75)	6.17 (2.28)

Note. Mean RAWs were not identical to the Observer Phase of Study 1 because participants were randomly assigned to the 50 buyers, allowing for some noise in the number of times each buyer's choices were observed. However, the RAWs were quite similar for Observers and Context-Free Observers, suggesting successful randomization.

Data Transformation.

As in Study 1, we applied a centered log-ratio transformation to RAWs and SAWs to account for their compositional structure (Cash & Oppenheimer, 2025a; Smithson & Broomell, 2024). We again refer to the transformed values as tRAWs and tSAWs.

Correlations between tRAWs, tSAWs, and AIRs

We first calculated the participant-level tRAW-tSAW correlations in the same way as we did in Study 1. The mean tRAW-tSAW correlation for No-Context Observers was $r = .18$ ($sd = .40$), which was lower than the mean for Buyers ($r = 0.52$, $t(428.56) = -9.32$, $p < .001$) and the mean for Observers ($r = 0.31$, $t(435.93) = -3.51$, $p < .001$). We next evaluated participant-level correlations between tRAWs and AIRs, using the same methods as Study 1. The mean tRAW-AIR correlation for No-Context Observers was $r = 0.20$ ($sd = 0.39$). This was significantly lower than the mean correlation for Buyers ($r = .52$, $t(404.48) = -9.02$, $p < .001$) and marginally lower than the mean correlation for Observers ($r = .27$, $t(419.78) = -1.94$, $p = .05$). These correlational results, summarized in Figure 2, unsurprisingly show that No-Context Observers' SAWs and AIRs were less accurate than those generated by Buyers. More interestingly, however, they also demonstrate that No-Context Observers' SAWs and AIRs were at least marginally less accurate than those generated by Observers. This suggests that domain-specific beliefs augment the accuracy of Observers' inferences about Buyers' decision-making processes, at least to some extent.

Euclidean Distance Between tRAWs and tSAWs

We calculated Euclidean Distances between tRAWs and tSAWs in the same way as in Study 1. The mean Euclidean Distance for No-Context observers was 0.31 ($sd = 0.13$).

Surprisingly, this was significantly lower than the mean Euclidean Distance for both Buyers (M = 0.34, t(436.55) = -2.41, p = .02) and Observers (M = 0.35, t(434.69) = -2.88, p = .004). Results were qualitatively similar when we replicated these results with City Block Distance instead of Euclidean Distance (see Supplemental Materials). At face value, and in contrast to all other measures, these results suggest that No-Context Observers' inferences about Buyers' attribute weights were more accurate than the estimates generated by not only Observers, but also Buyers themselves. However, this result is confounded by the fact that there was less variance in SAWs for No-Context Observers than in either of the other two conditions (see Tables 1 and 2). This suggests that No-Context Observers achieved a lower Euclidean Distance not by better understanding the choice behavior of the buyer to whom they were yoked (which would result in higher correlations AND lower Euclidean Distances), but instead by reporting SAWs that were reasonably close to some meaningful prior that was shared across participants.

Given that No-Context Observers had very little information, we suspected that this prior was a roughly equal distribution of weights across attributes (14.29% per attribute). To investigate this suspicion, we ran an exploratory analysis in which we calculated the percentage of SAWs between 10-20% reported by participants in each condition. For No-Context Observers, 61.1% of SAWs fell between 10-20%. For Observers, this percentage was 45.9% and for Buyers it was 38.65%. In this sense, we can think of No-Context Observers as anchoring at an equal distribution and insufficiently adjusting from that anchor because of a lack of information (Epley & Gilovich, 2006; Furnham & Boo, 2011). However, because all attributes were relatively important to Buyers (mean RAWs = 9.11 – 20.21), this ended up reducing absolute error – as reflected in the Euclidean Distance metric – without reflecting a more accurate understanding of

the particular Buyer to whom each No-Context Observer was yoked. However, this is a post-hoc interpretation and should be interpreted with appropriate skepticism.

RAW and SAW Choice Predictions

Using the same methods as in Study 1, we next evaluated the percentage of choices tasks for which RAWs (estimated from Buyers) and SAWs made the same prediction about which alternative the Buyer would select. No-Context Observers' SAWs predicted the same alternative as Buyers' RAWs in 76.8% (95% CI = 75.2% - 78.3%) of choice tasks. This was significantly lower than the percentage of choices tasks correctly predicted by Buyers' own SAWs (81.6%), $X^2(1, N = 6{,}024) = 21.27$, $p < .001$, but not significantly different than the proportion predicted by Observers' SAWs (78.5%), $X^2(1, N = 6{,}032) = 2.34$, $p = .13$. As an exploratory robustness check, we ran a logistic multi-level model (choice tasks nested within individuals) predicting the likelihood of RAWs and SAWs making different predictions, with all three conditions included. When No-Context Observers were treated as the reference group, the fixed effect of being a Buyer was significant ($B = -0.30$, $SE = 0.09$, $OR = 0.74$, $p < .001$) but the effect of being an Observer was not ($B = -0.09$, $SE = .09$, $OR = 0.92$, $p = .32$). The full model is reported in the Supplemental Materials. In contrast to the correlational results, these findings suggest that the lack of domain-specific beliefs only mildly inhibited the accuracy of No-Context Observers, as they were not significantly less accurate than Observers. Again unsurprisingly, No-Context Observers' relatively limited information made their inferences less accurate than Buyers' metacognitive knowledge. As an additional robustness check, we demonstrate in the Supplemental Materials that the pattern of results is replicated when comparing SAW predictions to the choices that Buyers actually made, rather than the choices predicted by their RAWs.

Confidence

Finally, we evaluated participants' confidence in the accuracy of their SAWs and AIRs, reported on a 0-100 scale. No-Context Observers reported a mean confidence of 62.45 (*sd* = 25.37) for SAWs, which was significantly lower than mean confidence for Buyers ($M = 74.84$; $t(421.37) = -5.52$, $p < .001$) but no different than mean confidence for Observers ($M = 59.25$; $t(426.55) = 1.40$, $p = .16$). Similarly, No-Context Observers reported a mean confidence of 63.65 (*sd* = 24.60) for AIRs, which was also significantly lower than mean confidence for Buyers ($M = 79.00$; $t(407.73) = -7.28$, $p < .001$) but not significantly different than mean confidence for Observers ($M = 63.11$; $t(422.98) = 0.24$, $p = .81$). This suggests that Buyers were significantly more confident than both sets of Observers, but that the two sets of Observers were rather similar in their confidence.

Study 2 Discussion

In Study 2, we found several interesting patterns in the accuracy of No-Context Observers' inferences about Buyers' attribute weights. First, we found that No-Context Observers were significantly less accurate than Buyers in terms of all measures of metacognitive knowledge, except for the confounded Euclidean Distance metric. This is unsurprising, given that Study 1 demonstrated that Buyers outperformed Observers who had more information than the No-Context Observers. More interestingly, we found that No-Context Observers were consistently less accurate than Observers (except in terms of Euclidean Distance), but that the difference was small and only rose to significance for the tRAW-tSAW correlation metric. This suggests that domain-specific beliefs can increase the accuracy of metacognitive knowledge of attribute weights, but that these beliefs provide relatively little value above and beyond that which is provided by statistical regularity monitoring. Critically, however, the correlational metrics for No-Context Observers were significantly above zero and their SAWs made the same

predictions as Buyers' RAWs in more than 3/4 of choice tasks, suggesting that they were able to extract some meaningful information through statistical regularity monitoring, even when domain-specific beliefs were removed. This again highlights the key finding that some inferences about cognitive processes can be made without access to privileged, metacognitive cues.

Another interesting pattern that arose was that No-Context Observers had less variance in their SAWs, instead putting relatively equal weights on each attribute. Though this slightly reduced their relative accuracy (e.g., the correlational measures)[3], it resulted in lower absolute error (Euclidean Distances) and did not significantly reduce the predictive accuracy of their SAWs relative to Observers who had contextual information. As we explore in the General Discussion, this could be an adaptive strategy in some decision environments.

