Assessing Metacognitive Knowledge in Subjective Decisions: The *Knowledge of Weights* Paradigm

Trent N. Cash* and Daniel M. Oppenheimer

Department of Social and Decision Sciences, Carnegie Mellon University, USA

Department of Psychology, Carnegie Mellon University, USA

Keywords: Metareasoning; Metacognitive Knowledge; Choice Based Conjoint; Subjective Decisions; Multiattribute Choice

Author Note

This is an original manuscript of an article published by Taylor & Francis in *Thinking & Reasoning* on November 8th, 2024, available at: https://doi.org/10.1080/13546783.2024.2426543

Data, supplemental materials, and analysis code for this manuscript are available at: https://doi.org/10.17605/OSF.IO/UZQK5. Study 2 [https://osf.io/mw3qg/?view_only=92e94f752db440beb2f563b674495cda] and Study 3 [https://aspredicted.org/V9W_T4V] were preregistered. This work was supported by a Graduate Student Small Grant from Carnegie Mellon University's Center for Behavioral and Decision Research, a DDRIG 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. 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. We have no competing interests to declare**.** We also acknowledge the support of numerous undergraduate research assistants, colleagues, committee members, reviewers, and loved ones.

Abstract

Subjective, multi-attribute choice decisions – such as whom to marry or which college to attend – play a substantial role in decision makers' long-term well-being. However, the metacognition literature lacks tools for assessing metacognitive capacities in subjective decisions. We present three studies in which we propose and validate the *Knowledge of Weights (KoW)* paradigm, a novel method for assessing metacognitive knowledge of attribute weights in subjective, multi-attribute choice decisions. In Study 1, we demonstrate the test-retest reliability of metrics generated by the *KoW* paradigm. In Study 2, we apply the *KoW* paradigm in four domains and show that it generates consistent results. In Study 3, we demonstrate that participants who perform better on the *KoW* paradigm make choices with which they are more satisfied, providing suggestive evidence of predictive validity. Use cases in cognitive psychology and beyond are discussed

Introduction

Imagine that you are at a car dealership, and a salesman asks: "*What are you looking for in a car?*" Your answer to this question critically influences the likelihood that you will end up with a car that you are happy with. For example, if you tell the salesman that your top priority is the ability to go off-roading, but really the most important thing to you is having a fuel-efficient vehicle, you'll be quite unhappy when it costs you $100 to fill up your gas tank each week because you bought a Jeep.

This example highlights a critical element of multi-attribute choice that is often overlooked: a decision maker's explicit knowledge of how strongly various factors influence their choices. Decision makers who have this explicit metacognitive knowledge of their attribute weights can make choices that better align with their preferences and may be able to more accurately communicate their preferences to others, such as the car salesman. Currently, however, the metacognition literature lacks a validated method for assessing participants' knowledge of the weights they place on various factors when making subjective decisions. This paper aims to fill that gap by introducing the *Knowledge of Weights (KoW)* paradigm.

Metacognition

Over the last half-century, a great deal of research has sought to evaluate the ways in which people understand and manipulate their own cognitive processes. This capacity to think about thinking is commonly referred to as metacognition (Brown, 1987; Flavell, 1979). Much of the extant literature in metacognition has focused on memory and learning (e.g., Aleven & Koedinger, 2002; Chua et al., 2009; Dodson et al., 2007; Hu et al., 2019; Perry et al., 2019; Zepeda et al., 2015). However, more recent research has also highlighted the importance of metacognition for a wide variety of cognitive processes, such as social cognition (Frith & Frith, 2012; Petty et al., 2007; Wright, 2002), belief updating (George & Mielicki, 2023; Stanovich & Toplak, 2023; van der Plas et al., 2022), self-regulation (Davis et al., 2010; Duckworth et al., 2014), and, central to this paper, reasoning (Ackerman & Thompson, 2017, Koriat, 2015).

In their seminal framework, Ackerman and Thompson (2017, p. 1) define metareasoning (i.e., metacognition in reasoning) as the "processes that monitor the progress of our reasoning and problem-solving activities and regulate the time and effort devoted to them." Through metacognitive monitoring, reasoners can evaluate how confident they are in their reasoning processes (Ackerman, 2014; De Neys et al., 2011; Jackson et al., 2016, 2017; Pennycook et al., 2017), assess the degree to which they intuitively feel like they have come to the right conclusions (Fernandez-Cruz et al., 2016; Gangemi et al., 2015; Thompson et al., 2011, 2013; Vega et al., 2021), and recognize when they have made reasoning errors (Fernandez-Cruz et al., 2016). Through metacognitive control, reasoners can determine whether they are satisfied with their reasoning processes (Ackerman, 2014; Ackerman et al., 2020; De Neys et al., 2013), and, if they are not, switch to a different reasoning strategy (Ackerman & Thompson, 2017; Cary & Reder, 2002; Haddara & Rahnev, 2022; Lieder & Griffiths, 2017). Unsurprisingly, high levels of metareasoning skill are associated with better reasoning (Batha & Carroll, 2007; Fleming & Daw, 2017; Ghazal et al., 2014).

Metareasoning skills are often assessed using one of a small set of paradigms that serve as analogues of the tasks used in the memory and learning literatures (for a detailed review of these paradigms, see Ackerman & Thompson, 2015, 2017). Metacognitive monitoring tasks typically compare participants' judgments about their performance on a reasoning task to their actual performance on that task. For example, participants might be asked to judge how solvable they think a problem is (Ackerman & Beller, 2017; Topolinski et al., 2016), how confident they are that they solved a puzzle correctly (Ackerman, 2014) or how strongly they feel that they arrived at a correct (or erroneous) solution (Fernandez-Cruz et al., 2016; Gangemi et al., 2015). These judgments are then compared to their actual performance and participants who provided more accurate judgments about their performance are said to have engaged in more effective metacognitive monitoring.

Metareasoning control tasks require decision makers to make decisions about how to proceed to best achieve a goal. For example, participants might be asked to complete a series of puzzles and make strategic decisions such as when to quit one puzzle and move onto the next one (Law et al., 2022; Payne & Duggan, 2011), how much costly evidence to collect before providing a final answer to a puzzle (De Neys et al., 2013; Thompson et al., 2013), or when to switch reasoning strategies to optimize their performance (Karpicke, 2009; Macaluso et al., 2022). Participants who engage in strategies (e.g., strategically allocating time) that allow them to more successfully achieve the goal (e.g., solving more puzzles) are then designated as having greater metareasoning control.

All of these paradigms share a common attribute: they require a correct or optimal performance measure (e.g., whether they solved the puzzle, how much time they wasted on an impossible task), to provide a straightforward means of assessing metacognitive performance. While these paradigms provide valuable insights into metareasoning, their dependence on objectively correct answers limits their functionality for assessing the role of metareasoning in subjective judgments and decisions, such as whom to marry, which house to purchase, which college to attend, or which medical treatment to undergo.

Metareasoning in Subjective Decisions.

We define subjective decisions as any decision in which the optimal or selected choice is determined by an individual's preferences, feelings, values, tastes, or beliefs, rather than fact or truth (e.g., Berman et al., 2018; Fishburn, 1981; Spiller & Belogolova, 2017; Weber & Federico, 2012). Because subjective decisions lack objectively correct answers, assessing the accuracy of participants' metacognitive judgments in such domains requires an entirely new set of methodological tools. Specifically, tools used to evaluate subjective metareasoning must assess the degree to which participants' metacognitive judgments are consistent with their subjective preferences and goals, which can be difficult to measure. One way in which this can be achieved is by having participants complete a reasoning task, then asking them to explain their reasoning process. The degree to which the participant's self-report aligns with their actual choice/judgment behavior reflects the degree of explicit, declarative metacognitive knowledge the participant has about their reasoning processes.

In line with this approach, a great deal of classic Judgment and Decision Making (JDM) work evaluated participants' knowledge of the factors that influenced their judgments and decisions. Much of this work relied on introspective verbal protocols (Ericsson & Simon, 1980) in which participants were asked to reflect on their decision-making processes and describe them to the researcher (e.g., Harte & Koele, 1995; Nisbett & Bellows, 1977; Wilson & Nisbett, 1978). This approach is perhaps best exemplified by a seminal series of studies in which Nisbett and Wilson (1977) found that participants were unable to provide accurate reports about the factors that influenced their judgments (e.g., film quality) or decisions (e.g., which socks to buy).

Taking a more quantitative approach, other classic JDM research sought to understand whether participants could accurately report the weight they placed on various cues when making subjective judgments and decisions (see Slovic & Lichtenstein, 1971 for a review).

Importantly, these judgments and decisions were multi-attribute in nature (i.e., the alternatives were compared on multiple dimensions), thus requiring the decision maker to make tradeoffs between the various attributes and determine how important each attribute was to them (Keeney & Raiffa, 1976; Soman, 2004). In essence, these studies asked participants to make a series of judgments then asked them to state how heavily they had weighted various cues when making their judgments. The participants' stated weights were then compared to objective weights estimated via linear regression (e.g., Hoepfl & Huber, 1970; Slovic, 1969; Slovic et al., 1972).

The Current Approach.

While these classic studies provided early insight into metareasoning in subjective decision making, they were limited to relatively simple methods that provided noisy estimates of participants' decision-making processes. Few studies have sought to systematically build upon these approaches using modern statistical and methodological approaches. To fill this gap in the literature we present a novel method for assessing participants' metacognitive knowledge of the cue weights they use when making subjective, multi-attribute choice decisions.

Before discussing our paradigm, it is worth noting that the term metacognitive knowledge has most frequently been used in the learning literature, where it describes students' knowledge of the strategies that they can use to learn most effectively (e.g., Pintrich, 2002; Vrugt & Oort, 2008). However, we use metacognitive knowledge in the broader sense of knowledge about any cognition, including judgment and decision making. This usage of the term is in line 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."

Furthermore, the literature has demonstrated that metacognitive knowledge is a relevant construct in non-learning domains, including problem solving (Antonietti et al., 2010), creativity (Jia et al., 2019) and, most relevantly, decision making (Basu & Dixit, 2022; Colombo et al., 2010). Using this framework, we contend that a decision maker's knowledge about the cognitive processes underlying their subjective decision making falls under the umbrella of metacognitive knowledge.

Our paradigm – which we call the *Knowledge Of Weights (KoW)* paradigm – is rooted in Choice-Based Conjoint Analysis (CBC), a technique from the marketing literature (e.g., Allenby et al., 1995; Hein et al., 2020; Lenk et al., 1996; Louviere & Woodworth, 1983; Sawtooth Software, Inc., 2017, 2021) in which participants are asked to make a series of choices between different sets of alternatives (e.g., 3 schools) that vary across a pre-determined set of attributes (e.g. graduation rates, crime rates, extra-curricular opportunities, etc.). CBC allows for estimation of the weight that each participant places on each attribute based on a small number of multi-attribute choices. We then compare these CBC weight estimates to participants' self-reports of the weights that they believe they used while making their choices. Importantly, these self-report items instruct participants to *retrospectively reflect* on the weights they used during their decision-making processes, not to make novel judgments about the importance of each attribute. This distinction is critical for interpreting the *KoW* paradigm as a metacognitive measure. Higher calibration or resolution between the estimated and self-reported weights indicates greater metacognitive knowledge (Fleming & Lau, 2014).The *KoW* paradigm is highly inspired by the classic metareasoning work (e.g., Nisbett & Wilson, 1977; Slovic, 1969; Slovic & Lichtenstein, 1971), but builds upon it in four ways:

First, CBC generates more precise estimates of participants' preferences and attribute weights by implementing hierarchical Bayesian models that, unlike simple regressions, can accurately capture non-linear (and even non-monotonic) utility functions. Because of this, CBC can detect, for example, that a decision maker prefers a house that is not too small and not too large, rather than assuming that all participants prefer larger homes. Similarly, non-monotonicity allows CBC to accurately model the preference order for categorical variables, which are often important in subjective decision making. Classic studies often avoided this concern by using dichotomous variables (e.g., Slovic, 1969, 1972), thus limiting their design choices.

Second, CBC only requires participants to make about a dozen choices – whereas the classic studies often required participants to complete more than 100 choice or judgment tasks (e.g., Slovic, 1969). Requiring participants to make so many choices likely reduced the ecological validity of the paradigms, induced fatigue, and pushed participants to use lexicographic strategies that may not be reflective of their typical decision-making processes (Bradley & Daly, 1994; Hirshleifer et al., 2019). Because CBC limits the number of choices participants are required to make, it is likely to generate more accurate estimates of their decision weights.

Third, modern advances in computing allow CBC programs to generate more complex designs (i.e., more attribute x level combinations; unique choice sets for each participant) than classic studies. These complex designs allow for additional randomization and orthogonality across attributes, choices, and participants, thus reducing the risk of biased designs and limiting the correlations between variables in the alternatives presented to the participants. This level of randomization simply was not possible when studies were run on paper (Slovic, 1969; Slovic & Lichtenstein, 1971).

Fourth, and perhaps most importantly, CBC allows the researcher to estimate weights at the individual level, rather than the sample level, even with a small number of data points from each participant (Allenby et al., 1995; Hein et al., 2020; Lenk et al., 1996; Louviere & Woodworth, 1983; Sawtooth Software, Inc., 2017, 2021). This is critical for assessing metacognitive knowledge, as averaging revealed weights across participants would make it impossible to assess the accuracy of individual participants' stated decision weights. Furthermore, averaging across the sample is likely to cause errors to average out, making the sample appear more accurate than the individuals truly are (Surowiecki, 2004). Using CBC eliminates this issue.

To demonstrate the utility of the present approach, we report the results of three studies. In Study 1, we show that the metrics generated by the *KoW* paradigm demonstrate test-retest reliability. In Study 2, we present evidence that the *KoW* paradigm generates similar results when applied across four distinct domains. And finally, in Study 3, we demonstrate that some of the metrics generated by the *KoW* paradigm are correlated with a decision maker's ability to make choices they are happier with, thus providing suggestive evidence of the predictive validity of the paradigm.

Study 1

The primary objective of Study 1 was to demonstrate the reliability of the novel *KoW* paradigm. To do so, participants completed the *KoW* paradigm and then completed the paradigm a second time 24-48 hours later. We then compared the results from the two phases to assess the test-retest reliability of the paradigm.

Methods

Participants

Simulations suggested that we needed a minimum of about 200 participants to complete both phases of the study to reliably estimate attribute weights. We recruited 275 Prolific participants to ensure that we would have sufficient data after accounting for attrition and incomplete responses. Of the 275 participants we recruited, 272 completed Phase 1 and were invited to complete Phase 2. 239 participants (87.9%) completed the second phase. Demographics of the sample are reported in the Supplemental Materials. The 12.1% of participants who attritted were statistically indistinguishable from the participants who completed both phases in terms of demographic characteristics (*p*s > .08, see Supplemental Materials). All participants in the final sample passed at least two out of four attention checks in Phase 1 (96.2% passed all four) and at least two out of three attention checks in Phase 2 (98.7% passed all three).

