Journal of Experimental Psychology: General


ISSN: 0096-3445

2026, Vol. 155, No. 2, 464–478
https://doi.org/10.1037/xge0001857

# Disentangling the Multifaceted Nature of Certainty in Evaluations

Julian Quandt[1, 2], Bernd Figner[2], Rob W. Holland[2], Maria Teresa Carere[2], Marijn Eversdijk[2, 3], and Harm Veling[2, 4]

[1] Institute for Cognition and Behavior, Vienna University of Economics and Business
[2] Behavioural Science Institute, Radboud University
[3] Department of Medical and Clinical Psychology, Tilburg University
[4] Consumption and Healthy Lifestyles, Wageningen University and Research

People hold their opinions and evaluations of objects with varying degrees of certainty. Understanding the nature of evaluation certainty is important, as it is a strong predictor of behavior. Several factors that impact certainty have been identified in past research, including value positivity, value extremity, and consistency in value-relevant evidence. However, whether these factors contribute uniquely to certainty or are statistically confounded remains unknown. The present work discusses four experiments (total $N = 372$) to disentangle these factors by leveraging the phenomenon of ensemble perception. We created pairs of ensembles (food baskets and retail good bundles) that are experimentally equalized on value positivity, value extremity, or evidence consistency, while systematically varying in evaluation certainty or evidence consistency. The results show independent contributions of value positivity and value extremity, but not evidence consistency, on ensemble evaluation certainty. Neither of the equalized factors fully accounts for evaluation certainty in the created ensembles, suggesting that it is a complex construct influenced by multiple independent factors.

***Public Significance Statement***

In today's world, characterized by an overwhelming variety of choice alternatives, understanding what drives certainty in the evaluations of these alternatives is crucial. High certainty in evaluations leads to more stable attitudes and behaviors, which are less likely to change over time. This research directly addresses the complex nature of evaluation certainty, a poorly understood yet critical aspect of decision making. By identifying and isolating three specific factors that could contribute to certainty—value positivity, value extremity, and evidence consistency—this study not only advances our theoretical understanding but also has practical implications. For instance, better insights into evaluation certainty can help in designing more effective interventions aimed at promoting sustainable and healthy behavior. Thus, by providing crucial insights into the underpinnings of evaluation certainty, this research paves the way for strategies that could influence people's evaluations and attitudes in more predictable and enduring ways.

*Keywords:* evaluation certainty, ensemble perception, value-based choice, confidence

*Supplemental materials:* https://doi.org/10.1037/xge0001857.supp



Many decisions that people make on a daily basis are instances of so-called *value-based decisions* that lack objectively correct answers, but instead depend on the subjective value that a decision-maker ascribes to the available alternatives, for instance, when selecting a tasty meal or an entertaining movie (Kahneman & Tversky, 1979; Martino & Cortese, 2023). According to current theory, subjective value is determined through an evidence accumulation process. In this process, a person samples value-relevant

This article was published Online First November 3, 2025.
Timothy Vickery served as action editor.
Julian Quandt https://orcid.org/0000-0002-3095-2710
All preregistrations, materials, and data are openly available on the Open Science Framework at https://osf.io/9y38n/. The work has previously been presented at lab meetings and seminars at Radboud University. The ideas or data have not been shared in any form outside of the authors' institutions and have not been uploaded outside of the Open Science Framework. The authors do not have any conflicts of interest to disclose. The authors thank all participants who took part in the studies for their time and effort.

Julian Quandt played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing–original draft, and writing–review and editing. Bernd Figner played a supporting role in conceptualization, formal analysis, investigation, methodology, supervision, and writing–original draft. Rob W. Holland played a supporting role in conceptualization, investigation, methodology, supervision, writing–original draft, and writing–review and editing. Maria Teresa Carere played a supporting role in conceptualization, investigation, and methodology. Marijn Eversdijk played a supporting role in conceptualization, investigation, and methodology. Harm Veling played a lead role in supervision and a supporting role in conceptualization, investigation, methodology, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Julian Quandt, Institute for Cognition and Behavior, Vienna University of Economics and Business, Welthandelsplatz 1, D5, 1020 Vienna, Austria. Email: julian_quandt@live.de

information from past experiences with an object. This information is then used to compare available alternatives (see Bakkour et al., 2019; Shadlen & Shohamy, 2016; Weber & Johnson, 2006; but see Hayden & Niv, 2021; Stewart et al., 2006, for alternative models that do not involve value comparisons). This evidence-sampling-from-memory account does not imply that evaluation is *entirely* subjective. For instance, when observing an apple, objective perceptual qualities of the apple, such as color, smell, and shape, can provide important information about the quality of the apple. However, these perceptual qualities can only provide information to an extent that the evaluator can relate these perceptual qualities to previous experiences (except potential innate associations, such as avoiding molded or rotten food). For example, without any prior experience on what perceptual qualities relate to the appetitiveness of an apple, a person would not be able to utilize this information, while a merely perceptual judgment of the color or size of the apple would not require such prior experiences. Indeed, instead of value- versus perceptual-based, evaluations of stimuli can be classified as representation-based (i.e., whether a mental representation of the stimulus is involved, e.g., evaluating the tastiness of food) versus stimulus-based (i.e., whether the evaluation can be made simply on the stimulus level, e.g., size or color, but also judging the pleasantness of a painting at the art gallery; Smith & Krajbich, 2021).