Critically, however, the conclusions that can be drawn from Studies 1 and 2 were limited by two methodological factors. First, participants were not randomly assigned to the Buyer, Observer, and No-Context Observer conditions. Indeed, multiple months passed between the conditions. While this is to some extent necessary because data must be collected from Buyers first, alternative sampling strategies could be used to maximize random assignment. Second, due to some logistical constraints, Observers and No-Context Observers were randomly assigned to study the choices made by one of 50 Buyers, rather than the entire subset. This partial matching may have led to unexpected artifacts in the data based on the randomly selected sample of Buyers. In Study 3, we replicate Studies 1 and 2, with modifications to overcome these design deficits.

[3] If participants assigned perfectly equal SAWs or AIRs for all seven attributes, the tRAW-tSAW and tRAW-AIR correlations would fall to a mean of zero (assuming noise) or be incalculable (no noise). We demonstrate this via simulations in the Supplemental Materials.

Study 3

Methods

Overview

In Study 3, we conducted a replication of Studies 1 and 2. To increase the internal validity of the replication and reduce sources of error variance, we made several modifications to the research design. First, we implemented a wave-based recruitment procedure that allowed us to randomly assign all but the first and last wave of participants to one of the three conditions. Second, we retained all Buyers as potential matches for the Observers and No-Context Observers to fully balance our design. Third, data was collected over the course of just 9 days, limiting concerns about recruitment delays. Study 3's analyses were largely the same as the prior studies and were pre-registered, unless otherwise noted: https://aspredicted.org/un8hi6.pdf. This study complies with APA ethical standards for the treatment of human participants and was approved by the Carnegie Mellon University IRB (STUDY2023_00000133).

Participants

To maximize random assignment, we recruited participants using a wave-based procedure. In Wave 1, we recruited 20 participants, all of whom were assigned to the Buyer condition. In Waves 2-11, we recruited 60 participants and randomly assigned them to act as Buyers, Observers, or No-Context Observers. In Wave 12, we recruited 40 participants and randomly assigned them to the Observer or No-Context Observer conditions. Only participants in Waves 1 and 12 did not experience full random assignment.

To match the sample size from the prior studies, we aimed to recruit 220 American Prolific participants per condition, for a total of 660 participants. Prolific recruited two more

participants than we requested. As a result, we ended up with 221 Buyers, 221 Observers, and 220 No-Context Observers. All participants passed at least 3 out of 5 attention checks (84.4% passed all 5), so, following our pre-registration, no participants were excluded from analyses. Across conditions, 487 participants (73.6%) identified as White, 67 (10.1%) as Black, 31 (4.7%) as Asian, 17 (2.6%) as Hispanic/Latino, two (0.3%) as Native American, four (0.6%) as an ethnicity that was not listed, and 47 (7.1%) as more than one race or ethnicity. Seven participants (1.1%) did not provide race or ethnicity information. The sample included 354 (53.5%) females, 292 (44.1%) males, 13 (2.0%) non-binary participants, and one participant (0.2%) who identified as another gender. Two participants (0.3%) did not report their gender. The average age of our sample was 43.0 years (sd = 13.4) and 406 (61.3%) of the participants had a college degree. Participants were paid $2.00. Condition-level demographic information is reported in the Supplemental Materials.

Procedure.

The procedures for Buyers were identical to the procedures for the Buyer condition in Study 1. The procedures for Observers and No-Context Observers were identical to Studies 1 and 2, respectively, except for two notable differences. First, Observers and No-Context Observers were randomly yoked to one of the 20 Buyers from the previous wave (e.g., Wave 2 Observers were yoked to a Wave 1 Buyer). Over the course of the entire study, this meant that the full sample of Buyers was able to be used for yoking, rather than a random subset of 50 Buyers. Second, we reduced the maximum amount of time that Observers and No-Context Observers had to study each choice from 60 seconds to 40 seconds. This maximum was reduced to better reflect the range of time that most Buyers spent on the choice task, falling between the 90th (36s) and 95th (42s) percentile of choice time in Study 1. We also introduced three new questions asking

participants about their subjective experiences with the task. Specifically, we asked participants how *motivated* they were to give accurate stated weights (SAWs and AIRs), how *difficult* they found it to give accurate stated weights, and how much *effort* they put into giving accurate stated weights.

Materials.

We randomly generated a new set of 300 CBC surveys for the Buyers, using the same attributes, levels, and settings as in Study 1. This re-randomization would not be expected to affect the results but offers a robustness check. Screenshots for the Observers and No-Context Observers were generated in the same way as in Studies 1 and 2, except that we labeled the alternatives the Buyer was evaluating as Option 1, Option, 2, and Option 3. Then, instead of using a blue checkmark to indicate the selected alternative, we added text above each screenshot stating: "Previous Participant's Selection: Option [1/2/3]." This change was necessary to implement the wave-based procedure in a reasonable timeframe, as it allowed us to prepare the screenshots – a labor-intensive process – prior to data collection.

Analysis & Results

Estimating Revealed Attribute Weights

Buyers' RAWs were estimated using the same procedures as in Study 1. To test the success of our yoking procedure – while accounting for the fact that some participants left the study after being yoked to a Buyer, resulting in imperfect randomization – we compared Buyers' mean RAWs to the mean RAWs of the Buyers to whom Observers and No-Context Observers were yoked. One-way ANOVAs indicated no significant differences across conditions for any attribute, Fs(2, 659) $\leq$ 0.22, ps > .80. Mean RAWs for each condition are reported in Table 3.

Table 3

Attribute	Buyers' Mean RAW (sd)	Observers' Mean RAW (sd)	No-Context Observers' Mean RAW (sd)
Commute Time (Alpha Score)	18.41 (8.20)	18.11 (8.34)	18.19 (8.13)
Home Size (Beta Score)	11.19 (5.13)	11.16 (5.17)	11.26 (5.08)
Mortgage (Gamma Score)	18.38 (10.13)	18.33 (9.97)	18.32 (10.15)
School Quality (Delta Score)	17.81 (11.54)	17.94 (11.44)	17.93 (11.77)
Attractiveness (Epsilon Score)	12.80 (7.49)	12.81 (7.61)	12.98 (7.65)
Lot Size (Zeta Score)	9.10 (4.11)	9.33 (4.02)	9.13 (3.98)
Parking (Omega Score)	12.32 (7.40)	12.33 (7.56)	12.20 (6.87)

Mean RAWs for Observers and No-Context Observers reflect the mean of the RAWs from the Buyers to whom the participants were yoked. Mean RAWs are not identical across conditions because of noise in the random assignment procedure (e.g., participants who started the study, got yoked to a Buyer, then didn't finish the study).

Descriptives.

Mean SAWs and AIRs for each condition are reported in Table 4.

Validating RAWs

Using the same methods as in Study 1, we found that Buyers' RAWs correctly predicted the choice they would make for 94.9% (95% CI = 94.0 – 95.6%) of choice tasks, suggesting again that our RAW estimation procedure was effective.

Data Transformation.

As in Studies 1 and 2, we applied a centered log-ratio transformation to RAWs and SAWs to account for their compositional structure (Cash & Oppenheimer, 2025a; Smithson & Broomell, 2024). We again refer to the transformed values as tRAWs and tSAWs.