In Phase 1, the first attention check required participants to choose their favorite season, then choose the holiday that occurs during the season they chose from a list of four options. The second attention check required participants to follow instructions and choose orange from a list of colors. The third attention check came after the CBC portion of the survey and asked participants to identify what kind of items (homes) they were picking between from a multiple-choice list with four options. The fourth attention check was a five-point Likert scale asking participants how strongly they agreed with the statement that they were born in the year 1250 AD (Strongly Disagree and Disagree were treated as correct). Phase 2 repeated attention checks 2-4, except that colors were replaced with fruits in the second attention check and “in the year 1250 AD” was replaced with “on Mars” in the fourth attention check.

Procedures

At the start of the survey, participants were informed that this was a two-phase study that required them to return the next day. After providing consent, participants completed a Choice Based Conjoint (CBC) survey in which they were instructed to imagine that they were looking to buy a house. Participants were given a brief description of six attributes that explained the scale on which each attribute was scored, listed the five possible levels of each attribute, and provided context regarding a typical score for that attribute in the real world (see Supplemental Materials for language). Participants were then shown 14 sets of three hypothetical homes and asked to pick which one they would be most interested in purchasing based on each home's scores on the six attributes. An example of one of these choice tasks is provided in Figure 1.

After completing the CBC survey, participants were asked to describe the weight that they believed they had put on each attribute while making their choices during the CBC task. Participants were asked to *retrospectively reflect* on the choices they had made during the CBC task – not to state how important they felt the attributes were at the time of self-report. This distinction is necessary to interpret the *KoW* paradigm as a metacognitive task.

Participants self-reported their attribute weights in two formats, presented in a random order. In one format, participants self-reported how important each attribute was to them on a scale of 1 (*Not at All Important*) to 9 (*Extremely Important*). We refer to these as Attribute Importance Ratings or **AIRs**. In the other format, participants self-reported the percentage of their decision-making process that was based on each attribute (i.e., weights). Their responses for the six attributes were required to sum to 100%, but otherwise could be allocated in any way they wanted – including putting 100% weight on one attribute. We refer to these as Stated Attribute Weights or **SAWs**. Notably, SAWs are compensatory in nature, while AIRs are not. Participants also self-reported their confidence that their SAWs and AIRs accurately reflected their decision-making process during the CBC task. These confidence judgments are meta-metacognitive judgments that provide some insight about participants' beliefs about the accuracy of their own metacognitive judgments. Finally, participants completed a standard set of demographic questions, marking the completion of Phase 1.

Twenty-four hours after the final participant completed Phase 1, participants were invited back to complete Phase 2. Participants were told that they had 24 hours to return for the second phase. We chose to use a relatively short delay of 24-48 hours to minimize attrition, which is critical because less-consistent participants are most likely to attrit, artificially inflating test-retest metrics in choice experiments with long delays (Rigby et al., 2016). A short delay was also beneficial because it is reasonable to expect that participants' preferences may change over longer delays, making it difficult to assess reliability. Phase 2 was identical to Phase 1, except that it had one fewer attention check and did not include demographic questions. Participants also self-reported the extent to which they believed their attribute weights changed from Phase 1 to Phase 2 on a scale of 1 (*Stayed exactly the same*) to 5 (*Changed drastically*). On average, participants self-reported that their weights changed very little ($M = 1.99$, $sd = 0.89$).

Materials

In each phase, the homes were described in terms of the following six attributes: 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. These attributes were selected to reflect the attributes of a home that can be identified on a home listing website, like Zillow. We based the Mortgage attribute on percent of income instead of absolute price to avoid confounds based on individual differences in wealth and regional costs of living. We did not include pictures of the homes because pictures carry information about several different attributes, thus hindering our ability to isolate individual attributes.

We then created five discrete levels for each attribute, as CBC cannot use continuous variables. These levels were designed to reflect typical homes that would be found in American neighborhoods (see Supplemental Materials for details). We avoided extreme values so that no pairs would seem unreasonable when presented together (e.g., a 20,000 square foot home for 10% of your income). Using these levels, we generated 300 versions of the CBC survey for each phase. Each version contained a unique 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 CBC versions than participants, each participant experienced a unique set of choices. Since CBCs were generated separately for each phase, participants made different sets of choices during each of the two phases – thus providing the most stringent test of the reliability of the *KoW* paradigm. We generated the CBC versions using Lighthouse Studio (Sawtooth Software, Inc., 2023), but many alternative software tools with similar functionalities exist, including Conjointly (Analytics Simplified Pty Ltd, 2023), and the R package cbcTools (Helveston, 2023). Study 1 was not pre-registered, but all data, materials, and code have been made publicly available on OSF:

https://osf.io/uzqk5/?view_only=2c37ad40b8ad4b0c8be6a0982ca655a6

Analysis & Results

Estimating Revealed Attribute Weights (RAWs)

Before we could assess participants' metacognitive knowledge of their attribute weights, we first had to estimate the weights revealed by their choices on the CBC task – which we will refer to as Revealed Attribute Weights, or **RAWs**. The first step in calculating RAWs was to estimate the part-worth utility (henceforth, just utility) that each participant placed on each level of each attribute. We did so using Hierarchical Bayes (HB) estimation, which is considered the gold standard for analyzing CBC data (Eggers et al., 2022; Hein et al., 2020; Orme, 2002). We estimated the utilities using only the data from participants who completed both phases of the study (*n* = 239). Because we used a standard HB procedure, we provide only a brief plain-language overview here. For a detailed technical overview, see Sawtooth Software, Inc. (2021).

The HB model begins with the assumption that nothing is known about the utility of each level of each attribute (i.e., all utilities are set to 0). The model then generates a new set of utilities (as well as variances and covariances) 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 and used to inform the next set of utilities. This continuous updating process – known as a Monte Carlo Markov Chain – leads the model to eventually converge on accurate utility estimates. For this study, we allowed the model to generate 20,000 iterations, the first 10,000 of which were used to calibrate the model and the remaining 10,000 of which were averaged to generate point estimates of utilities. Importantly, the strength of HB arises from the fact that its upper-model uses sample-level mean utility estimates to inform individual-level utility estimates, which are then applied at the lower-level to estimate the likelihood of participants' choices given a provided set of utilities. This hierarchical procedure allows for precise utility estimates with relatively little data.

When the HB process is complete, it produces point estimates of the utility of each level of each attribute for each participant (6 attributes x 5 levels = 30 utilities for each participant).

These utilities must then be converted into a single RAW for each attribute. In adherence to standard CBC practice (e.g., Eggers et al., 2022; Orme, 2002), we do so using the following equation, where U_j is a vector containing the utility values for the five levels of attribute *j* (the attribute for which the RAW is being calculated), 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 six attributes from the CBC, including *j:*

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

RAWs are estimated separately for each participant, thus allowing participants to have varying preferences regarding the highest- and lowest-utility level of each attribute. Notably, this method of estimating RAWs only considers the highest- and lowest-utility levels of each attribute, and thus disregards the shape of the utility function between those two points (e.g., Eggers et al., 2022; Orme, 2002). This could be considered a limitation because it simplifies participants' preferences and limits the insights that can be made about the value placed on each level of each attribute but could also be considered a strength because it eliminates the need to make assumptions about the shape of the utility function. For this study, both the HB estimation and RAW calculations were conducted using Lighthouse Studio (Sawtooth Software, Inc., 2023). We provide a detailed, step-by-step walk-through of the process for converting utilities into RAWs in the Supplemental Materials.

Predicting Choices Using RAWs and SAWs

To ensure that the estimated RAWs were accurately capturing participants' attribute weights, we sought to assess how frequently they could predict participants' actual choices. To do so, we analyzed each CBC task that each participant completed (*n* = 3,346 tasks) and estimated the utility that they would have assigned to each of the three homes by multiplying their RAW for each attribute by the level of the attribute (scored as 1-5) and summing across the six attributes. The attribute level scores were transformed to match each participant's rank-order preferences for the five levels of each attribute, with their lowest-utility level being scored as a 1 and their highest-utility level being scored as a 5, assuming a linear utility function between the levels of each attribute.[1] This process was conducted for Phase 1 and Phase 2. We found that Phase 1 RAWs accurately predicted 92.89% ($\kappa = .89$) of participants' actual choices from Phase 1 and Phase 2 RAWs accurately predicted 93.37% ($\kappa = .90$) of participants' actual choices from Phase 2, suggesting that RAWs are a good estimate of participants' true decision weights. We then made cross-phase comparisons, testing the ability of RAWs from one phase to predict choices from the other phase. We found that Phase 1 RAWs accurately predicted 83.56% of Phase 2 choices ($\kappa = .75$) and Phase 2 RAWs accurately predicted 83.14% of Phase 1 choices ($\kappa = .75$). Unsurprisingly, Phase 1 RAWs, $X^2(1, N\text{s} = 3{,}346) = 139.12, p <.001$, and Phase 2 RAWs, $X^2(1, N\text{s} = 3{,}346) = 167.63, p <.001$, were significantly less-accurate predictors of cross-phase (i.e., out-of-sample) choices than within-phase (i.e., in-sample) choices.

We also repeated this process using SAWs as decision weights instead of RAWs. Phase 1 SAWs accurately predicted 82.45% ($\kappa = .74$) of choices from Phase 1 and Phase 2 SAWs accurately predicted 83.01% ($\kappa = .75$) of choices from Phase 2. Unsurprisingly, the proportion of within-phase accurate predictions was significantly higher for RAWs than SAWs in both Phase 1, $X^2(1, N\text{s} = 3{,}265 – 3{,}346) = 166.28, p <.001$, and Phase 2, $X^2(1, N\text{s} = 3{,}267 – 3{,}346) = 169.88, p <.001$. We then evaluated cross-phase predictions. Phase 1 SAWs accurately predicted 81.52% of

[1] For the choice predictions, we chose to assume linearity across levels instead of using the utility values we estimated for each level so that we could test how accurate the RAWs were without additional information. This allowed us to more fairly compare the accuracy of RAWs and SAWs, which were only collected for each attribute, not each level. As noted earlier, RAWs themselves were estimated without linearity/monotonicity assumptions.

Phase 2 choices (κ = .72) and Phase 2 SAWs accurately predicted 80.76% of Phase 1 choices (κ = .71). Both Phase 1 SAWs, X^2(1, *N*s = 3,274 – 3,346) = 4.65, *p* =.03, and Phase 2 SAWs, X^2(1, *N*s = 3,270 – 3,346) = 6.18, *p* =.01, were less accurate in making cross-phase choice predictions than their respective RAWs. However, the predictive power of RAWs was much closer to the predictive power of SAWs when making out-of-sample predictions than in-sample predictions, an unsurprising finding given that RAWs are directly estimated from in-sample choices.

Compositional Transformation

We next sought to evaluate the reliability of our three measures of participants' attribute weights. Before we could do so, however, we had to address the statistical challenge that decision weights – such as RAWs and SAWs, but not AIRs – are compositional in nature, meaning that all values for a single participant are constrained to sum to a constant value (e.g., all RAWs add to 100%). Because of this, weights for an individual participant are not independent of one another. This dependency can induce spurious relationships between weights, particularly when comparing one set of weights to another (Aitchison, 1982; Smithson & Broomell, 2024).

Here, we mitigate these dependencies by transforming the weights into numbers that exist in unrestricted Euclidean space. There are multiple ways to achieve this, but we will follow the Centered Log-Ratio Transformation Method, as described by Smithson & Broomell (2024). For our case, we will separately transform RAWs and SAWs using a four-step process:

1) Divide weights by 100 to place them on a 0-1 scale;
2) Replace weights of 0 and 1 with 0.01 and 0.99, respectively, to avoid logs that are zero or undefined (i.e., the simple replacement method);
3) Calculate the log of each weight;

4) Subtract the mean of the six log-transformed weights from each log-transformed weight.

The remainder from Step 4 is the transformed weight for each attribute. We will use these transformed weights – which we call tRAWs and tSAWs – in place of their non-transformed counterparts in all analyses in which we compare two sets of weights. We provide a step-by-step walkthrough (with examples) of the Centered Log-Ratio Transformation Method in the Supplemental Materials. It is worth noting that non-transformed RAWs and SAWs will still be used whenever the analyses do not compare two sets of weights (e.g., when making choice predictions).

Reliability of tRAWs, tSAWs, and AIRs.

Having completed the compositional transformation, we could now evaluate the reliability of our three measures of participants' attribute weights – tRAWs, tSAWs, and AIRs. To do so, we evaluated the correlations between individual participants' tRAWs, tSAWs, and AIRs across the two phases. Higher correlations indicate greater reliability. Attribute-level correlations between Phase 1 tSAWs and Phase 2 tSAWs ranged from r = .52 - .89, with an average of r = .76. Participants' confidence in the accuracy of their SAWs was high in both phases (see Table 1) and highly correlated across timepoints (r = .81, p < .001). Attribute-level correlations between Phase 1 AIRs and Phase 2 AIRs ranged from r = .70 - .88, with an average of r = .78. Participants' confidence in the accuracy of their AIRs was also high in both phases (see Table 1) and highly correlated across timepoints (r = .77, p < .001). These results suggest that tSAWs, AIRs, and both confidence metrics were reliable across phases.

Attribute-level correlations between Phase 1 tRAWs and Phase 2 tRAWs ranged from r = .33 - .60, with an average correlation of r = .47. While this average falls well below the typical reliability threshold of r = .70, it is not far below reliability metrics for other behavioral measures commonly used in the decision sciences, including the Columbia Card Task (r = .57; Buelow & Barnhart, 2018) and the Domain Specific Risk-Taking Scale (subscale rs = .42 - .80, Mean r ~ .61; Weber et al., 2002).

The sub-optimal reliability of the tRAWs could arise from two distinct sources: measurement error or participants applying their preferences inconsistently across the two timepoints, which could be a result of the very metacognitive errors we are trying to capture. To disentangle these explanations, we used participants' utilities from Phase 1 to calculate the utility of each alternative they evaluated during Phase 2. We then estimated what each participant's Phase 2 tRAWs would have been if they had always chosen the alternatives that maximized their Phase 1 utilities. The correlation between these predicted Phase 2 tRAWs and the Phase 1 tRAWs for each attribute ranged from r = .64 - .85 (ps < .001), with an average of r = .75. These high correlations suggest that the *KoW* paradigm itself generates reliable tRAW estimates and that a great deal of the sub-optimal reliability of tRAWs (the gap between r = .48 and r = .75) can be attributed to participants' limited ability to apply their preferences consistently across phases. This provides greater confidence in the reliability of the *KoW* paradigm.