Importantly, even when initial perceptual information is available to guide value-based (or representation-based) evaluations, the available value-relevant evidence is imperfect. A priori, the taste of any specific apple can neither be perfectly predicted solely based on color and smell nor solely based on past experiences. Hence, people will generally feel some degree of uncertainty about their evaluations of an object (Koriat, 2024).

Certainty in evaluations is a key predictor of the stability and consistency of choices (Folke et al., 2016), attitudes, and opinions (DeMarree et al., 2020; Petrocelli et al., 2007; Tormala & Rucker, 2018). Certainty in evaluations can be influenced by a variety of factors (see Brus et al., 2021; Kiani et al., 2014; Koriat, 2012; Kvam & Pleskac, 2016), but particularly strong predictors of evaluation certainty are the positivity of object values (Lebreton et al., 2015; Lee & Hare, 2023) and value extremity (Polanía et al., 2019), with more positive and extreme values relating to higher certainty.

Yet, it is unclear whether these relations are causal or correlational (Folke et al., 2016), and despite the well-documented impacts of certainty on subjective values and resulting decisions (Tormala & Rucker, 2018), pinpointing the exact contributors to evaluation certainty is difficult. One key issue is that experimenters often lack access to the experiences underpinning evaluations and certainty. These experiences, however, could be important to understand how certainty arises, and it has been suggested that the consistency of such experiences mainly determines certainty (Quandt et al., 2022).

Interestingly, the apparent relation between value extremity and certainty could also be caused by precisely this consistency of experiences. Specifically, in most experiments of preferences, people report their evaluations on any bounded scale (e.g., Bakkour et al., 2019; Krajbich et al., 2010; Lee & Coricelli, 2020). If we posit that these evaluations represent the average value of the sampled evidence, and assuming that the sampled values fall inside the same scale, for values to be extreme, they also must be very *consistently* extreme. For objects that are not of extreme value, however, the potential variance of evidence samples is less constrained, and there need not be a relation between the degree of extremity and certainty. Thus, if certainty were mainly determined by the consistency of values, this would automatically result in a correlation between observed evaluation extremity and evaluation certainty.

Figure 1 (left) presents a simulation to illustrate this point. For the exact details of the simulation, see the Supplemental Material. Assuming that evidence is sampled on a bounded evaluation scale[1] and that people report the average value of the sampled experience as their evaluation, the variance that samples of value-relevant evidence can take is limited by the position of the evaluation on the scale. The possible variance of evidence samples is very small close to the endpoints of the scale and increases strongly toward the midrange. This results in a strong association between evaluation extremity and evaluation certainty, which is, however, driven by the necessity of having consistent (i.e., low variance) evidence samples at the extreme ends of the evaluation scale. For most of the scale range, where this necessity is less pronounced, there is only a weak association between evaluation extremity and evaluation certainty. Yet, in this range, where evidence samples are allowed to be consistent or inconsistent up to the possible variance on this point of the scale, evaluation certainty is still strongly determined by the consistency of evidence samples. As a result, the relation between evaluation extremity and evaluation certainty is not a direct one, but rather a byproduct of the necessity of consistent evidence samples for extreme evaluations. However, as we can usually not observe the consistency of evidence samples, it is difficult to estimate the exact contribution of evaluation extremity on evaluation certainty.

To provide further intuition, imagine a person evaluating the appetitiveness of a food item as 95 on a 100-point scale. Now, if we assume that the evaluation represents the average of some underlying sample of value-relevant evidence on the same 100-point scale, the possible variance of the sampled evidence must be very small to result in an average of 95. This means that the sampled evidence must be very consistent, which would lead to a high certainty in the evaluation. However, if the person evaluates a food item with a value of 50, the possible variance of the sampled evidence is much larger, and the sampled evidence can be much less consistent, which would lead to lower certainty in the evaluation. The simulation demonstrates this point by showing how observing a given evaluation restricts the possible variance of the sampled evidence.