Table 4

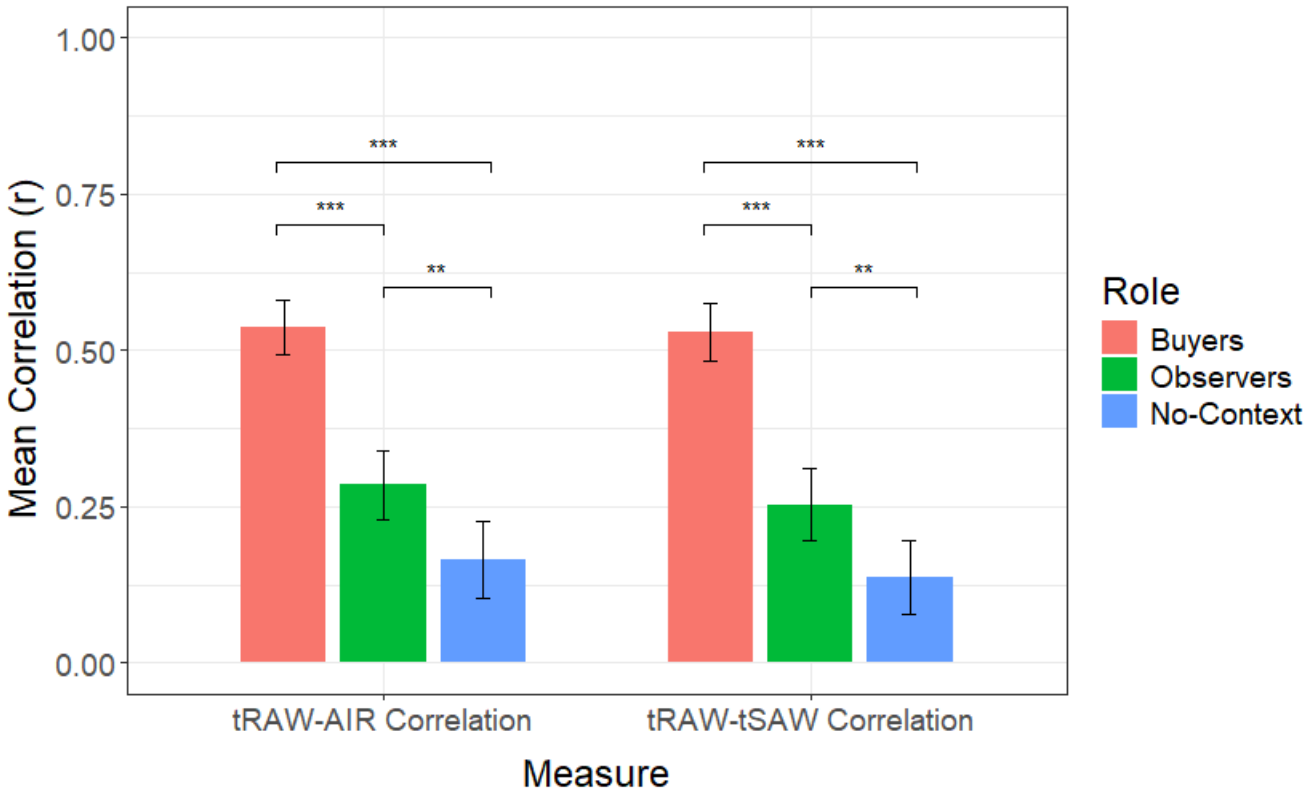
Attribute	Buyers	Observers	No-Context
Mean SAW (sd)			
Commute Time (Alpha Score)	18.57 (14.81)	15.72 (17.03)	15.02 (11.39)
Home Size (Beta Score)	12.85 (11.66)	14.77 (10.61)	14.69 (10.83)
Mortgage (Gamma Score)	22.02 (16.33)	17.96 (16.13)	14.11 (13.60)
School Quality (Delta Score)	14.19 (18.24)	14.31 (17.84)	15.85 (15.43)
Attractiveness (Epsilon Score)	10.86 (11.98)	14.74 (16.39)	14.46 (12.52)
Lot Size (Zeta Score)	10.13 (9.94)	12.51 (10.86)	12.28 (9.25)
Parking (Omega Score)	11.37 (11.56)	9.99 (10.63)	13.58 (11.41)
Mean AIR (sd)			
Commute Time (Alpha Score)	6.19 (2.32)	5.82 (2.37)	5.71 (2.10)
Home Size (Beta Score)	5.53 (2.24)	5.93 (2.10)	5.71 (2.24)
Mortgage (Gamma Score)	6.80 (2.10)	6.16 (2.30)	5.49 (2.14)
School Quality (Delta Score)	4.56 (2.98)	5.23 (2.61)	5.63 (2.14)
Attractiveness (Epsilon Score)	4.81 (2.46)	5.53 (2.25)	5.39 (2.13)
Lot Size (Zeta Score)	4.97 (2.45)	5.41 (2.18)	4.96 (2.21)
Parking (Omega Score)	5.00 (2.41)	4.97 (2.41)	5.41 (2.23)

Correlations between tRAWs, tSAWs, and AIRs

We calculated the participant-level tRAW-tSAW and tRAW-AIR correlations in the same way as in Studies 1 and 2. The mean tRAW-tSAW correlation was $r = 0.53$ ($sd = .35$) for Buyers, $r = .25$ ($sd = .44$) for Observers, and $r = .14$ ($sd = .44$) for No-Context Observers. A one-way ANOVA indicated that mean participant-level correlations were significantly different across conditions, $F(2, 659) = 53.36$, $p < .001$. Pairwise comparisons made using Tukey's HSDs indicated that Buyers' mean participant-level correlation was significantly higher than that of both Observers ($p < .001$) and No-Context Observers ($p < .001$). Observers' mean correlation was significantly higher than that of No-Context Observers ($p = .008$). These results are summarized in Figure 5.

The mean participant-level tRAW-AIR correlation was $r = 0.54$ ($sd = .33$) for Buyers, $r = .29$ ($sd = .41$) for Observers, and $r = .17$ ($sd = .45$) for No-Context Observers. A one-way ANOVA indicated that mean participant-level correlations were significantly different across conditions, $F(2, 636) = 49.10$, $p < .001$. Pairwise comparisons using Tukey's HSDs indicated that Buyers' mean participant-level correlation was significantly higher than that of both Observers ($p < .001$) and No-Context Observers ($p < .001$). Observers' mean correlation was significantly higher than No-Context Observers' mean correlation ($p = .006$). These results, summarized in Figure 5, largely replicate the findings from Studies 1 and 2, except that the difference between Observers' and No-Context Observers' mean tRAW-AIR correlations was significant in Study 3, whereas it was only marginally significant in Study 2. This shift from marginal significance to significance may be attributable to the decrease in error variance resulting from the implementation of random assignment in Study 3. Comparisons of attribute-level correlations (which were pre-registered) are reported in the Supplemental Materials.

Figure 5. tRAW-tSAW and tRAW-AIR Correlations, by Condition



***p < .01, ***p < .001.*

Euclidean Distances

We calculated Euclidean Distances between tRAWs and tSAWs in the same way as in Studies 1 and 2. The mean Euclidean Distance was 0.33 (*sd* = .12) for Buyers, 0.37 (*sd* = .15) for Observers, and 0.33 (*sd* = .14) for No-Context Observers. A one-way ANOVA indicated that these means were significantly different from one another, $F(2, 659) = 6.82$, $p = .001$. Pairwise comparisons using Tukey's HSDs indicated that Observers had a significantly higher mean Euclidean Distance than both Buyers ($p = .004$) and No-Context Observers ($p = .004$). These results, summarized in Table 6, are directionally similar to the results from the prior studies, but the magnitude of the differences failed to replicate in two ways. First, the difference between

Buyers and Observers rose from non-significant in Study 1 to significant in Study 3. Second, the difference between Buyers and No-Context Observers, which was significant in Study 2, became non-significant in Study 3. This suggests some unreliability in condition-level differences on the Euclidean Distance metric. However, one finding that was consistent across studies was the surprising finding that No-Context Observers significantly outperformed Observers. Results were qualitatively similar when we replicated these results with City Block Distance instead of Euclidean Distance (see Supplemental Materials).