The reliability metrics for tRAWs, tSAWs, AIRs, and each confidence measure are summarized in Table 1. Attribute-level means and cross-phase correlations for tRAWs, tSAWs, and AIRs are reported in the Supplemental Materials. Reliability metrics for the non-transformed RAWs and SAWs are also reported in the Supplemental Materials.

Assessment of Metacognitive Knowledge

In this section, we will present 4 metrics that can be used to assess participants' metacognitive knowledge of the attributes weights they used. Our first measure is the correlation between tRAWs and tSAWs (henceforth, tRAW-tSAW Correlations). Higher correlations indicate greater metacognitive calibration (Fleming & Lau, 2014), and therefore greater metacognitive knowledge. Like all correlations, tRAW-tSAW correlations are theoretically bounded at -1 and +1. Decisions makers choosing randomly would be theoretically expected to achieve average correlations of about 0, assuming the attributes are independent of one another. In Phase 1 of this study, tRAW-tSAW correlations for each attribute ranged from *r* = .34 - .64 and the average of the six correlations was *r* = .54. In Phase 2, tRAW-tSAW correlations for each attribute ranged from *r* = .42 - .60 and the average of the six correlations was *r* = .53. The average tRAW-tSAW correlations from Phase 1 and Phase 2 were not significantly different (*z* = 0.10, *p* = .92).

Our second metric is the correlation between tRAWs and AIRs (henceforth, tRAW-AIR Correlations). Again, higher correlations indicate greater metacognitive calibration (Fleming & Lau, 2014), and therefore greater metacognitive knowledge. tRAW-AIR correlations are also theoretically bounded at -1 and +1 and decision makers choosing randomly would be theoretically expected to achieve average correlations of about 0, assuming the attributes are independent of one another. In Phase 1 of this study, tRAW-AIR Correlations for each attribute ranged from *r* = .22 - .66 and the average of the six correlations was *r* = .50. In Phase 2, tRAW-AIR correlations for each attribute ranged from *r* = .43 - .62 and the average of the six correlations was *r* = .52. The average tRAW-AIR correlations from Phase 1 and Phase 2 were also not significantly different from one another(*z* = -0.32, *p* = .75). These results suggest that performance does not improve when the paradigm is completed a second time (i.e., no evidence of practice effects). Attribute-level comparisons of tRAW-tSAW and tRAW-AIR correlations across phases are reported in the Supplemental Materials.

Our third metric is the Euclidean distance between each participant's tRAWs and tSAWs. Following Smithson & Broomell (2024), we calculated these Euclidean distances according to the following formula, 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:

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

Lower Euclidean distances indicate greater metacognitive resolution (Fleming & Lau, 2014), and therefore greater metacognitive knowledge. The theoretical lower bound for Euclidean distances is zero, but there is no theoretical upper bound (Smithson & Broomell, 2024). Simulations presented in Study 2 demonstrate that participants making random choices across four domains achieve mean Euclidean distances of 0.42 - 0.46, which can be interpreted as a benchmark range akin to chance. However, this benchmark is likely to be sensitive to the parameters of the task, such as the number of attributes used.

In this study, a paired t-test indicated that there was no significant difference between the mean Euclidean distances from Phase 1 (*M* = 0.34, *sd* = 0.14) and Phase 2 (*M* = 0.36, *sd* = 0.15, *t*(238) = -1.78, *p* = .08), again providing no evidence of practice effects. The correlation between participants' Euclidean distances from Phase 1 and Phase 2 was *r* = .59, which falls below the standard threshold of reliability (*r* = .70). This is unsurprising, given that the reliability of Euclidean distances is functionally bounded by the reliability of two other metrics (tRAWs and tSAWs), one of which (tRAWs) has sub-optimal reliability due to participants' inconsistent application of preferences. While our evidence does not indicate that the Euclidean distances metric is highly reliable, we contend that it is still a meaningful metric, as a correlation of *r* = .59

suggests that we are capturing some – admittedly noisy – individual differences in metacognitive knowledge.

Our fourth and final metric is the percentage of CBC task for which RAWs and SAWs would predict different choices (see *Predicting Choices Using RAWs & SAWs*). Lower percentages indicate greater metacognitive calibration of outcomes, and therefore metacognitive knowledge. Values for this metric are theoretically bounded at 0% and 100% but, accounting for chance, RAWs and SAWs would be theoretically expected to make different predictions only 66.67% ($\kappa = 0$) of the time for a decision maker behaving randomly. In this study, we found that RAWs and SAWs predicted different choices in 15.72% of Phase 1 choice tasks ($\kappa = .76$) and 16.16% of Phase 2 choice tasks ($\kappa = .76$). These proportions were not significantly different from one another, $X^2(1, Ns = 3{,}264 – 3{,}267) = 0.21, p = .65$, providing no evidence of practice effects. The reliability of each of the metacognitive metrics is summarized in Table 1.

Study 1 Discussion

At the individual level, Study 1 demonstrated that the *KoW* paradigm produces reliable estimates of participants' self-reported attribute weights (tSAWs and AIRs). Participants' revealed weights (tRAWs) were less reliable ($r = .47$) than their self-reported weights, but this sub-optimal reliability largely arose from participants' inconsistent application of their preferences across phases, not measurement error. The Euclidean distance metric was only marginally reliable ($r = .59$), suggesting that it is a noisy – albeit potentially useful – individual differences metric.

At the sample level, Study 1 demonstrated that the *KoW* paradigm's four key metacognitive metrics – tRAW-tSAW correlations, tRAW-AIR correlations, mean Euclidean distances, and the percentage of tasks for which RAWs and SAWs would predict different choices – were highly consistent across timepoints, providing no evidence of improvement in performance from repeated exposure to the task (i.e., practice effects). In all, the primary conclusion of Study 1 is that the *KoW* paradigm is a fairly reliable tool for assessing metacognitive knowledge of attribute weights in subjective domains, though its reliability is hindered by participants' inconsistent behaviors and metacognitive limitations. We also encourage future researchers to consider that test-retest reliability may not be an appropriate assessment of the *KoW* paradigm over long periods of time or in domains where preferences change frequently. Tastes and preferences are often unstable, and metacognitive knowledge of those tastes and preferences may be as well.

Study 2

In Study 1, we explored the *KoW* paradigm in one domain: homebuying. In Study 2, we explore whether the *KoW* paradigm generates similar results across other subjective domains that, like homebuying, are highly familiar and often thought about in terms of tradeoffs across attributes. We do so by randomly assigning participants to complete the *KoW* paradigm in one of four domains (Homebuying, Dating, College Choice, Job Selection) and comparing sample-level metacognitive metrics across the domains. Our pre-registered hypothesis was that participants would have similar levels of metacognitive knowledge in each domain – which should be reflected in the metrics generated by the *KoW* paradigm. If this prediction holds, it would further support the conclusion that the *KoW* paradigm is a valid and consistent measure of participants' metacognitive knowledge. This study and its analyses were pre-registered on OSF: https://osf.io/mw3qg/?view_only=92e94f752db440beb2f563b674495cda

Methods

Participants

As in Study 1, simulations suggested that we needed a minimum of about 200 participants per domain to reliably estimate their attribute weights. We recruited 850 Prolific participants to ensure that we would have sufficient data after accounting for incomplete responses. 825 participants completed the study and were randomly assigned to complete the *KoW* paradigm in one of four domains: Job Selection (i.e., Jobs; n = 202), Homebuying (i.e., Homes; n = 214), Romantic Partner Selection (i.e., Dating; n = 196) or College Choice (i.e., Colleges; n = 213). Demographics of each sample are reported in the Supplemental Materials. Participants in the four domains were statistically indistinguishable in terms of gender, age, race, ethnicity, income, and education (ps > .22; see Supplemental Materials for descriptives and significance tests). We used the same attention checks as in Phase 1 of Study 1. All participants correctly answered the attention check (which also functioned as a manipulation check) asking what type of items (e.g., homes, romantic partners) they picked between during the CBC task and passed at least two out of the three other attention checks (97.1% passed all three).

Procedures

All procedures were identical to the first phase of Study 1, except that Study 2 had no longitudinal component and participants were randomly assigned to make choices in one of four different domains (Jobs, Homes, Dating, Colleges). In each domain, participants were given a brief prompt asking them to imagine that they were actively choosing between different domain-relevant alternatives (e.g., in the dating domain, participants were told to imagine that they were single and had recently downloaded a dating app like Tinder). The full text for these prompts is provided in the Supplemental Materials. As in Study 1, participants then completed tasks that measured their RAWs, SAWs, AIRs, and confidence.

Materials

For each domain, the alternatives presented during the CBC task were described in terms of six attributes. For the Homes domain, the attribute set was the same as Study 1. For Jobs, the attribute set was inspired by the information that might be found on a job listing (e.g., company size, commute time) or a company review site, like Glassdoor.com (e.g., a rating of company culture). For the Dating domain, the attribute set was inspired by the information available on dating apps like Tinder or Bumble (e.g., education level, political affiliation). Finally, for the Colleges domain, the attributes were inspired by the information that might be found on college ranking websites, like *U.S. News & World Report* (e.g., number of students, ranking). As in Study 1, we created five discrete levels for each attribute that were used to randomly generate the alternatives for the CBC task. The full list of attributes and levels used for each domain and the language used to describe them to participants is provided in the Supplemental Materials.

Analysis & Results

Deviation from Pre-Registration

The pre-registration for Study 2 stated that we would use metrics based on non-transformed RAWs and SAWs in all our analyses. However, reviewers pointed out that these pre-registered analyses were not always ideal given the compositional nature of our data. As such, we chose to report results using tRAWs, tSAWs, and Euclidean distances in place of their non-transformed equivalents when doing so is more statistically appropriate (i.e., when comparing two sets of weights). We believe that this deviation from our pre-registration will increase the validity of our inferences, thus making it justifiable (Lakens, 2024). For completeness and in the interest of transparency and open science, the pre-registered analyses using the non-transformed values (RAWs, SAWs, and RAW-SAW Differences) are reported in the online supplement.

Regardless of which strategy is used, the results are qualitatively similar, and the inferences and conclusions that we draw are substantively the same.

Calculating Metrics

RAWs and RAW/SAW Choice Predictions were calculated in the same way as in Study 1. Full descriptives for RAWs, SAWs, AIRs, SAW/AIR Confidence and RAW/SAW Choice Predictions are reported in the Supplemental Materials. RAWs and SAWs were once again transformed into tRAWs and tSAWs using the Centered Log-Ratio Transformation Method, and Euclidean distances were calculated using the same procedure described in Study 1.

RAW/SAW Choice Prediction Accuracy

To ensure that our RAWs were accurately capturing participants' attribute weights, we first evaluated the accuracy of the RAW and SAW Choice Predictions. In each domain, RAWs accurately predicted between 91.90 – 93.93% of participants' actual choices (κs = .88 - .91), whereas SAWs only accurately predicted between 81.66 – 84.11% (κs = .72 - .76) of participants' choices. Two-sample tests for equality of proportions (which were not pre-registered) indicated that, as expected, RAWs were significantly better predictors of participants' choices than SAWs for all domains, $X\text{s}^2(1, N\text{s} = 2{,}689 – 2{,}996) > 88.46$, $p\text{s} < .001$. This finding gave us confidence that the RAWs were accurately estimating participants' attribute weights.

Confidence Across Domains

We next sought to evaluate whether participants had similar confidence in the accuracy of their SAWs and AIRs across domains. All analyses in this section were pre-registered. Participants' average confidence in the accuracy of their SAWs (*M*s = 78.49 – 81.92) and AIRs (*M*s = 80.28 – 83.72) was high in all domains. There were no significant differences across domains in SAW confidence, $F(3, 820) = 2.16$, $p = .09$, or AIR confidence, $F(3, 821) = 2.26$, $p =$

.08. The two confidence measures were highly correlated in all domains (*r*s = .71-.86, *p*s < .001). These results indicate that participants had similar levels of perceived familiarity with their weights across the domains, giving further reason to believe that the metacognitive knowledge metrics should be similar across domains.

Metacognitive Metrics Across Domains

We then compared the metacognitive metrics generated by the *KoW* paradigm across domains. All analyses in this section were pre-registered, with the caveat that RAWs and SAWs were replaced with tRAWs and tSAWs when comparing multiple sets of weights.

In each domain, the average tRAW-tSAW correlation was between *r* = .48 - .51 (see Figure 2). Pairwise fisher's *r* to *z* tests indicated no significant differences in average tRAW-tSAW correlations across domains (*n*s = 196-214; *z*s < 0.29, *p*s > .77). Average tRAW-AIR correlations for each domain ranged from *r* = .45 - .50 (see Figure 2). Pairwise fisher's *r* to *z* tests indicated that there were no significant differences in average tRAW-AIR correlations across domains (*n*s = 196-214; *z*s < 0.59, *p*s ≥ .56). Average tRAW-AIR correlations were not significantly different than Average tRAW-tSAW correlations in any domain (*n*s = 196-214; *z*s < 0.62, *p*s > .54)[2].

Average Euclidean distances ranged from 0.31 – 0.37 across domains. An omnibus ANOVA indicated that there were significant differences in average Euclidean distances across domains, *F*(3, 821) = 6.94, *p* < .001. Post-hoc pairwise comparisons using Tukey's HSDs indicated that the Jobs domain (*M* = 0.31, *sd* = 0.13) had a significantly lower average Euclidean distance than the Dating (*M* = 0.37, *sd* = 0.14, *p* < .001), College (*M* = 0.35, *sd* = 0.13, *p* = .02),

[2] Attribute-level tRAW-AIR, tRAW-tSAW, and tSAW-AIR correlations are reported in the Supplemental Materials.

and Homes domains (*M* = 0.36, *sd* = 0.15, *p* = .001). All other pairwise comparisons were non-significant (*p*s $\geq$ .65; see Figure 2).

RAWs and SAWs predicted different choices in 14.39 – 17.34% (κs = .74 - .78) of CBC tasks across domains (see Figure 2). An omnibus four-sample test of equality of proportions indicated that these proportions were significantly different from one another, X^2(3, *N*s = 2,690 – 2,942) = 9.82, *p* = .02. Post-hoc pairwise tests of equality of proportions using Bonferroni corrections indicated that RAWs and SAWs made different predictions for a smaller proportion of choices in the College domain (14.39%) than the Homes domain (17.34%, *p* = .01). All other pairwise comparisons were non-significant (*p*s > .24).

Optimal Metacognition Simulations

As a check of the sensitivity of the *KoW* paradigm, we investigated how participants' metacognitive knowledge metrics would have changed if participants had optimal metacognitive knowledge of their attribute weights. To do so, we created new versions of the CBC survey for each domain and completed one CBC survey (14 choices) for each human participant, simulating which alternative each participant would have chosen if they had used their SAWs as their true decision weights. These simulations were conducted using the same procedure as the RAW/SAW Choice Prediction analyses. The predicted choices were entered into Lighthouse Studio (Sawtooth Software, Inc., 2023) by research assistants. We then estimated the RAWs and tRAWs for each simulated optimal participant and calculated the three SAW/tSAW-based metacognitive metrics for the simulated optimal respondents. Optimal tRAW-AIR correlations could not be meaningfully calculated because there is no objectively optimal mapping between RAWs and AIRs.

The average tRAW-tSAW correlation for the simulated optimal respondents in each domain ranged from *r* = .59 - .65. These correlations were greater than the average tRAW-tSAW correlations for the human participants in all domains, but the difference was only significant in the Homes domain (*n* = 214, *z* = 2.47, *p* = .01; *p*s for other domains = .08 - .22). The mean Euclidean distance for the simulated optimal respondents in each domain ranged from 0.27 – 0.33. These means were significantly lower than the mean Euclidean distances for human participants in all domains (*t*s = 2.32 - 4.86, *df*s = 388.65 - 425.99, *p*s < .02). Across domains, RAWs and SAWs predicted different choices in 5.98 – 7.31% (κs = .89 - .91) of choice tasks completed by the optimal simulated respondents. These proportions were significantly lower than the proportions for human participants in all domains, *X*s^2(1, *N*s = 2,690 – 2,959) > 103.23, *p*s < .001. These results indicated that when participants have better (in this case perfect) metacognitive knowledge, the metacognitive knowledge metrics improve as well, suggesting that they are sensitive measures of metacognitive knowledge. Notably, however, the tRAW-tSAW correlation metric was not as sensitive as the other metrics. Attribute-level tRAW-tSAW correlations for the simulated optimal respondents are reported in the Supplemental Materials.

Random Simulations

Finally, we conducted similar simulations for each domain in which simulated respondents made completely random choices – thus reflecting zero metacognitive knowledge. The random responses were generated via Lighthouse Studio (Sawtooth Software, Inc., 2023). We then estimated RAWs and tRAWs for these simulated participants and paired each simulated respondent's RAWs and tRAWs with a randomly selected human respondent's SAWs, tSAWs and AIRs, and calculated the four metacognitive metrics. The tRAW-tSAW correlation, tRAW-AIR correlation, and RAW/SAW Different Choice Prediction metrics have clear theoretical predictions for a random agent (*r* = 0, *r* = 0, and 66.67%, respectively), so these simulations can be interpreted as sanity checks. For the Euclidean distance metric, the theoretical prediction is not self-evident, so these simulations were useful for establishing an empirical standard of comparison for human performance. These random simulations were not pre-registered.

As expected, Average tRAW-tSAW (*r*s = -.01 - .04) and tRAW-AIR correlations (*r*s = -.01 - .04) for each domain were effectively zero. These correlations were significantly lower than the correlations achieved by human participants (*n*s = 196-214, *z*s < -4.74, *p*s < .001). RAWs and SAWs predicted different choices in 66.06 – 67.09% (κs = -.01 - .01) of choice tasks. These proportions were significantly greater than the proportions for human participants in all domains, *X*s^2(1, *N*s = 2,690 – 2,943) > 1386.70, *p*s < .001. Average Euclidean distances for the simulated random participants ranged from 0.42 – 0.46. These means were significantly higher than the means for human participants in each domain (*t*s = 3.49 - 8.23, *df*s = 366.47 - 401.59, *p*s < .001). The results of these random simulations provided confidence that our metacognitive knowledge metrics were functioning as expected and demonstrated that human participants were, as expected, outperforming random agents. Attribute-level tRAW-tSAW correlations and tRAW-tAIR correlations for the simulated random respondents are reported in the Supplemental Materials.

Study 2 Discussion

The primary conclusion from Study 2 is that the *KoW* paradigm generates similar results in four distinct multi-attribute choice domains. While there were some differences across domains on two of the metacognitive knowledge metrics, the effects were scattered across different domains, therefore providing little evidence that overall metacognitive knowledge was consistently greater or worse in any particular domain. Study 2 also demonstrated that the *KoW* paradigm is sensitive to changes in metacognitive knowledge by showing that simulated participants with optimal metacognitive knowledge do in fact score better on the *KoW* paradigm's metrics than (sub-optimal) human participants. Furthermore, Study 2 demonstrated that, as expected, human participants' metacognitive knowledge was greater than would be expected if they had made random choices during the CBC task. Taken together, these results provide further evidence that the *KoW* paradigm is an effective tool for measuring participants' metacognitive knowledge of attribute weights in subjective decisions. Future research should seek to test the *KoW* paradigm in other domains that are qualitatively different from those tested here – such as decisions that are not typically thought about in terms of tradeoffs across attributes (e.g., which friend to hang out with).

Study 3

Having shown that the *KoW* paradigm is reliable (Study 1) and produces consistent results across distinct domains (Study 2), we now turn our attention to the validity of the paradigm. One way to test the validity of the paradigm is to demonstrate that measures generated by the paradigm are predictive of outcomes of functional importance (i.e., predictive validity). One important metric for evaluating subjective decisions is whether a participant's choices achieve their personal goals. Study 3 will focus on music choices, which are a convenient domain in that nearly every music consumer has the same goal: personal enjoyment. Study 3 seeks to determine whether individual differences in performance on the *KoW* paradigm – operationalized as Euclidean distances – predict whether participants can effectively choose music that maximizes their personal enjoyment.

Methods

Participants

To ensure that we would have at least 200 usable observations, we recruited 220 Prolific participants to complete a version of the *KoW* paradigm about songs. All 220 participants completed the study and passed at least two out of four attention checks (93.2% passed all four). We used the same attention checks as Phase 1 of Study 1. Demographics of the sample are reported in the Supplemental Materials

Procedures

The procedure for Study 3 was largely the same as for the previous two studies, except that the domain was song choices. During the CBC portion of study, participants made 15 choices between three hypothetical pop songs, each of which was described in terms of six attributes that could take one of five discrete levels (see *Materials* below). Participants were asked "which of these songs do you think you would most enjoy listening to." As in Studies 1 and 2, the first 14 choices were randomly generated permutations of the possible levels of the six attributes. The 15th choice, however, was a fixed task, meaning that all participants saw the same three songs, which aligned with three real songs.

After completing the SAW and AIR tasks, participants listened to all three songs that corresponded to the songs presented in the Fixed Task, presented in a random order. After listening to each song, participants rated how much they enjoyed the song on a scale of 0 (*Did not enjoy at all*) to 100 (*Enjoyed greatly*) and self-reported whether they had heard the song before. After listening to all three songs, participants completed a multiple-choice item indicating which song they enjoyed listening to most. Participants then answered demographic questions. This study and its analyses were pre-registered on AsPredicted: https://aspredicted.org/V9W_T4V

Materials

In the CBC task, each song was described in terms of the following six attributes: 1) Acousticness (how acoustic or electric a song is); 2) Danceability (how easy it is to dance to a song); 3) Tempo; 4) Length; 5) Decade Released; and 6) Artist Type (group composition and singer gender). The first three attributes came from Spotify, which evaluates the audio features of every song in their library and makes the data public through their API, Spotify for Developers (Spotify, Inc., 2023). These attributes were selected over other possible Spotify attributes because they captured qualitatively different elements of songs and are not strongly correlated with one another (see Supplemental Materials). Spotify describes Acousticness and Danceability on a 0-1 scale, but we multiplied the scale by 10 for ease of participant interpretation. Length was chosen because online participants are highly sensitive to how long they spend completing a task. The last two attributes were selected based on suggestions from participants in generative pilot tests. In a final pilot test, participants self-reported average SAWs between 10% - 22% for each attribute, suggesting no attribute was dominant or irrelevant (see Supplemental Materials). The instructions used to describe the attributes and their levels to participants are provided in the Supplemental Materials.

We pilot tested 8 songs (see Supplemental Materials) to be used in the fixed task. Our goal was to identify three songs that 1) Scored very differently on the six focal attributes, maximizing variation; 2) Were not well-known by pilot participants; and 3) Were similarly enjoyed by pilot participants. Based on these criteria, we chose *First Day of Summer*, by Jesse Ruben (2018), *Prisoner of Love*, by Miami Sound Machine (1984), and *Seasons*, by Grace Slick (1980). A screenshot of the Fixed Task is provided in Figure 3.

Analysis and Results.

Deviation from Pre-Registration

As in Study 2, the pre-registration for Study 3 stated that we would use metrics based on non-transformed RAWs and SAWs in all our analyses. However, we have chosen to report results using tRAWs, tSAWs, and Euclidean distances in place of their non-transformed equivalents when doing so is more statistically appropriate (i.e., when comparing two sets of weights). The pre-registered analyses using the non-transformed values are reported in the online supplement. The results are qualitatively similar regardless of strategy.

Metacognitive Knowledge Metrics

RAWs, tRAWs, tSAWs, RAW/SAW Choice Predictions, and our four metacognitive knowledge metrics were calculated using the same procedures as in Study 1 and Study 2. Following CBC convention, RAWs were estimated based only on the 14 random tasks. Cross-domain comparisons of these metrics to the same metrics from Study 2 are provided in the Supplemental Materials. All analyses in this section were pre-registered as exploratory.

Attribute-level tRAW-tSAW correlations ranged from $r = .23 - .57$, with an average of $r = .39$. Participant confidence in the accuracy of their SAWs was high ($M = 78.00$, $sd = 19.15$). Attribute-level tRAW-AIR correlations ranged from $r = .18 - .52$, with an average of $r = .32$. Participant confidence in the accuracy of their AIRs was also high ($M = 78.21$, $sd = 20.19$). The mean Euclidean distance was 0.34 ($sd = 0.14$). RAWs accurately predicted 93.47% of choice tasks ($\kappa = .90$), whereas SAWs accurately predicted only 81.96% of choice tasks ($\kappa = .73$). RAWs and SAWs predicted different choices in 16.45% of choice tasks ($\kappa = .75$).

Choice Satisfaction

We next evaluated whether participants with greater metacognitive knowledge of their attribute weights made choices on the fixed CBC task that better aligned with their actual enjoyment of the songs. Descriptives regarding participants' choices on the fixed CBC task and their enjoyment of each song are provided in Table 2.

We first explored whether participants who gave the highest enjoyment rating to the song that they chose during the CBC task had lower Euclidean distances than participants who did not. Thirteen participants had a two- or three-way tie for the song they rated as most enjoyable, so they were excluded from these analyses. 40.1% of participants rated the song they chose during the CBC task as most enjoyable.

Participants who rated the song they chose during the CBC task as most enjoyable had slightly lower Euclidean distances ($M = 0.33$) than those who didn't ($M = 0.35$), but a t-test indicated that the effect was not significant ($t(181.33) = 1.28$, $p = .20$). Excluding participants who had heard any of the songs before ($t(157.39) = 1.01$, $p = .31$) did not meaningfully alter the results. Additional robustness analyses reported in the Supplemental Materials show similar patterns. Though we initially hypothesized that participants whose CBC choices were consistent with their enjoyment ratings would have lower Euclidean distances – thus reflecting the benefit of greater metacognitive knowledge – in retrospect the lack of effect is unsurprising, given the limited power of assessing a single binary outcome (success/failure).

We next turned our attention to the nearly 60% of participants who did not rate the song they chose during the CBC task as most enjoyable. Using this sample, we explored whether participants with greater metacognitive knowledge made smaller errors than participants with worse metacognitive knowledge. To do so, we calculated the difference in enjoyment between the song the participant rated as most enjoyable and the song they chose during the CBC task. This metric – which we will refer to as *error magnitude* – captures how much enjoyment the participant would have lost by listening to the song they chose during the CBC task, rather than the song they enjoyed most. Excluding participants who had ties in their ratings (*n* = 13), the mean error magnitude among participants who made an error was 26.69 points (*sd* = 20.09). A simple linear regression (Model 1) indicated that participants with greater Euclidean distances had greater error magnitudes (*B* = 30.44, *SE* = 12.14, *p* = 0.01), though the correlation was weak (*r* = .22; see Figure 4). This suggests that greater metacognitive knowledge as measured by the *KoW* paradigm is associated with making choices that result in greater utility maximization.

Our pre-registration did not specify that this analysis would look only at the participants who made errors, so we ran several robustness checks to demonstrate that the effect holds under other specifications. First, we re-ran the regression excluding participants who had heard any of the three songs before (Model 2). The effect persisted and remained significant (*B* = 29.83, *SE* = 13.11, *p* = 0.02, *r* = .21). Next, we re-ran the regression controlling for age, college education, hours spent listening to music each day, and whether the participant likes pop music (Model 3). The positive relationship between Euclidean distance and error magnitude remained significant (*B* = 30.08, *SE* = 12.56, *p* = .02). None of the covariates were significant predictors of error magnitude (*p*s > 0.37). Next, we re-ran the regression, this time excluding participants whose error magnitudes were more than three standard deviations above the mean (*n* = 2; Model 4). The positive relationship between Euclidean distance and error magnitude remained significant (*B* = 22.56, *SE* = 11.30, *p* = .048, *r* = .18). Finally, we re-ran the regression including participants who successfully rated the song they chose during the CBC as most enjoyable, assigning them an error magnitude of 0 (Model 5). The positive relationship between Euclidean distance and error magnitude persisted (*B* = 26.81, *SE* = 9.70, *p* = .01, *r* = .19). Full regressions and additional robustness checks are reported in the Supplemental Materials.

As another robustness check, we calculated a separate *multiple choice error magnitude* by taking the difference between the enjoyment rating for the song chosen during the post-listening multiple choice task and the enjoyment rating for the song chosen during the CBC task. Excluding participants who made the same choice both times, the average multiple choice error magnitude was 24.54 (*sd* = 23.12). Notably, this included nine participants who enjoyed the song they chose during the multiple-choice task *less* than the song they chose during the CBC task, resulting in a negative multiple-choice error magnitude. Excluding these participants, the mean rose to 27.88 (*sd* = 20.09). Regardless of whether these participants were included (*B* = 42.69, *SE* = 14.01, *p* = .002, *r* = .26) or excluded (*B* = 31.36, *SE* = 12.64, *p* = .01, *r* = .23), simple linear regressions indicated that participants with greater Euclidean distances had greater multiple choice error magnitudes, further supporting our hypothesis that individuals with greater metacognitive knowledge make choices that they are happier with. This effect also held when participants who made the same choice both times (multiple choice error magnitude = 0) were included in the regression (*B* = 27.99, *SE* = 10.04, *p* = .01, *r* = .19).