To further investigate our hypothesis that this difference occurs because No-Context Observers exhibit less variance in their weights than Observers, we again calculated the percentage of SAWs between 10-20% reported by participants in each condition. For No-Context Observers, 56.6% of SAWs fell between 10-20%. For Observers, this percentage was 45.2% and for Buyers it was 39.2%. This again suggests that No-Context Observers achieved a lower Euclidean Distance than Observers not by better understanding the choice behavior of the Buyer to whom they were yoked, but instead by treating weights as being approximately evenly distributed across attributes. This analysis was pre-registered for this study, thus lending more confidence to its interpretation.

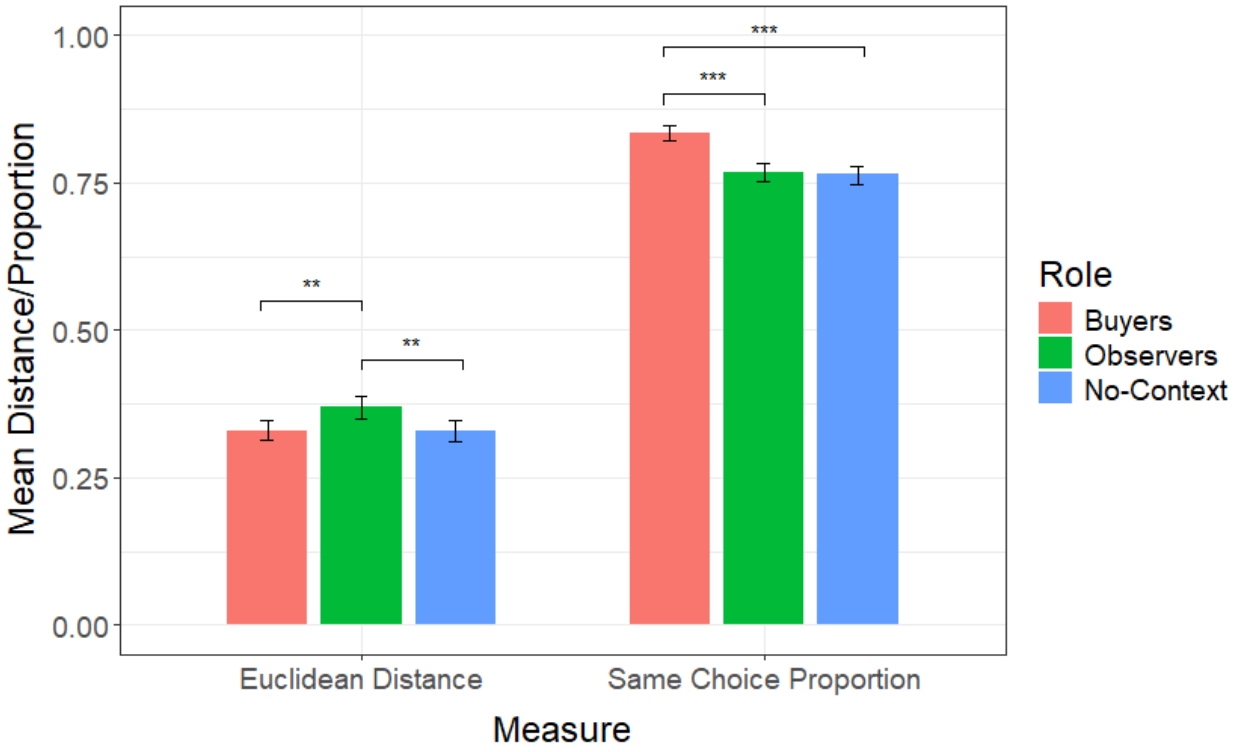
RAW and SAW Choice Predictions

Using the same methods as in the previous studies, we next evaluated the percentage of choices tasks for which RAWs (estimated from Buyers) and SAWs made the same choice prediction. Buyers' SAWs predicted the same alternative as their RAWs in 83.5% of choice tasks (95% CI = 82.1% - 84.8%). Observers' SAWs predicted the same alternative as the Buyer's RAWs in 76.7% of choice tasks (95% CI = 75.2% - 78.2%). For No-Context Observers, this percentage was 76.4% (95% CI = 74.8% - 77.9%). Chi-squared tests of equal proportions

indicated that the percentage for Buyers was significantly higher than the percentage for Observers, $X^2(1, N = 6{,}059) = 24.42$, $p < .001$, and No-Context Observers, $X^2(1, N = 6{,}020) = 27.70$, $p < .001$. The percentages for Observers and No-Context Observers were not significantly different from one another, $X^2(1, N = 6{,}045) = 0.09$, $p = .76$. These results are summarized in Figure 6.

As a robustness check, we ran a logistic multi-level model (choice tasks nested within individuals) predicting the likelihood of RAWs and SAWs making different predictions. The model included two planned contrasts. The first contrast compared the Buyer condition to the mean of the Observer and No-Context Observer conditions (coded as Buyers = 1, Observers = -0.5, No-Context Observers = -0.5). This contrast indicated that the likelihood of RAWs and SAWs making different predictions was significantly lower than the mean likelihood for Observers and No-Context Observers ($B = -0.30$, $SE = 0.05$, OR = .74, $p < .001$). The second contrast compared Observers to No-Context Observers (coded as Buyers = 0, Observers = -1, No-Context Observers = 1). This contrast indicated that there was no significant difference in the likelihood of RAWs and SAWs making different predictions when comparing Observers and No-Context Observers ($B = -0.01$, $SE = 0.04$, OR = .99, $p = .79$). The full model is reported in the Supplemental Materials. Replicating the prior studies, this pattern of results suggests that Observers and No-Context Observers' lack of access to privileged metacognitive cues reduces the accuracy of their inferences relative to Buyers, and that access to domain-specific beliefs does not improve the absolute accuracy of Observers relative to No-Context Observers. As an additional robustness check, we also demonstrate in the Supplemental Materials that this pattern of results holds when comparing SAW predictions to the choices Buyers actually made, rather than the choices predicted by their RAWs.

Figure 6. Mean Euclidean Distance and Same Choice Proportion, By Condition



***p < .01, ***p < .001.*

Confidence

We next evaluated participants' confidence in the accuracy of their SAWs and AIRs, reported on a 0-100 scale. Mean confidence for SAWs was 73.58 (*sd* = 21.43) for Buyers, 57.61 (*sd* = 23.02) for Observers, and 45.70 (*sd* = 26.06) for No-Context Observers. A one-way ANOVA indicated that these means were significantly different from one another, $F(2, 659) = 77.63$, $p < .001$. Pairwise comparisons using Tukey's HSDs indicated that Buyers were significantly more confident than both Observers ($p < .001$) and No-Context Observers ($p < .001$) and that Observers were more confident than No-Context Observers ($p < .001$). Mean confidence for AIRs was 76.51 (*sd* = 19.00) for Buyers, 61.54 (*sd* = 22.61) for Observers, and

47.66 (*sd* = 26.44) for No-Context Observers. The difference between means was significant, $F(2,659) = 87.70$, $p < .001$ and, as with SAWs, Buyers were more confident than both Observers ($p < .001$) and No-Context Observers ($p < .001$), while Observers were more confident than No-Context Observers ($p < .001$). These results replicate the earlier finding that Buyers were more confident than both sets of Observers but diverge from the prior studies in that No-Context Observers were less confident than Observers, whereas the No-Context Observers in Study 2 were equally as confident as Observers. Given that the No-Context Observers had less information than Observers, and that their metacognitive knowledge was generally (though not consistently significantly) less accurate, their lower confidence seems reasonable. As such, it is possible that this effect that was obscured in Studies 1 and 2 by the greater error variance resulting from non-random assignment to conditions.