Study 3 Discussion

The primary finding from Study 3 was that participants with better metacognitive knowledge – as measured by the *KoW* paradigm – make choices that they are happier with. While our results were insufficiently powered to demonstrate that participants with greater metacognitive knowledge were more likely to make the *best* possible choice, we demonstrated that decision makers with greater metacognitive knowledge made smaller errors, thus minimizing the reduction in utility they experience from making a mistake. In all, Study 3 highlights the importance of metacognition in subjective decision making and provides suggestive evidence of the predictive validity of the *KoW* paradigm. However, given that this conclusion was built partially on exploratory findings, the results may be considered weaker than the results from Studies 1 and 2. As such, we encourage future researchers to replicate this study and further assess the predictive validity of the *KoW* paradigm.

General Discussion

Many of the most important decisions that we make – such as whom to marry, which home to purchase, and which college to attend – are subjective and multi-attribute in nature. To make these choices effectively, it is critical for decision makers to know how important the various attributes by which the alternatives vary are to them. Having this knowledge allows decision makers to weight the attributes appropriately in their decision-making processes (Keeney & Raiffa, 1976; Soman, 2004) and accurately communicate their preferences to others (e.g., Slovic & Lichtenstein, 1971; Nisbett & Wilson, 1977). Developing and maintaining this explicit knowledge requires a decision maker to actively monitor their beliefs, values, and decision-making processes, and thus can be considered a metacognitive task (Dunlosky & Metcalfe, 2009; Flavell, 1979; McCormick, 2003). However, the extant metareasoning literature has largely focused on objective decisions, not subjective choices (Ackerman & Thompson, 2017) and therefore lacks methods for assessing metacognitive knowledge of attribute weights in subjective decisions.

To fill this gap in the literature, we created the novel *KoW* paradigm that allows for the assessment of metacognitive knowledge in subjective, multi-attribute choice. In the studies presented here, we demonstrated that the *KoW* paradigm generates measures of metacognitive knowledge that are reliable (Study 1), resistant to practice effects (Study 1), consistent across domains (Study 2), sensitive to known increases in metacognitive knowledge (Study 2) and predictive of choice satisfaction (Study 3), though the evidence for predictive validity was weaker than the evidence for reliability and consistency across domains. Together, these results provide evidence that the *KoW* paradigm is a reliable and reasonably valid method of assessing metacognitive knowledge in subjective multi-attribute choice. The *KoW* paradigm thus represents a useful methodological addition to the metacognitive toolkit.

Comparison to Existing Approaches

Though the *KoW* paradigm fills a novel experimental niche, it shares similarities to other empirical approaches that are pervasive in the literature. Here, we will discuss some of these similar approaches and highlight the novelty of the *KoW* paradigm. The most obvious point of comparison for the *KoW* paradigm is the line of scholarship – primarily in decision science, economics, and marketing – demonstrating that individuals' preferences are inconsistent across measurement modalities (e.g., Borcherding et al., 1991; Pöyhöyen & Hämäläinen, 2001; Suk & Yoon, 2012). One common finding in this literature is that individuals' stated preferences are inconsistent with their revealed preferences (e.g., Barlas, 2003; Harte & Koele; 1995; Heeler et al., 1979; Riquelme, 2001). While the *KoW* paradigm inherently relies on comparisons of stated and revealed preferences (measured via decision weights), it is unique from the existing literature in that it does not seek to elicit multiple distinct measures of participants' decision weights, but rather explicitly instructs participants to articulate the weights that they believe reflect their choice behavior (i.e., their revealed weights). This nuance transforms the approach from a measure of behavioral (in)consistency to a measure of participants' explicit knowledge of how they made their choices. This unique emphasis on decision makers' explicit knowledge of their own decision processes is what makes *KoW* a novel metacognitive paradigm (Dunlosky & Metcalfe, 2009; Flavell, 1979; McCormick, 2003).

The *KoW* paradigm also shares similarities to psychological lens model paradigms that compare judges' weighting of cues (i.e., cue utilization) to the true validity of each cue as a predictor of an underlying construct (Brunswik, 1952; Hammond 1955; Karelaia & Hogarth, 2008; Nestler & Back, 2013). However, unlike the lens model which compares cue utilization to cue validity, the *KoW* paradigm compares participants' beliefs about their cue utilization to their revealed cue utilization. As such, the *KoW* paradigm can be thought of as a metacognitive variant of the lens model in which both sides of the model are generated by the judge and alignment is driven by metacognitive knowledge (Dunlosky & Metcalfe, 2009; Flavell, 1979; McCormick, 2003)

In a similar vein, a recent paper by Ackerman (2023) introduced the *BEVoCI* method, a lens model-based approach that compares the influence that various factors have on participants' task performance to the influence that the same factors have on their metacognitive judgments. *BEVoCI* is similar to the *KoW* paradigm in that it uses cue weights to evaluate participants' metacognitive judgments, but the two paradigms have very different goals and approaches. *BEVoCI* leverages within-participant variability in success and confidence (or other metacognitive judgments) across similar items, with the goal of untangling various potential sources of bias. In contrast, The *KoW* paradigm leverages differences in stated and revealed cue weights generated by individual participants to quantify each participant's knowledge of how they make multi-attribute choices. While the two paradigms have clear synergies, they each provide distinct contributions to the metareasoning literature.

Another related empirical strategy that is commonly used in the metacognition literature is cue integration. In this approach, multiple unique cues are experimentally manipulated to assess whether and to what extent each cue affects participants' performance and their metacognitive judgments (e.g., Jang & Nelson, 20005; Koriat, 1997; Undorf et al., 2018, 2020). For example, in a study of word learning, Undorf et al. (2018) systematically manipulated different factors (e.g., repetitions, font size) to assess which factors affected participants' performance and Judgments of Learning (JoLs). Like *KoW*, the cue integration approach can be used to compare the cues that affect participants' behavior (performance for JoLs, choice for *KoW*) to a metacognitive judgment about those cues. However, *KoW* differs from cue integration in that it asks participants to explicitly describe their beliefs about how the cues affected their behavior, rather than using variation in the cues to predict an intermediary metacognitive judgment, such as JoLs. This is necessary because *KoW* is designed to be implemented for subjective decisions, which do not have objective performance metrics to which typical metacognitive judgments, such as JoLs, can be compared. Cue integration and *KoW* should be considered as complementary methods.

Use Cases for the *KoW* Paradigm

To inspire future scholars to adopt the *KoW* paradigm, we will now highlight several ways in which the *KoW* paradigm can be used to better understand how decision makers engage in metareasoning. First, we can use the *KoW* paradigm to identify characteristics of individuals or groups that are predictive of metacognitive knowledge. For example, developmental psychologists may use the *KoW* paradigm to explore the development of metacognitive knowledge throughout the lifespan (Metcalfe et al., 2010) or political psychologists may use the *KoW* paradigm to compare the metacognitive knowledge of Republicans and Democrats in voting contexts (Anson, 2018). The *KoW* paradigm may also be used to evaluate how metacognitive knowledge covaries with other individual difference metrics, such as intelligence (Ohtani & Hisasaka, 2018), need for cognition (Coutinho et al., 2005), creativity (Kaufman et al., 2016), or personality traits (Bidjerano & Dai, 2007). We believe it will be of particular interest for future research to investigate the ability of the *KoW* paradigm to discriminate between general intelligence and metacognitive knowledge.

Second, the *KoW* paradigm can be used to explore how metacognitive knowledge varies across decision making domains. For example, marketers may use the *KoW* paradigm to compare consumers' metacognitive knowledge across categories of goods and services (Schwarz, 2004). Scholars may also be interested in diving deeply into individual domains that are of substantial individual or societal importance. For example, public policy scholars may be interested in using the *KoW* paradigm to study how metacognitive knowledge of voting preferences influences election outcomes (Rollwage et al., 2018) or finance scholars may use the *KoW* paradigm to assess how metacognitive knowledge influences household budgeting (Sunderaraman et al., 2020).

Third, the *KoW* paradigm can be used to explore contextual factors that influence decision makers' metacognitive knowledge. For example, educational psychologists may use the *KoW* paradigm to study how classroom environments influence the development of metacognitive knowledge among students (Callender et al., 2016) or decision scientists may use it to evaluate how aspects of the decision environment, such as the number of alternatives, impact metacognitive knowledge (Hadar et al., 2014). Scholars can also use the *KoW* paradigm to assess the efficacy of various interventions – such as mindfulness (Vickery & Dorjee, 2016) and numeracy interventions (Muncer et al., 2022) – that may improve metacognitive knowledge.

Fourth, the *KoW* paradigm can be used to study the interplay between individual differences and the decision environment. For example, social scientists may use the *KoW* paradigm to study how metacognitive knowledge covaries with participants' domain-specific motivation to make a good decision (Efklides, 2001). Experimenters could also manipulate motivation using incentives (Miller & Geraci, 2011). Similarly, scholars may be interested in evaluating how familiarity with or expertise in a domain (Veenman & Elshout, 1999) may correlate with metacognitive knowledge. As an example, a recent paper using a preliminary version of the *KoW* paradigm demonstrated that parents of high-school aged children had no greater metacognitive knowledge of the weights they placed on various attributes when comparing highs schools than a convenience sample of parents and non-parents, suggesting that familiarity with a domain is not necessarily associated with greater metacognitive knowledge, at least in the domain of school choice (Cash & Oppenheimer, 2024). Researchers could also experimentally manipulate participant expertise through training (Batha & Carroll, 2007). Studies of this nature will help us to better understand the sensitivity of the *KoW* paradigm.

Fifth and finally, the *KoW* paradigm can be used to explore the psychological mechanisms underlying metacognitive knowledge. For example, the *KoW* paradigm could be systematically altered to answer theoretical questions about the role of metacognitive monitoring (Ackerman, 2014; De Neys et al., 2011), metacognitive control (Ackerman et al., 2020; De Neys et al., 2013), and top-down knowledge (Sherman et al., 2015) in the development and deployment of metacognitive knowledge. These examples highlight only a small subset of the potential empirical questions that could be explored using the *KoW* paradigm, but we hope that they inspire creative applications of the paradigm in a wide variety of domains.

Limitations

One limitation of the *KoW* paradigm is that it is forced to estimate weights for each attribute, and thus may generate less-accurate RAWs for participants who put most or all of their weight on one attribute (e.g., Newell & Shanks, 2003). The primary concern is that, mathematically, RAWs tend to be condensed away from extremely high (e.g., 100%) and extremely low (e.g., 0%) weights because the underlying utilities are estimated using Bayesian models that treat sample mean utilities as priors. Across studies, the lowest RAW estimated for any attribute was 0.53% and the highest RAW was 70.18%. This concern is unlikely to be problematic in the present studies, as only 8.14% of participants in the studies presented here reported a SAW of greater than 70% for any attribute, suggesting that strong non-compensatory strategies were relatively uncommon. However, future researchers using the *KoW* paradigm should consider methods for better assessing non-compensatory strategies. One potentially fruitful avenue for doing so is to increase the number of choices participants make to boost the amount of evidence available to update away from sample mean priors.

A second and related limitation of the *KoW* paradigm is that it assumes that decision makers are using weights at all, which is not necessarily true of all participants. Some decision makers may use non-weight-based decision strategies. For example, participants may use fast-and-frugal heuristic-based strategies (Gigerenzer et al., 1999; but see Krefeld-Schwalb et al, 2019; Oppenheimer, 2003 for criticisms of this approach), lexicographic decision rules (Fishburn, 1974) or other satisficing approaches (Simon, 1956). Participants may also make decisions based on more idiosyncratic factors – such as intuitive reactions to familiar alternatives (Klein, 1993; 2015), holistic judgments about each alternative (Arkes et al., 2010), or unique preferences about combinations of attributes (e.g., I will accept a small house if it has a big yard). The first concern raised by the possibility of participants using non-weight-based strategies is that the *KoW* paradigm may not generate RAWs that accurately reflect these decision makers' choices. However, we consistently found that RAWs were able to predict more than 90% of participants' choices. This suggests that most participants made choices that could be described using weights – even if that's not necessarily how they thought they were making their choices.

Another concern related to this limitation is that participants who made non-weight-based decisions may have had metacognitive knowledge of how they made their choices, but struggled to provide meaningful SAWs because they were not thinking in terms of weights. It is certainly possible that some participants had this challenge, but two pieces of evidence suggest that it was not a widespread problem. First, we consistently found that average tRAW-tSAW correlations were about the same as average tRAW-AIR correlations. If participants were engaging in non-weight-based decision strategies, the non-weight-based AIRs should be easier to accurately self-report than the weight-based SAWs. Second, participants were highly confident in the accuracy of their SAWs. If participants were completely blindsided by the concept of decision weights, they would likely report very low confidence in the accuracy of their SAWs.

In general, this evidence suggests that participants in our studies were using weight-based strategies – or at least were able to approximately translate the strategies they were using into weights. This is consistent with the extant JDM literature, which shows that participants often do make multi-attribute choices using decision weights (Huber, 1974; Keeney & Raiffa, 1976; Soman, 2004; Weiss et al., 2010). However, we strongly encourage future researchers to adapt the *KoW* paradigm to better identify and account for participants who use non-weight-based strategies, especially in contexts in which non-weighting strategies are known to be more prevalent, such as when participants are under time pressure (Böckenholt & Kroeger, 1993), experiencing a high cognitive load (Deck & Jahedi, 2015), or are in affective states that discourage deliberation (Lewinsohn & Mano, 1993). It may also be worth exploring whether simply warning participants that they will be asked to self-report decisions weights or allowing participants to complete a free response item describing their decision process might influence how well *KoW* captures cue weighting.

Conclusion

Across three studies, we presented and validated the *KoW* paradigm, a novel method for assessing metacognitive knowledge of attribute weights in subjective, multi-attribute choice decisions. Evidence for reliability and consistency across domains was strong, while evidence of predictive validity was slightly weaker. The *KoW* paradigm is unique from existing metareasoning paradigms in that it does not require participants' metacognitive judgments to be compared to an objectively correct answer, thus opening the door to metareasoning research in subjective decision domains. Given the prevalence of such decisions in our daily lives, these domains have significant impacts on our overall well-being and are worthy of study. The *KoW* paradigm has numerous applications for studying metareasoning across a wide variety of domains. When you judge how interesting this paper is, take a minute to think… do you really know what makes a paper interesting to you? If you don't think this paper is interesting, then maybe it's time to reconsider your weights.

Acknowledgements

This work was supported by a Graduate Student Small Grant from Carnegie Mellon University's Center for Behavioral and Decision Research, 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. 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 declare no conflicts of interest. We would like to thank our research assistants – Jannie Du, Aleksaundra Handrinos, Piper Ligotti, Claire Liu, Brett Nyman, Joanna Tan, and Jess Zheng – who helped run the simulations. We would also like to thank the first author's committee members – Sharon Carver, Julie Downs, and Bonnie Nozari – for their feedback throughout the research process.

Funding: This research was supported by a Graduate Student Small Grant from the Center for Behavioral and Decision Research at Carnegie Mellon University and a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (Award #2333553). Free use of the Lighthouse Studio software was awarded via an Academic Research Grant from Sawtooth Software, Inc. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the grantors. Neither the grant from Carnegie Mellon University nor the grant from Sawtooth Software, Inc. was assigned a grant number.

Conflicts of Interest: The authors report that there are no competing interests to declare.

Data Availability Statement: Data, materials, and code for all studies presented in this manuscript are available at https://doi.org/10.17605/OSF.IO/UZQK5

References

Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General, 143*(3), 1349-1368. https://doi.org/10.1037/a0035098

Ackerman, R. (2023). Bird's-eye view of cue integration: Exposing instructional and task Design factors which bias problem solvers. *Educational Psychology Review, 35*(2), Article 55. https://doi.org/10.1007/s10648-023-09771-z

Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgements during reasoning and memorisation. *Thinking & Reasoning, 23*(4), 376–408. https://doi.org/10.1080/13546783.2017.1328373

Ackerman, R., Douven, I., Elqayam, S., & Teodorescu, K. (2020). Satisficing, meta-reasoning, and the rationality of further deliberation. In S. Elqayam, I. Douven, J. St. B. T. Evans, & N. Cruz (Eds.), *Logic and Uncertainty in the Human Mind.* Routledge. https://doi.org/10.4324/9781315111902-2

Ackerman, R., & Thompson, V. A. (2015). Meta-reasoning: What can we learn from meta-memory? In A. Feeney & V. A. Thompson (Eds.), *Reasoning as memory* (pp. 164–182). Psychology Press.

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

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139–160.

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26*(2), 147-179. https://doi.org/10.1207/s15516709cog2602_1

Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research, 32*(2), 152-162. https://doi.org/10.2307/3152044

Analytics Simplified Pty Ltd (2023). *Conjointly*. https://conjointly.com/

Anson, I. G. (2018). Partisanship, political knowledge, and the Dunning-Kruger effect. *Political Psychology, 39*(5), 1173-1192. https://doi.org/10.1111/pops.12490

Antonietti, A., Ignazi, S., & Perego, P. (2010). Metacognitive knowledge about problem-solving methods. *British Journal of Educational Psychology, 70*(1), 1-16. https://doi.org/10.1348/000709900157921

Arkes, H. R., Gonzálaez-Vallejo, C., Bonham, A. J., Kung, Y.-H., & Bailey, N. (2010). Assessing the merits and faults of holistic and disaggregated judgments. *Journal of Behavioral Decision Making, 23*(3), 250–270. https://doi.org/10.1002/bdm.655

Barlas, S. (2003). When choices give in to temptations: Explaining the disagreement among importance measures. *Organizational Behavior and Human Decision Processes, 91*(2), 310-321. https://doi.org/10.1016/S0749-5978(02)00515-0

Basu, S., & Dixit, S. (2022). Role of metacognition in explaining decision-making styles: A study of knowledge about cognition and regulation of cognition. *Personality and Individual Differences, 185*, Article 111318. https://doi.org/10.1016/j.paid.2021.111318

Batha, K., & Carroll, M. (2007). Metacognitive training aids decision making. *Australian Journal of Psychology, 59*(2), 64-69. https://doi.org/10.1080/00049530601148371

Berman, J. Z., Barasch, A., Levine, E. E., & Small, D. A. (2018). Impediments to Effective Altruism: The Role of Subjective Preferences in Charitable Giving. *Psychological Science, 29*(5), 834-844. https://doi.org/10.1177/0956797617747648

Bidjerano, T., & Dai, D. Y. (2007). The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and Individual Differences, 17*(1), 69-81. https://doi.org/10.1016/j.lindif.2007.02.001

Böckenholt, U., & Kroeger, K. (1993). The effect of time pressure in multiattribute binary choice tasks. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 195–214). Plenum Press. https://doi.org/10.1007/978-1-4757-6846-6_14

Borcherding, K., Eppel, T., & von Winterfelt, D. (1991). Comparison of weighting judgments in multiattribute utility measurement. *Management Science, 37*(12), 1513-1654. https://doi.org/10.1287/mnsc.37.12.1603

Bradley, M., & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation, 21,* 167-184. https://doi.org/10.1007/bf01098791

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Erlbaum.

Brunswik, E. (1952). *The conceptual framework of psychology.* University of Chicago Press.

Buelow, M. T., & Barnhart, W. R. (2018). Test–retest reliability of common behavioral decision making tasks. *Archives of Clinical Neuropsychology, 33*(1), 125–129. https://doi.org/10.1093/arclin/acx038

Callender, A.A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*, 215-235. https://doi.org/10.1007/s11409-015-9142-6

Cary, M., & Reder, L. M. (2002). Metacognition in strategy selection: Giving consciousness too much credit. In P. Chambres, M. Izaute, & P.-J., Marescaux (Eds.), *Metacognition: Process, Function, and Use.* Springer. https://doi.org/10.1007/978-1-4615-1099-4_5

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

Chua, E. F., Schacter, D. L., & Sperlin, R. A. (2009). Neural correlates of metamemory: A comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience, 21*(9), 1751-1765. https://doi.org/10.1162/jocn.2009.21123

Colombo, B., Iannello, P., & Antonietti, A. (2010). Metacognitive knowledge of decision-making: An explorative study. In A. Efklides & P. Misailidi (Eds.), *Trends and prospects* *in metacognition research* (pp. 445–472). Springer Science + Business Media. https://doi.org/10.1007/978-1-4419-6546-2_20

Coutinho, S., Wiemer-Hastings, K., Skowronski, J. J., & Britt, M. A. (2005). Metacognition, need for cognition, and use of explanations during ongoing learning and problem solving. *Learning and Individual Differences, 15*(4), 321-337. https://doi.org/10.1016/j.lindif.2005.06.001

Davis, E. L., Levine, L. J., Lench, H. C., & Quas, J. A. (2010). Metacognitive emotion regulation: Children's awareness that changing thoughts and goals can alleviate negative emotions. *Emotion, 10*(4), 498–510. https://doi.org/10.1037/a0018428

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE,* 6(1), Article e15954. https://doi.org/10.1371/journal.pone.0015954

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20*(2), 269-273. https://doi.org/10.3758/s13423-013-0384-5

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13-21. https://doi.org/10.1016/j.intell.2006.02.001

Deck, C., & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review, 78,* 97-119. https://doi.org/10.1016/j.euroecorev.2015.05.004

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging, 22*(1), 122–133. https://doi.org/10.1037/0882-7974.22.1.122

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2014). Self-control in school-age children. *Educational Psychologist, 49*(3), 199-217. https://doi.org/10.1080/00461520.2014.926225

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

Efklides, A. (2001). Metacognitive experiences in problem solving: Metacognition, motivation, and self-regulation. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and prospects in motivation research* (pp. 297–323). Kluwer Academic Publishers.

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. https://doi.org/10.1007/978-3-319-57413-4_23

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Fernandez-Cruz, A. L., Arrango-Muñoz, S., & Volz, K. G. (2016). Oops, scratch that! Monitoring one’s own errors during mental calculation. *Cognition, 146*, 110-120. https://doi.org/10.1016/j.cognition.2015.09.005

Fishburn, P. C. (1974). Lexicographic orders, utilities and decision rules: A survey. *Management Science, 20*(11), 1442-1471.

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision, 13*, 139-199. https://doi.org/10.1007/BF00134215

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., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, Article 443. https://doi.org/10.3389/fnhum.2014.00443

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology, 63*, 287-313. https://doi.org/10.1146/annurev-psych-120710-100449

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning – in search of a phenomenon. *Thinking & Reasoning, 21*(4), 383-396. https://doi.org/10.1080/13546783.2014.980755

George, T., & Mielicki, M. K. (2023). Bullshit receptivity, problem solving, and metacognition: Simply the BS, not better than all the rest. *Thinking & Reasoning, 29*(2), 213-249. https://doi.org/10.1080/13546783.2022.2066724

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making, 9*(1), 15–34.

Gigerenzer, G., Todd, P. M., & The ABC Research Group. (1999). *Simple heuristics that make us smart.* Oxford University Press.

Hadar, L., & Sood, S. (2014). When knowledge is demotivating: Subjective knowledge and choice overload. *Psychological Science, 25*(9), 1739-1747. https://doi.org/10.1177/0956797614539165

Haddara, N., & Rahnev, D. (2022). The impact of feedback on perceptual decision-making and metacognition: Reduction in bias but no change in sensitivity. *Psychological Science, 33*(2), 259-275. https://doi.org/10.1177/09567976211032887

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62*(4), 255–262. https://doi.org/10.1037/h0046845

Harte, J. M., & Koele, P. (1995). A comparison of different methods for the elicitation of attribute weights: Structural modeling, process tracing, and self-reports. *Organizational Behavior and Human Decision Processes, 64*(1), 49-64. https://doi.org/10.1006/obhd.1995.1089

Heeler, R. M., Okechuku, C., & Reid, S. (1979). Attribute importance: Contrasting measurements. *Journal of Marketing Research, 16*(1), 60-63. https://doi.org/10.2307/3150875

Helveston, J. (2023). *cbcTools* (Version 0.5.0) [R Package]. https://cran.r-project.org/web/packages/cbcTools/cbcTools.pdf

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,* 1-36. https://doi.org/10.1007/s11573-019-00927-4

Hirshleifer, D. Levi, Y., Lourie, B., & Teoh, S. H. (2019). Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics, 133*(1), 83-98. https://doi.org/10.1016/j.jfineco.2019.01.005

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

Hu, X., Luo, L., & Fleming, S. M. (2019). A role for metamemory in cognitive offloading. *Cognition, 193*, Article 104012. https://doi.org/10.1016/j.cognition.2019.104012

Huber, G. P. (1974). Multi-Attribute Utility Models: A Review of Field and Field-Like Studies. *Management Science, 20*(10), 1323-1411. https://doi.org/10.1287/mnsc.20.10.1393

Jackson, S. A., Kleitman, S., Howie, P., & Stankov, L. (2016). Cognitive abilities, monitoring confidence, and control thresholds explain individual differences in heuristics and biases. *Frontiers in Psychology, 7*, Article 1559. https://doi.org/10.3389/fpsyg.2016.01559

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2017). Individual differences in decision making depend on cognitive abilities, monitoring and control. *Journal of Behavioral Decision Making, 30*(2), 209-223. https://doi.org/10.1002/bdm.1939

Jang, Y., & Nelson, T. O. (2005). How Many Dimensions Underlie Judgments of Learning and Recall? Evidence From State-Trace Methodology. *Journal of Experimental Psychology: General, 134*(3), 308–326. https://doi.org/10.1037/0096-3445.134.3.308

Jia, X., Li, W., & Cao, L. (2019). The role of metacognitive components in creative thinking. *Frontiers in Psychology, 10*, Article 2404. https://doi.org/10.3389/fpsyg.2019.02404

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469–486. https://doi.org/10.1037/a0017341

Kaufman, J. C., Beghetto, R. A., & Watson, C. (2016). Creative metacognition and self-ratings of creative performance: A 4-C perspective. *Learning and Individual Differences, 51*, 394-399. https://doi.org/10.1016/j.lindif.2015.05.004

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. Cambridge University Press.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138–147). Ablex Publishing.

Klein, G. A. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition, 4*(3), 164-168. https://doi.org/10.1016/j.jarmac.2015.07.001

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. (2015). Metacognition: Decision making processes in self-monitoring and self-regulation. In G. Keren & G. Wu (Eds.), *Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 356-379). Wiley-Blackwell. https://doi.org/10.1002/9781118468333.ch12

Krefeld-Schwalb, A., Donkin, C., Newell, B. R., & Scheibehenne, B. (2019). Empirical comparison of the adjustable spanner and the adaptive toolbox models of choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(7), 1151–1165. https://doi.org/10.1037/xlm0000641

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, 10(1), 117094. https://doi.org/10.1525/collabra.117094

Law, M. K. H., Stankov, L., & Kleitman, S. (2022). I choose to opt-out of answering: Individual differences in giving up behaviour on cognitive tests. *Journal of Intelligence, 10*(4), 86. https://doi.org/10.3390/jintelligence10040086

Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science, 15*(2), 173-191. https://doi.org/10.1287/mksc.15.2.173

Lewinsohn, S., & Mano, H. (1993). Multi-attribute choice and affect: The influence of naturally occurring and manipulated moods on choice processes. *Journal of Behavioral Decision Making, 6*(1), 33–51. https://doi.org/10.1002/bdm.3960060103

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review, 124*(6), 762–794. https://doi.org/10.1037/rev0000075

Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research, 20*(4), 350-367. https://doi.org/10.2307/3151440

Macaluso, J. A., Beuford, R. R., & Fraundorf, S. H. (2022). Familiar strategies feel fluent: The role of study strategy familiarity in the misinterpreted-effort model of self-regulated learning. *Journal of Intelligence, 10*(4), 83. https://doi.org/10.3390/jintelligence10040083

McCormick, C. B. (2003). Metacognition and learning. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of psychology: Educational psychology* (Vol. 7, pp. 79–102). John Wiley & Sons, Inc. https://doi.org/10.1002/0471264385.wei0705

Miami Sound Machine (1984). Prisoner of Love [Song]. On *Eyes of Innocence*. Epic Records.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

Muncer, G., Higham, P. A., Gosling, C. J., Cortese, S., Wood-Downie, H., & Hadwin, J. A. (2022). A meta-analysis investigating the association between metacognition and math performance in adolescence. *Educational Psychology Review, 34*, 301-334. https://doi.org/10.1007/s10648-021-09620-x

Metcalfe, J., Eich, T. S., & Castel, A. D. (2010). Metacognition of agency across the lifespan. *Cognition, 116*(2), 267-282. https://doi.org/10.1016/j.cognition.2010.05.009

Nestler, S., & Back, M. D. (2013). Applications and Extensions of the Lens Model to Understand Interpersonal Judgments at Zero Acquaintance. *Current Directions in Psychological Science*, *22*(5), 374-379. https://doi.org/10.1177/0963721413486148

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 53–65. https://doi.org/10.1037/0278-7393.29.1.53

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

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: a meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning, 13*, 179-212. https://doi.org/10.1007/s11409-018-9183-8

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition, 90*(1), B1–B9. https://doi.org/10.1016/S0010-0277(03)00141-0

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

Payne, S. J., & Duggan, G. B. (2011). Giving up problem solving. *Memory & Cognition, 39*(5), 902-913. https://doi.org/10.3758/s13421-010-0068-6

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review, 24*, 1774-1784. https://doi.org/10.3758/s13423-017-1242-7

Perry, J., Lundie, D., & Golder, G. (2019). Metacognition in schools: What does the literature suggest about the effectiveness of teaching metacognition in schools? *Educational Review, 71*(4), 483-500. https://doi.org/10.1080/00131911.2018.1441127

Petty, R. E., Briñol, P., Tormala, Z. L., & Wegener, D. T. (2007). The role of metacognition in social judgment. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 254–284). The Guilford Press.