Subjective Experience

In addition to confidence, we investigated other measures by which participants' subjective experiences may have differed across the three conditions. First, we asked participants to report how difficult they found it to generate accurate weights, on a scale of 0 (Not Difficult At All) to 100 (Extremely Difficult). The average difficulty rating was 45.90 (*sd* = 29.69) for Buyers, 63.79 for Observers (*sd* = 28.09), and 74.76 (*sd* = 24.92) for No-Context Observers. These means were significantly different, $F(2, 659) = 61.26$, $p < .001$, and Tukey's HSDs indicated that all pairwise differences were significant (ps $< .001$). Difficulty was not consistently correlated with any measure of absolute or relative accuracy (see Supplemental Materials).

We next asked participants how much effort they put into generating accurate weights, on a scale of 0 (No Effort) to 100 (A Great Deal of Effort). The mean effort value was 88.49 (*sd* =

13.69) for Buyers, 88.54 (*sd* = 13.78) for Observers, and 83.41 (*sd* = 19.43) for No-Context Observers. These means were significantly different from one another, $F(2.\ 659) = 7.61$, $p = .001$. Tukey's HSDs indicated that Buyers and Observers put in an equal amount of effort ($p = 1.00$), and that both put in more effort than No-Context Observers (ps = .002). Effort was not consistently correlated with participant-level tRAW-tSAW or tRAW-AIR correlations, but was weakly, positively correlated with Euclidean Distance (rs $\leq .20$; see Supplemental Materials).

Lastly, we asked participants how motivated they felt to generate accurate weights, on a scale of 0 (Not Motivated At All) to 100 (Extremely Motivated). The mean motivation value was 88.07 (*sd* = 15.13) for Buyers, 86.57 (*sd* = 16.33) for Observers, and 79.61 (*sd* = 22.85) for No-Context Observers. The means were significantly different from one another, $F(2, 659) = 13.24$, $p < .001$. Tukey's HSDs indicated that Buyers and Observers were not significantly different in their level of motivation ($p = .67$), but that both were more motivated than No-Context Observers (ps < .001). Motivation was not consistently correlated with any measure of absolute or relative accuracy (see Supplemental Materials)

In conjunction, these results suggest that No-Context Observers found the task to be substantially more difficult than Buyers and Observers, and that this higher perceived difficulty was associated with reduced motivation and lower effort. However, effort and motivation were still quite high in the No-Context Observer condition, and the between-condition differences in effort and motivation were substantially smaller than the differences in perceived difficulty, suggesting that the higher difficulty did not lead No-Context Observers to give up on the task completely. Importantly, though, these measures were self-reported, and participants may have been loath to report low effort or motivation, for fear of losing out on their pay. This could be addressed in future studies by directly informing participants that their renumeration will not be

affected by their responses to the motivation and effort questions. Furthermore, causal connections cannot be drawn, as motivation, effort, and difficulty were not manipulated.

Study 3 Discussion

In all, Study 3 provides a convincing replication of the results of Studies 1 and 2, with methodological procedures that maximized internal validity and reduced sources of error variance that limited the strength of conclusions that could be made based on the initial studies. As in the prior studies, Study 3 demonstrated that Observers and No-Context Observers made less accurate inferences about Buyers' decision weights than Buyers did about their own decision weights, likely due to their lack of access to privileged metacognitive cues. However, Observers and No-Context Observers were not totally incapable of generating meaningful inferences; Observers' and No-Context Observers' SAWs and AIRs were significantly (albeit weakly) correlated with Buyers' RAWs, and their SAWs made the same choice predictions as Buyers' RAWs more than 75% of the time. We also replicated our earlier finding that Observers' inferences were more accurate than No-Context Observers' inferences, but only when assessed using measures of relative accuracy. This suggests that domain-specific beliefs may improve relative metacognitive accuracy, but that statistical regularity monitoring alone is sufficient to achieve a given level of absolute accuracy.

Study 3 also probed the subjective experiences of the participants in each of the three conditions to identify differences in how they felt about and approached the task. Buyers found the task to be easiest, No-Context Observers found the task to be hardest, and Observers were in the middle. This is unsurprising, as it directly reflects the amount of information available to participants in each condition, with the task feeling easier when more information is available. More interestingly, however, we did not find the same pattern of results when measuring effort or

motivation. For both measures, we found that Buyers and Observers were not significantly different from one another, whereas No-Context Observers were less motivated and put in less effort. These divergent findings are explored in the General Discussion.

General Discussion

The question of the extent to which humans have and use privileged insights into the inner workings of their own mind has been debated for decades. (Bem, 1972; Byrne, 2005; Nisbett & Wilson, 1977; Smithies & Stoljar, 2012; Vermersch, 2009). Most modern metacognitive scholars suggest that people do not have direct introspective access to their cognitive and neural processes, but rather that they generate metacognitive judgments by evaluating a wide variety of metacognitive cues (Ackerman & Thompson, 2017; Koriat, 1997). One implication of this cue-utilization approach is that some of these cues could be observed by others to make inferences about your cognitive processes. To this end, the metamemory literature suggests that observers can, in fact, make reasonably accurate inferences about others' learning processes through the evaluation of observable cues (Koriat & Ackerman, 2010; Serra & Ariel, 2014; Undorf & Erdfelder, 2011; Vesonder & Voss, 1985; Wei et al., 2026). However, little is known about the ability of observers to make inferences about others' cognitive experiences outside of the domain of memory and learning. Here, we expand this literature to the reasoning domain of multi-attribute choice.

In three studies, we investigated the ability of both decision makers and observers to make inferences about the weight that decision makers placed on various attributes in a multi-attribute choice task. In Study 1, we found that decision makers could more accurately estimate the weight they placed on various attributes than an observer watching their choices. Unsurprisingly, this indicated that privileged metacognitive cues improve the accuracy of

metacognitive knowledge. These results align with recent findings from Morris et al. (2025), which similarly demonstrate that decision makers can more accurately estimate their own decision weights than fully informed Observers. However, we also demonstrated that observers were able to generate estimates of the decision makers' attribute weights that were correlated with their true weights. This suggests that observers could use observable cues to make reasonably accurate inferences about a decision maker's underlying cognitive processes, even if these inferences were not quite as accurate as decision makers' own metacognitive knowledge.

In Study 2, we built upon these results by probing the metacognitive mechanisms by which observers make these inferences. Specifically, we eliminated observers' access to contextual information so that they could not use domain-specific beliefs to inform their inferences (Aydogan, 2021; Heiphetz et al., 2013). Thus, the No-Context Observers could only infer the decision makers' attribute weights through statistical regularity monitoring processes. Interestingly, we found that removing observers' access to domain-specific beliefs reduced the relative accuracy of their inferred weights, but either had no effect on their absolute accuracy, or in some cases improved their absolute accuracy. This suggests that domain-specific beliefs only increase the relative accuracy of inferences about others' decision-making processes, not absolute accuracy. However, Studies 1 and 2 were limited in that participants were not randomly assigned to conditions, thus introducing threats to the internal validity of the results.

In Study 3, we replicated Studies 1 and 2 with improved methodological procedures, including random assignment, designed to maximize internal validity and reduce sources of error variance. The results were highly similar to the prior studies, again indicating that Buyers had the most accurate metacognitive knowledge and that Observers' inferences were only more accurate than those of No-Context Observers when assessed using measures of relative accuracy. We also

probed participants' subjective experiences during the task and discovered that No-Context Observers found the task more difficult than Observers, who in turn found the task more difficult than Buyers. However, Observers and Buyers reported equal levels of motivation and effort, both of which were reduced among No-Context Observers. This may suggest that the task was too difficult for No-Context Observers, leading to a decrease in motivation and effort (Brehm & Self, 1989). The decrease in effort and motivation could also be a function of the No-Context Observers' task feeling less realistic or relevant, given that we would rarely observe someone making a decision and have no idea what they were choosing between. One possible alternative account for the results is that No-Context Observers' metacognitive knowledge may have been more accurate if they had felt more motivated and/or put in more effort. However, we find this account to be unlikely because even the most motivated No-Context Observer would still lack the domain-specific cues that were available to Observers and the privileged metacognitive information that was accessible to Buyers. Furthermore, we did not see a strong relationship between motivation or effort and accuracy in the Study 3 data (see Supplemental Materials).