Pintrich, P. R. (20020> The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice, 41*(4), 219-225. https://doi.org/10.1207/s15430421tip4104_3

Pöyhöyen, M., & Hämäläinen, R. P. (2001). On the convergence of multiattribute weighting methods. *European Journal of Operational Research, 129*(3), 569-585. https://doi.org/10.1016/S0377-2217(99)00467-1

Rigby, D., Burton, M., & Pluske, J. (2016). Preference Stability and Choice Consistency in Discrete Choice Experiments. *Environmental and Resource Economics, 65*, 441-461. https://doi.org/10.1007/s10640-015-9913-1

Riquelme, H. (2001). Do consumers know what they want? *Journal of Consumer Marketing, 18*(5), 437-448. https://doi.org/10.1108/07363760110398772

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology, 28,* 4014–4021. https://doi.org/10.1016/j.cub.2018.10.053

Ruben, J. (2018). First Day of Summer [Song]. On *Hope.* MCA Music.

Sawtooth Software, Inc. (2017). *The CBC system for choice-based conjoint analysis V9.* https://sawtoothsoftware.com/resources/technical-papers/cbc-technical-paper

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. (2023). *Lighthouse Studio* (Version 9.14.2) [Computer software]. https://sawtoothsoftware.com/lighthouse-studio

Schwarz, N. (2008). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*(4), 332-348. https://doi.org/10.1207/s15327663jcp1404_2

Sherman, M. T., Seth, A. K., Barrett, A. B., & Kanai, R. (2015). Prior expectations facilitate metacognition for perceptual decision. *Consciousness and Cognition, 35*, 53-65. https://doi.org/10.1016/j.concog.2015.04.015

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138. https://doi.org/10.1037/h0042769

Slick, G. (1980). Seasons [Song]. On *Dreams*. RCA Records.

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., Flessner, D., & Bauman, W. S. (1972). Analyzing the use of information in investment decision making. A methodological proposal. *The Journal of Business, 45*(2), 283-301.

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

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

Soman, D. (2004). The effect of time delay on multi-attribute choice. *Journal of Economic Psychology, 25*(2), 153-175. https://doi.org/10.1016/j.joep.2003.09.002

Spiller, S. A., & Belogolova, L. (2017). On Consumer Beliefs about Quality and Taste. *Journal of Consumer Research, 43*(6), 970-991. https://doi.org/10.1093/jcr/ucw065

Spotify, Inc. (2023). *Spotify for Developers.* https://developer.spotify.com/documentation/web-api

Stanovich, K. E., & Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence, 11*(2), 27. https://doi.org/10.3390/jintelligence11020027

Suk, K., & Yoon, S.-O. (2012). The moderating role of decision task goals in attribute weight convergence. *Organizational Behavior and Human Decision Processes, 118*(1), 37-45. https://doi.org/10.1016/j.obhdp.2011.12.002

Sunderaraman, P., Chapman, S., Barker, M. S., & Cosentino, S. (2020). Self-awareness for financial decision-making abilities in healthy adults. *PLoS One*, *15*(7), Article e0235558. https://doi.org/10.1371/journal.pone.0235558

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday & Co.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107-140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, V. A., Prowse Turner, J. A., 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

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

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 van der Plas, E., Zhang, S., Dong, K., Bang, D., Li, J., Wright, N. D., & Fleming, S. M. (2022). Identifying cultural differences in metacognition. *Journal of Experimental Psychology: General, 151*(12), 3268–3280. https://doi.org/10.1037/xge0001209

Veenman, M., & Elshout, J. J. (1999). Changes in the relation between cognitive and metacognitive skills during the acquisition of expertise. *European Journal of Psychology of Education, 14*(4), 509–523. https://doi.org/10.1007/BF03172976

Vega, S., Mata, A., Ferreira, M. B., & Vaz, A. R. (2021). Metacognition in moral decisions: Judgment extremity and feeling of rightness in moral intuitions. *Thinking & Reasoning, 27*(1), 124-141. https://doi.org/10.1080/13546783.2020.1741448

Vickery, C. E., & Dorjee, D. (2016). Mindfulness training in primary schools decreases negative affect and increases meta-cognition in children. *Frontiers in Psychology, 6*, Article 2025. https://doi.org/10.3389/fpsyg.2015.02025

Vrugt, A., & Oort, F. J. (2008). Metacognition, achievement goals, study strategies and academic achievement: Pathways to achievement. *Metacognition and Learning, 3*(2), 123–146. https://doi.org/10.1007/s11409-008-9022-4

Weber, C. R., & Federico, C. M. (2012). Moral Foundations and Heterogeneity in Ideological Preferences. *Political Psychology, 34*(1), 107-126. https://doi.org/10.1111/j.1467-9221.2012.00922.x

Weber, E. K., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15*, 263-290. https://doi.org/10.1002/bdm.414

Weiss, J. W., Weiss, D. J., & Edwards, W. (2010). A descriptive multi-attribute utility model for everyday decisions. *Theory and Decision, 68*(1-2), 101–114. https://doi.org/10.1007/s11238-009-9155-1

Wilson, T. D. C., & Nisbett, R. E. (1978). The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology, 41*(2), 118-131. https://doi.org/10.2307/3033572

Wright, P. (2002). Marketplace metacognition and social intelligence. *Journal of Consumer Research, 28*(4), 677-682. https://doi.org/10.1086/338210

Zepeda, C. D., Richey, J. E., Ronevich, P., & Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology, 107*(4), 954–970. https://doi.org/10.1037/edu0000022

Table 1: Summary of Reliability Metrics

Sample-Level Metrics	**Phase 1**	**Phase 2**	***p*** **(Difference)**
Average tRAW-tSAW Correlation[5]	*r* = .54	*r* = .53	.92[1]
Average tRAW-AIR Correlation[5]	*r* = .50	*r* = .52	.75[1]
Confidence in SAWs	*M* = 80.46 *sd* = 17.00	*M* = 80.60 *sd* = 16.99	.84[2]
Confidence in AIRs	*M* = 83.20 *sd* = 15.38	*M* = 83.09 *sd* = 15.05	.87[2]
Different Choices Predicted by RAWs and SAWs (% of Tasks)	15.72%	16.16%	.65[3]
Euclidean Distance	*M* = 0.34 *sd* = 0.14	*M* = 0.36 *sd* = 0.15	.08[2]
Participant-Level Metrics	**Correlation Between Phase 1 and Phase 2**		***p*** **(Correlation)**
tSAWs[5]	.76		< .001[4]
AIRs[5]	.78		< .001[4]
tRAWs[5]	.47		< .001[4]
Confidence in SAWs	.81		< .001[4]
Confidence in AIRs	.77		< .001[4]
Euclidean Distances	.59		< .001[4]

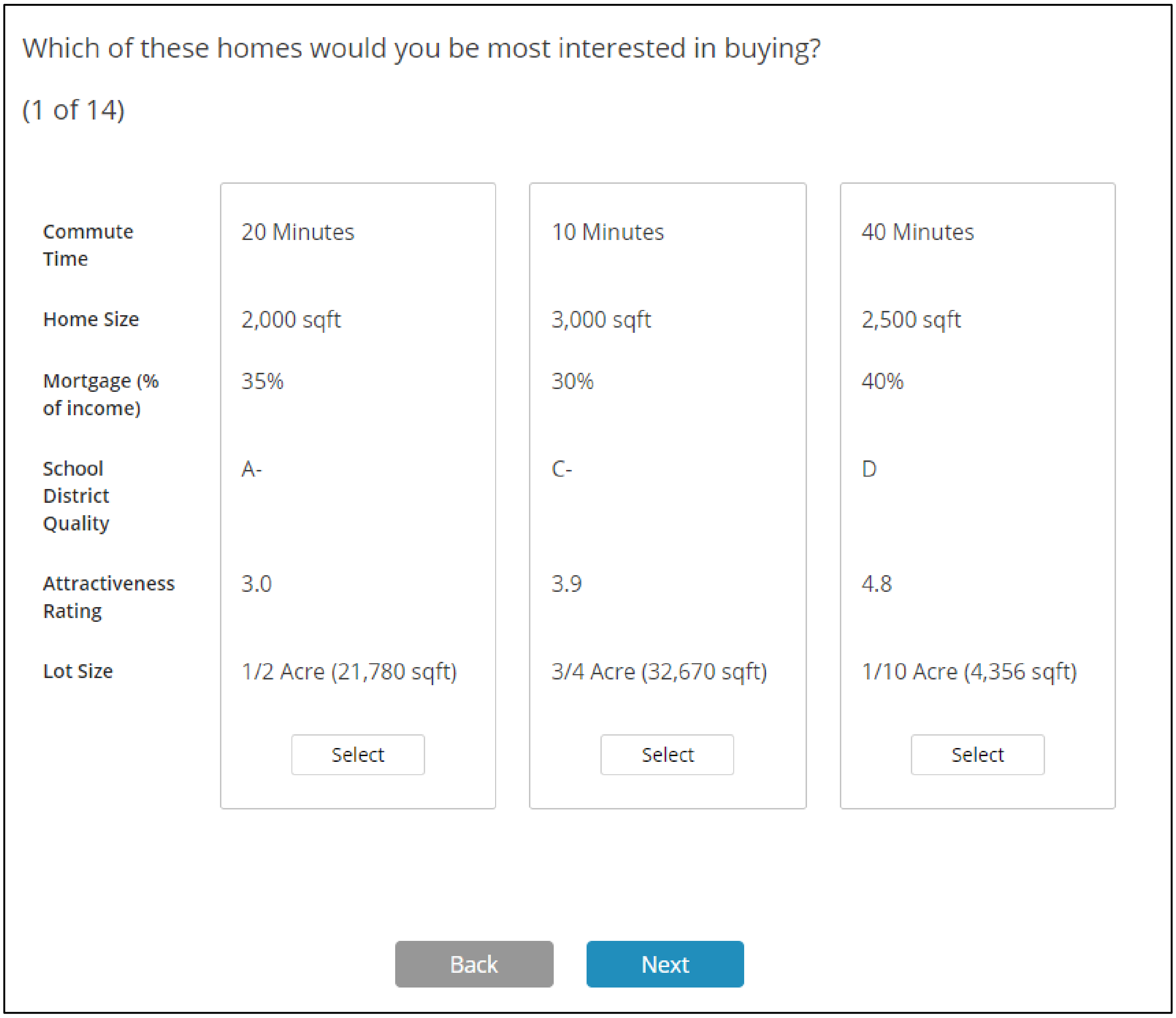
[1]Calculated via Fisher's r-to-z test; [2]Calculated via paired t-test; [3]Calculated via 2-sample test for equality of proportions; [4]Calculated via correlation test; [5]Averaged across attributes.

Table 2: Participant Song Ratings and Song Choices

	First Day of Summer	***Prisoner of Love***	***Seasons***
Chosen During CBC Task (% of participants)	35.5%	34.5%	30.0%
Average Enjoyment Rating (*sd*)	57.85 (29.76)	56.85 (26.92)	47.81 (28.82)
Given Highest Enjoyment Rating (% of participants)[1]	42.5%	34.8%	22.7%
Chosen as Most Enjoyable on Multiple Choice Item (% of participants)	41.8%	34.1%	24.1%

Note. [1]*Participants who reported two songs as being tied for most enjoyable were excluded from this calculation.*

Figure 1: Sample CBC Task



Which of these homes would you be most interested in buying?

(1 of 14)

Commute Time	20 Minutes	10 Minutes	40 Minutes
Home Size	2,000 sqft	3,000 sqft	2,500 sqft
Mortgage (% of income)	35%	30%	40%
School District Quality	A-	C-	D
Attractiveness Rating	3.0	3.9	4.8
Lot Size	1/2 Acre (21,780 sqft)	3/4 Acre (32,670 sqft)	1/10 Acre (4,356 sqft)