Absolute vs. Relative Accuracy

One interesting finding from Studies 2 and 3 was that No-Context Observers demonstrated less variance in their SAWs than Observers, suggesting that they treated different Buyers as having more similar attribute weights to one another. On one hand, doing so reduced relative accuracy, resulting in lower correlations between RAWs and SAWs. In other words, No-Context Observers did not as effectively distinguish which attributes were more or less important to their specific Buyer. However, this behavior simultaneously increased the absolute accuracy of their attribute weights, as demonstrated by lower Euclidean Distances, and had no negative effect

on the extent to which No-Context Observers' SAWs were predictive of Buyers' choice behavior. So, this begs the question: are relative and absolute accuracy equally important?

Like most psychological questions, the answer is: it depends. If you are making a decision in which preferences are strong and you are motivated to make the best possible decision – such as buying the perfect engagement ring for your partner – then relative accuracy is quite important. Your partner likely would not be thrilled if they were a diamond boycotter who preferred sapphires, but you bought a diamond ring and explained that it was the most popular and common choice. However, if you are making a decision in an environment where information is scarce and others' preferences are relatively similar – such as giving generic gifts to others you do not know well (Steffel & Le Boeuf, 2014) – heuristics like equal-weighting may be adaptive (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1978). In Studies 2 and 3, it would appear that, given the lack of information available, No-Context Observers did not make strong inferences about their specific Buyer, and that doing so paid off in terms of absolute accuracy, which could be interpreted as a rational and adaptive strategy. This also aligns with previous research suggesting that access to information can reduce task accuracy if that information is used sub-optimally (Hall et al., 2007).

Critically, however, applying relatively equal weights to all attributes is only a valid strategy in domains where all attributes are relatively equally important. If the choice task included both important (e.g., mortgage) and unimportant (e.g., front door color) attributes, then applying equal weights is likely to reduce both relative and absolute accuracy. Similarly, if Buyers had highly heterogeneous preferences, then equal weighting should also lead to reductions in both types of accuracy. It is also plausible, however, that No-Context Observers would move away from the equal-weighting strategy if they realized that their Buyer had strong

enough preferences. This would depend on whether the No-Context Observers in our studies intentionally and strategically implemented the low-variance strategy, or if it was simply a signal that they could not glean much from the choices they observed. We strongly encourage future researchers to explore unequal weighting environments, heterogeneity, and strategic reasoning as potential moderators of the absolute and relative accuracy of No-Context Observers' inferences.

Limitations and Future Directions

One limitation of the studies presented here is that we only isolated statistical regularity monitoring, without isolating domain-specific beliefs. As such, it is currently unclear how a group of "No-Monitoring Observers" who made inferences based only on domain-specific beliefs, without observing the choices made by the Buyers, would perform relative to Observers and No-Context Observers. This could be investigated in future studies by showing participants the choices that a Buyer faced, without telling them which alternative the Buyer selected. The No-Monitoring Observer would then guess the weights that they believe the Buyer would place on each of the homebuying attributes and these estimates could then be compared to the Buyer's RAWs, and any correlations between the two would be entirely the result of domain-specific beliefs. Alternatively, No-Monitoring Observers could be asked to estimate the weight that a "typical" Buyer would place on each attribute without being shown any choices. Their weights could then be randomly paired with Buyers' weights (e.g., via bootstrapping) and correlations could be calculated. We believe that this presents a fruitful direction for further isolating the cues that Observers use to infer others' decision weights.

A related limitation of the studies presented here is that the only cues available to the Observers were the choice task itself and contextual information. In many real-world scenarios, an Observer may have access to other cues as well. For example, the Observer could be told how

long the Buyer spent on each choice (Mata & Almeida, 2014; Undorf & Erdfelder, 2011) or could watch the Buyer and see how stressed they looked while making each choice. Similarly, the Observer could be given idiosyncratic information about the Buyer (Tullis & Feder, 2023; Tullis & Fraundorf, 2017), such as their demographic characteristics, or, in an extreme case, information on the decision maker's explicit beliefs and preferences. For example, the Buyer could write an essay about what they value in a home or share pictures of the homes they've lived in, thus minimizing the information that is privileged to the Buyer. It is reasonable to suspect that providing these additional cues to Observers would reduce the gap between Buyers and Observers, but empirical work is needed to fully explore this possibility.

Another limitation of the studies presented here is that we cannot fully distinguish whether our metacognitive knowledge metrics captured the ability of participants to generate accurate knowledge about how Buyers made their choices, or their ability to remember this knowledge after they generated it. As such, it is plausible that Buyers, Observers, and No-Context Observers were equally able to generate accurate estimates of the weights that Buyers placed on each attribute, but that Observers and No-Context Observers simply forgot this information more quickly. This explanation would align with prior research in the memory literature suggesting that people better remember information about themselves (Symons & Johnson, 1997) and are less likely to forget actions they took themselves, as compared to actions taken by observers (Manzi & Nigro, 2008). Though we sought to minimize differential forgetting across conditions by eliciting SAWs and AIRs immediately after participants completed the observation or choice task, future research should seek to explicitly isolate the ability of participants to generate attribute weights from their ability to remember the weights, perhaps by manipulating the length of the delay before weight elicitation.

Another limitation of the studies presented here is that we only explored choices in the context of homebuying, a domain where decision makers likely have strong and established preferences, beliefs, and values. Future researchers should seek to replicate these results in other multi-attribute choice domains, such as choosing between complex medical procedures, buying unfamiliar products, or making low-stakes decisions, such as where to go for lunch. Furthermore, participants made only hypothetical decisions and thus may not have been as motivated to engage with the study as they would have been if the choices had real-world implications. Despite this, participants in Study 3 self-reported relatively high motivation across all three conditions. We also sought to limit this concern by incentivizing accuracy with monetary bonuses. However, future research exploring the accuracy of metacognitive knowledge in incentive-compatible choices would be worthwhile.

Conclusions

The studies presented here highlight the importance of privileged metacognitive cues for the generation of accurate metacognitive knowledge in the context of multi-attribute choice. However, they also demonstrate that inferences about cognitive processes can be reasonably generated by observers using non-privileged processes, such as statistical regularity monitoring (Perkovic & Orquin, 2018; Ruffman et al., 2012; Sherman et al., 2020) and domain-specific beliefs (Aydogan, 2021; Heiphetz et al., 2013; Lee et al., 2018). We further investigated these non-privileged processes and found that observers can make significant inferences through statistical regularity monitoring alone, but that access to domain-specific beliefs increased the relative accuracy of their inferences. Interestingly, however, removing access to domain-specific beliefs led observers to generate more conservative estimates of Buyers' attribute weights, which in some cases resulted in greater accuracy on an absolute scale, suggesting that additional

contextual information, if used sub-optimally (Hall et al., 2007), may actually reduce the accuracy of inferences. These findings highlight critical distinctions that should inform future investigations into the role that non-privileged, observable cues play in the generation of metacognitive knowledge. They also provide additional support for the notion that cue-utilization theory applies to reasoning processes (Ackerman, 2019; Ackerman & Thompson, 2017; Koriat, 1997) and extend the expansive metacognitive literature on observer inferences about others' cognitive processes (Tullis & Fraundorf, 2017; Undorf et al., 2022; Vesonder & Voss, 1985; Wei et al., 2026) to the domain of multi-attribute choice.

Acknowledgements

This work was supported by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (Award #2333553), and an Academic Research Grant from Sawtooth Software, Inc. that provided us with free access to their Lighthouse Studio CBC software. Trent N. Cash was also supported by the Lupina foundation. No grantor had any role other than financial support. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the grantors. The authors would like to recognize the support of two undergraduates – Piper Ligotti and Mira Devgan – who assisted with stimulus design. The authors would also like to thank their wives, Elizabeth and Nicole, for their support and feedback.

References

Ackerman, R. (2019). Heuristic Cues for Meta-Reasoning Judgments: Review and Methodology. *Psihologijske Teme*, *28*(1), 1–20. https://doi.org/10.31820/pt.28.1.1

Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review*, *19*(6), 1187–1192. https://doi.org/10.3758/s13423-012-0305-z

Allwood, C. M. (1994). Confidence in own and others' knowledge. *Scandinavian Journal of Psychology*, *35*(3), 198–211. https://doi.org/10.1111/j.1467-9450.1994.tb00944.x

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Alter, A. L., Oppenheimer, D. M., & Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: Response to Thompson et al. *Cognition*, *128*(2), 252–255. https://doi.org/10.1016/j.cognition.2013.01.006

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569–576. https://doi.org/10.1037/0096-3445.136.4.569

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA

Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Aydogan, I. (2021). Prior Beliefs and Ambiguity Attitudes in Decision from Experience. *Management Science*, *67*(11), 6934–6945. https://doi.org/10.1287/mnsc.2020.3841

Bainbridge, L. (1979). Verbal reports as evidence of the process operator's knowledge. *International Journal of Man-Machine Studies*, *11*(4), 411–436. https://doi.org/10.1016/S0020-7373(79)80035-8

Bem, D. J. (1972). Self-Perception Theory. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 6, pp. 1–62). Elsevier. https://doi.org/10.1016/S0065-2601(08)60024-6

Boring, E. G. (1953). A history of introspection. *Psychological Bulletin*, *50*(3), 169–189. https://doi.org/10.1037/h0090793

Brehm, J. W., & Self, E. A. (1989). The Intensity of Motivation. *Annual Review of Psychology*, *40*(1), 109–131. https://doi.org/10.1146/annurev.ps.40.020189.000545

Byrne, A. (2005). Introspection. *Philosophical Topics*, *33*(1), 79–104.

Cash, T. N., & Oppenheimer, D. M. (2026). *Exploring sources of metacognitive knowledge in multi-attribute choice.* Open Science Framework. https://osf.io/t8sbd/?view_only=87691388e51a4291ac1feb65f8f97fb6

Cash, T. N., & Oppenheimer, D. M. (2024). Parental rights or parental wrongs: Parents' metacognitive knowledge of the factors that influence their school choice decisions. *PLOS ONE*, *19*(4), e0301768. https://doi.org/10.1371/journal.pone.0301768

Cash, T. N., & Oppenheimer, D. M. (2025a). Assessing metacognitive knowledge in subjective decisions: The *knowledge of weights* paradigm. *Thinking & Reasoning*, *31*(3), 331–373. https://doi.org/10.1080/13546783.2024.2426543

Cash, T. N., & Oppenheimer, D. M. (2025b). *Worth the Weight: Modern LLMs Demonstrate Accurate Metacognitive Knowledge of Decision Weights in Multi-Attribute Choice*. PsyArXiv. https://doi.org/10.31234/osf.io/zq6md_v1

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429–437. https://doi.org/10.3758/MC.36.2.429

Chaxel, A.-S. (2015). How Do Stereotypes Influence Choice? *Psychological Science*, *26*(5), 641–645. https://doi.org/10.1177/0956797615569354

Craig, M. A., Zou, L. X., Bai, H., & Lee, M. M. (2022). Stereotypes About Political Attitudes and Coalitions Among U.S. Racial Groups: Implications for Strategic Political Decision-Making. *Personality and Social Psychology Bulletin*, *48*(9), 1349–1366. https://doi.org/10.1177/01461672211037134

Cruz, S. M. (2017). The relationships of political ideology and party affiliation with environmental concern: A meta-analysis. *Journal of Environmental Psychology*, *53*, 81–91. https://doi.org/10.1016/j.jenvp.2017.06.010

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106. https://doi.org/10.1037/h0037613

Descartes, R. (1641). *Meditations on first philosophy*. Hackett Publishing Company.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage.

Eggers, F., Sattler, H., Teichert, T., & Völckner, F. (2022). Choice-Based Conjoint Analysis. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of Market Research* (pp. 781–819). Springer International Publishing. https://doi.org/10.1007/978-3-319-57413-4_23

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*(5), 395–416. https://doi.org/10.1037/0033-295X.85.5.395

Engelbert, M., & Carruthers, P. (2010). Introspection. *WIREs Cognitive Science*, *1*(2), 245–253. https://doi.org/10.1002/wcs.4

Epley, N., & Gilovich, T. (2006). The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Fabre, E. F., Causse, M., Pesciarelli, F., & Cacciari, C. (2015). Sex and the money – How gender stereotypes modulate economic decision-making: An ERP study. *Neuropsychologia*, *75*, 221–232. https://doi.org/10.1016/j.neuropsychologia.2015.06.013

Fiedler, K., Ackerman, R., & Scarampi, C. (2019). Metacognition: Monitoring and Controlling One's Own Knowledge, Reasoning and Decisions. In R. J. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought* (pp. 89–111). Heidelberg University Publishing. https://doi.org/10.17885/heiup.470

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

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(JULY). https://doi.org/10.3389/fnhum.2014.00443

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910. https://doi.org/10.1016/j.tics.2021.07.008

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, *103*(2), 277–290. https://doi.org/10.1016/j.obhdp.2007.01.003

Halverson, C. (2020). *Sample size rule of thumb for choice-based conjoint (CBC)*. Sawtooth. https://sawtoothsoftware.com/resources/blog/posts/sample-size-rules-of-thumb

Hein, M., Kurz, P., & Steiner, W. J. (2020). Analyzing the capabilities of the HB logit model for choice-based conjoint analysis: A simulation study. *Journal of Business Economics*, *90*(1), 1–36. https://doi.org/10.1007/s11573-019-00927-4

Heiphetz, L., Spelke, E. S., Harris, P. L., & Banaji, M. R. (2013). The development of reasoning about beliefs: Fact, preference, and ideology. *Journal of Experimental Social Psychology*, *49*(3), 559–565. https://doi.org/10.1016/j.jesp.2012.09.005

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22–34. https://doi.org/10.1037/0278-7393.29.1.22

Hoch, E., Sidi, Y., Ackerman, R., Hoogerheide, V., & Scheiter, K. (2023). Comparing Mental Effort, Difficulty, and Confidence Appraisals in Problem-Solving: A Metacognitive

Perspective. *Educational Psychology Review*, *35*(2), 61. https://doi.org/10.1007/s10648-023-09779-5

Hoepfl, R. T., & Huber, G. P. (1970). A study of self-explicated utility models. *Behavioral Science*, *15*(5), 408–414. https://doi.org/10.1002/bs.3830150503

Ioannou, C. A., & Romero, J. (2014). A generalized approach to belief learning in repeated games. *Games and Economic Behavior*, *87*, 178–203. https://doi.org/10.1016/j.geb.2014.05.007

Jordan, J. S. (1995). Bayesian Learning in Repeated Games. *Games and Economic Behavior*, *9*(1), 8–20. https://doi.org/10.1006/game.1995.1002

Kant, I. (1781). *The critique of pure reason*. Cambridge University Press.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, and E. Thompson (Eds.), *The Cambridge Handbook of Consciousness.* Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.012

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Lauterman, T., & Ackerman, R. (2024). Initial judgment of solvability: Integrating prior expectations with experience-based heuristic cues. *Thinking & Reasoning*, *30*(1), 135–168. https://doi.org/10.1080/13546783.2023.2214378