Figure 2 (Panel A): Average Metacognitive Knowledge Correlations Across Domains

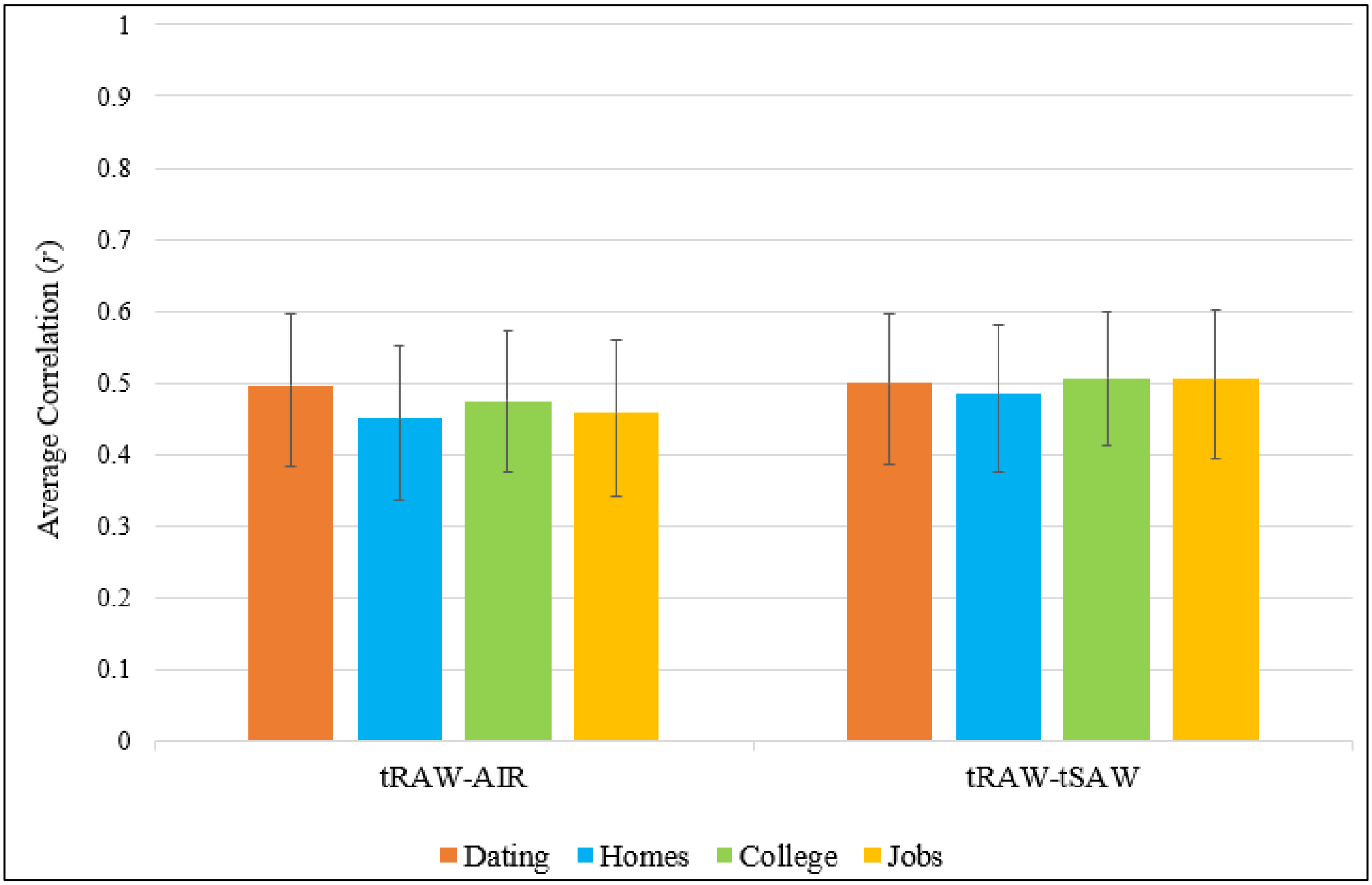


Figure 2 (Panel B): Average Euclidean Distances and RAW/SAW Different Choice Prediction Proportions Across Domains

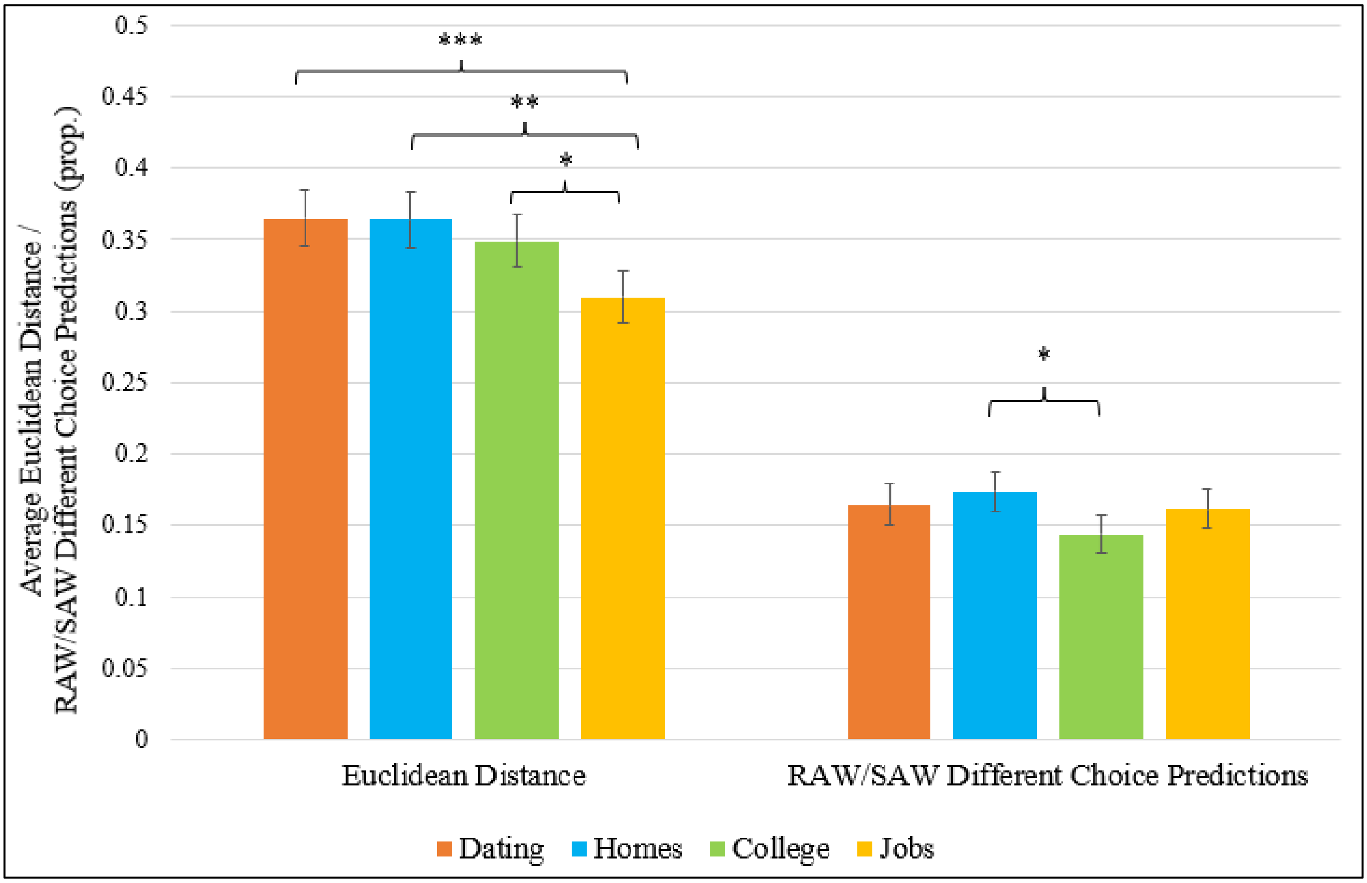
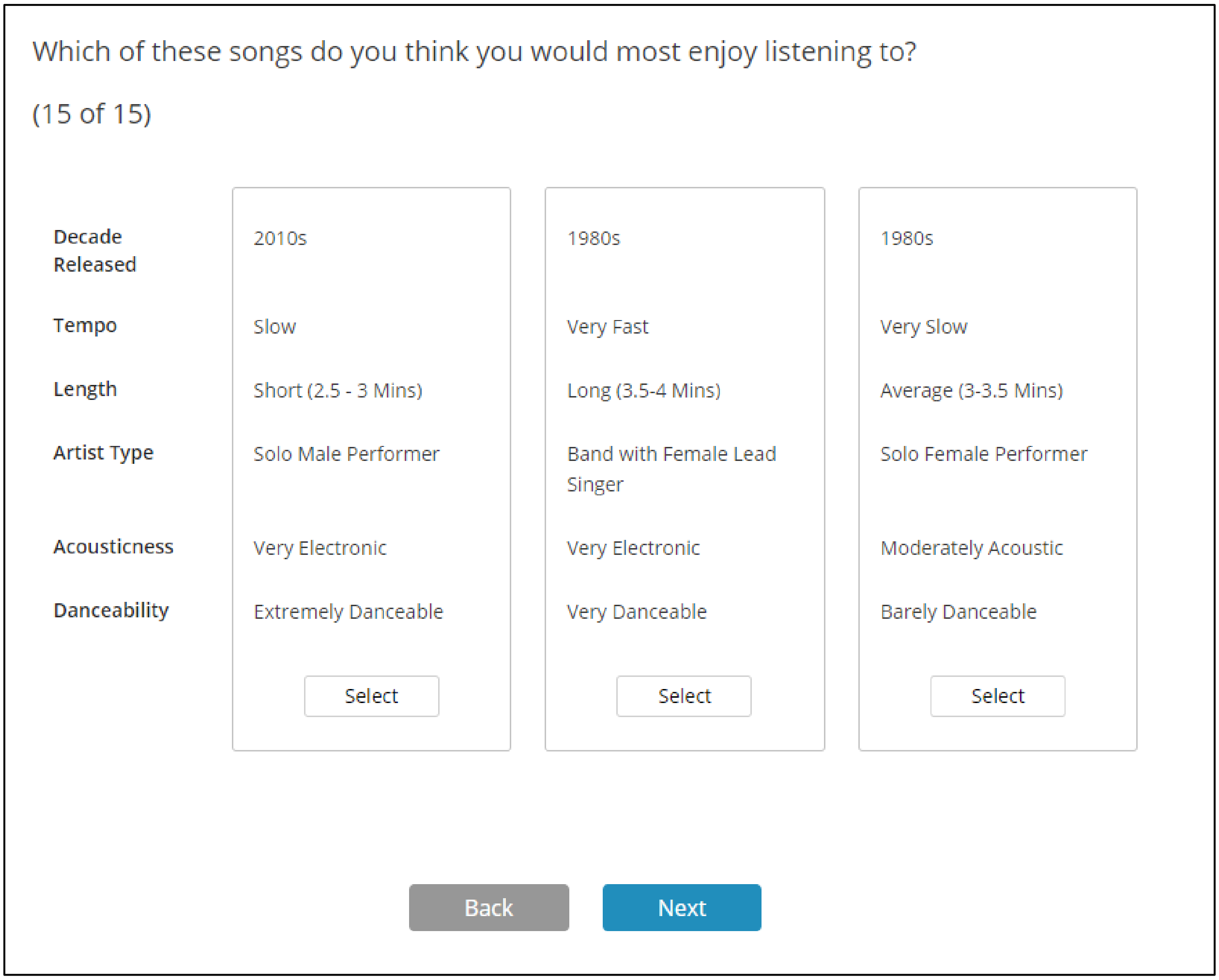


Figure 3: Fixed Song Choice Task



Which of these songs do you think you would most enjoy listening to?

(15 of 15)

Decade Released	2010s	1980s	1980s
Tempo	Slow	Very Fast	Very Slow
Length	Short (2.5 - 3 Mins)	Long (3.5-4 Mins)	Average (3-3.5 Mins)
Artist Type	Solo Male Performer	Band with Female Lead Singer	Solo Female Performer
Acousticness	Very Electronic	Very Electronic	Moderately Acoustic
Danceability	Extremely Danceable	Very Danceable	Barely Danceable

Figure 4: Error Magnitude by Euclidean Distance

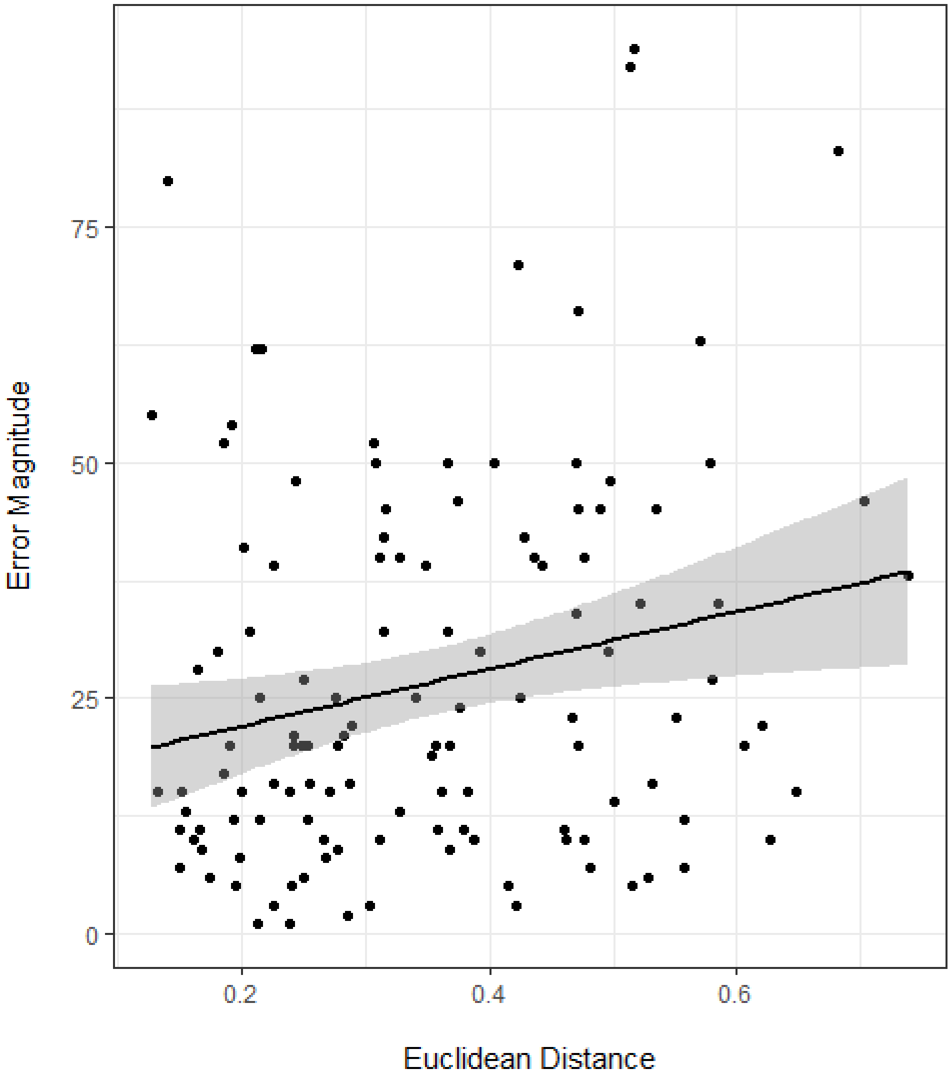


Figure Captions (List Format)

Figure 1: Sample CBC Task

No Caption

Figure 2 (Panel A): Average Metacognitive Knowledge Correlations Across Domains

*Note. Errors bars reflect 95% confidence intervals. *p < .05; **p < .01; ***p < .001*

Figure 2 (Panel B): Average Euclidean Distances and RAW/SAW Different Choice Prediction Proportions Across Domains

*Note. Errors bars reflect 95% confidence intervals. *p < .05; **p < .01; ***p < .001*

Figure 3: Fixed Song Choice Task

The song on the left is First Day of Summer *by Jesse Ruben (2018); the song in the middle is* Prisoner of Love *by Miami Sound Machine (1984); the song on the right is* Seasons *by Grace Slick (1980).*

Figure 4: Error Magnitude by Euclidean Distance

In correspondence with Model 1, this scatterplot does not include participants who had an error magnitude of zero.