Lee, J. C., Kim, J., & Kwak, K. (2018). A Multi-Attribute Examination of Consumer Conformity in Group-Level Ordering. *Australasian Marketing Journal*, *26*(1), 41–48. https://doi.org/10.1016/j.ausmj.2017.11.004

Manzi, A., & Nigro, G. (2008). Long-term memory for performed and observed actions: Retrieval awareness and source monitoring. *Memory*, *16*(6), 595–603. https://doi.org/10.1080/09658210802070749

Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & Cognition*, *43*(4), 681–693. https://doi.org/10.3758/s13421-014-0488-9

Mata, A., & Almeida, T. (2014). Using metacognitive cues to infer others' thinking. *Judgment and Decision Making*, *9*(4), 349–359. https://doi.org/10.1017/S1930297500006215

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition*, *29*(2), 222–233. https://doi.org/10.3758/BF03194916

Moore, D., & Schatz, D. (2007). The three faces of overconfidence. *Social and Personality Psychology Compass, 11*, e12331. https://doi.org/10.1111/spc3.12331

Morris, A., Carlson, R. W., Kober, H., & Crockett, M. (2025). Introspective access to value-based multi-attribute choice processes. *Nature Communications, 16*, Article 3733. https://doi.org/10.1038/s41467-025-59080-y

Nachbar, J. H. (1997). Prediction, Optimization, and Learning in Repeated Games. *Econometrica*, *65*(2), 275. https://doi.org/10.2307/2171894

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. *Journal of Personality and Social Psychology*, *35*(9), 613–624. https://doi.org/10.1037/0022-3514.35.9.613

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–241. https://doi.org/10.1016/j.tics.2008.02.014

Orme, B. (2002). Interpreting conjoint analysis data. *Sawtooth Software, Inc.* https://leeds-faculty.colorado.edu/ysun/MKTG4825_files/Interpreting_CA_Data.pdf

Perkovic, S., & Orquin, J. L. (2018). Implicit Statistical Learning in Real-World Environments Leads to Ecologically Rational Decision Making. *Psychological Science*, *29*(1), 34–44. https://doi.org/10.1177/0956797617733831

Pronin, E. (2009). The Introspection Illusion. In *Advances in Experimental Social Psychology* (Vol. 41, pp. 1–67). Elsevier. https://doi.org/10.1016/S0065-2601(08)00401-2

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Roebers, C. M., Mayer, B., Steiner, M., Bayard, N. S., & Van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: A longitudinal study. *Developmental Psychology*, *55*(10), 2077–2089. https://doi.org/10.1037/dev0000776

Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology*, *30*(1), 87–104. https://doi.org/10.1111/j.2044-835X.2011.02045.x

Sanchez, C., & Dunning, D. (2020). Decision fluency and overconfidence among beginners. *Decision*, *7*(3), 225–237. https://doi.org/10.1037/dec0000122

Sawtooth Software, Inc. (2021). *CBC/HB system technical paper V5.6: The CBC/HB system for Hierarchical Bayes Estimation*. https://sawtoothsoftware.com/resources/technical-papers/cbc-hb-technical-paper

Sawtooth Software, Inc. (2024). *Lighthouse Studio* (Version 9.14.2) [Computer software]. https://sawtoothsoftware.com/lighthouse-studio

Schwarz, N., Jalbert, M., Noah, T., & Zhang, L. (2021). Metacognitive experiences as information: Processing fluency in consumer judgment and decision making. *Consumer Psychology Review*, *4*(1), 4–25. https://doi.org/10.1002/arcp.1067

Schwitzgebel, E. (2024). Introspection. *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/Entries/introspection/

Serra, M. J., & Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for judgments of learning. *Memory & Cognition*, *42*(8), 1260–1272. https://doi.org/10.3758/s13421-014-0431-0

Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, *32*, 15–20. https://doi.org/10.1016/j.cobeha.2020.01.015

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, *34*(3), 619–632. https://doi.org/10.3758/BF03193584

Slovic, P. (1969). Analyzing the expert judge: A descriptive study of a stockbroker's decision process. *Journal of Applied Psychology*, *53*(4), 255–263. https://doi.org/10.1037/h0027773

Slovic, P., Fleissner, D., & Bauman, W. S. (1972). Analyzing the Use of Information in Investment Decision Making: A Methodological Proposal. *The Journal of Business*, *45*(2), 283. https://doi.org/10.1086/295448

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, *6*(6), 649–744. https://doi.org/10.1016/0030-5073(71)90033-X

Smithies, D., & Stoljar, D. (Eds.). (2012). *Introspection and consciousness*. Oxford University Press.

Smithson, M., & Broomell, S. B. (2024). Compositional data analysis tutorial. *Psychological Methods*, *29*(2), 362–378. https://doi.org/10.1037/met0000464

Steffel, M., & Le Boeuf, R. A. (2014). Overindividuation in Gift Giving: Shopping for Multiple Recipients Leads Givers to Choose Unique but Less Preferred Gifts. *Journal of Consumer Research*, *40*(6), 1167–1180. https://doi.org/10.1086/674199

Susser, J. A., & Mulligan, N. W. (2019). Exploring the intrinsic-extrinsic distinction in prospective metamemory. *Journal of Memory and Language*, *104*, 43–55. https://doi.org/10.1016/j.jml.2018.09.003

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371–394. https://doi.org/10.1037/0033-2909.121.3.371

Thiessen, E. D. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160056. https://doi.org/10.1098/rstb.2016.0056

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for

initiating analytic thinking. *Cognition*, *128*(2), 237–251. https://doi.org/10.1016/j.cognition.2012.09.012

Topolinski, S., Bakhtiari, G., & Erle, T. M. (2016). Can I cut the Gordian tnok? The impact of pronounceability, actual solvability, and length on intuitive problem assessments of anagrams. *Cognition*, *146*, 439–452. https://doi.org/10.1016/j.cognition.2015.10.019

Topolinski, S., & Reber, R. (2010). Immediate truth – Temporal contiguity between a cognitive problem and its solution determines experienced veracity of the solution. *Cognition*, *114*(1), 117–122. https://doi.org/10.1016/j.cognition.2009.09.009

Tullis, J. G., & Feder, B. (2023). The "curse of knowledge" when predicting others' knowledge. *Memory & Cognition*, *51*(5), 1214–1234. https://doi.org/10.3758/s13421-022-01382-3

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124–137. https://doi.org/10.1016/j.jml.2017.03.003

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629–642. https://doi.org/10.1177/1747021819882308

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., Navarro-Báez, S., & Bröder, A. (2022). "You don't know what this means to me" – Uncovering idiosyncratic influences on metamemory judgments. *Cognition*, *222*, 105011. https://doi.org/10.1016/j.cognition.2021.105011

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition*, *46*(4), 507–519. https://doi.org/10.3758/s13421-017-0780-6

Vermersch, P. (2009). Describing the practice of introspection. *Journal of Consciousness Studies*, *16*(10–12), 20–57.

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language*, *24*(3), 363–376. https://doi.org/10.1016/0749-596X(85)90034-8

Wei, Y., Soderstrom, N. C., & Meade, M. L. (2026). Making judgments of learning (JOLs) for oneself versus others: A review and proposed model. *Psychonomic Bulletin & Review*, *33*(1), 14. https://doi.org/10.3758/s13423-025-02816-0

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org/

Wilson, T. D., & Nisbett, R. E. (1978). The Accuracy of Verbal Reports About the Effects of Stimuli on Evaluations and Behavior. *Social Psychology*, *41*(2), 118. https://doi.org/10.2307/3033572

Yang, C., Huang, T. S.-T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language*, *99*, 99–110. https://doi.org/10.1016/j.jml.2017.11.005

Zheng, Z., & Wang, J. (2023). Co-actors represent each other's task regularity through social statistical learning. *Cognition*, *235*, 105411. https://doi.org/10.1016/j.cognition.2023.105411