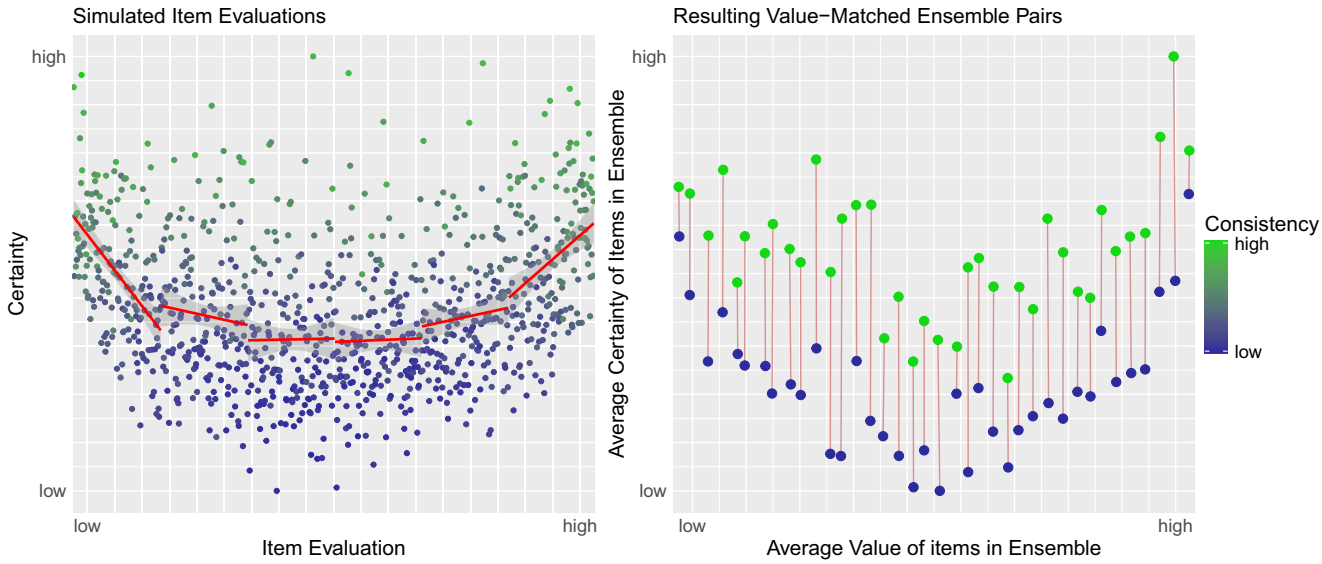
To some extent, this might also explain observed correlations between value positivity and value certainty in the literature (Lebreton et al., 2015; Lee & Hare, 2023). Specifically, when people's evaluation of most items in a study is positive, extremity and positivity would be positively correlated. Hence, any correlation between extremity and consistency could spill over to value positivity in a set of mostly positively evaluated items. This pattern complicates any conclusions about independent contributors to value certainty from experiments in which certainty, positivity, and extremity have not been explicitly unconfounded.

The main objective of this article is to systematically eliminate potential confounds across predictors of evaluation certainty, such as evaluation extremity and evaluation positivity, to infer whether these factors are unique contributors to evaluation certainty. To this

[1] It should be noted that when not assuming that the sampled evidence is restrained to the scale of the reported evaluations, the argument about limited variance of evidence samples at the scale endpoints would not necessarily apply, and there could be different reasons for a relation between evaluation extremity and certainty (Polanía et al., 2019).

**Figure 1**

*Simulated Item Evaluations With Certainty and Resulting Value-Matched Ensemble Pairs*




*Note.* Left panel: Simulation of object evaluations (*x*-axis) and evaluation certainty (*y*-axis) by a theoretical evaluator. The color of dots indicates the consistency of the sampled evidence used in each object evaluation. The consistency of sampled evidence is constrained by evaluation extremity (see the Supplemental Material for details on simulation). Certainty is directly determined by consistency plus an error term representing unobserved influences on certainty. The red regression lines indicate the relationship between evaluation and evaluation certainty, divided into six bins across the *x*-axis. There is a strong relation between evaluations and certainty toward the endpoints of the scale (i.e., for high evaluation extremity), but no significant relationship in the midrange. The color progression across the *y*-axis shows how certainty is determined by consistency throughout the entire scale range. Right panel: Pairs of 20 ensembles are created from the simulated objects by clustering objects with similar evaluations and grouping them within clusters to maximize the difference in evaluation certainty within pairs. This results in value-matched pairs of ensembles that can be used to investigate the impact of consistency on evaluation certainty between the matched pairs (connected by light-red lines), while experimentally eliminating the impact of the matching variable (i.e., item evaluations). See the online article for the color version of this figure.



end, we leverage the phenomenon of ensemble perception. Ensemble perception describes the phenomenon that the visual system tends to encode summary representations (i.e., statistical moments) of grouped stimuli, especially when an overarching representation can be formed from these objects (Whitney & Yamanashi Leib, 2018). Intuitive examples of ensemble perception are perceiving a combination of many trees as a forest or a group of people as a crowd. Experimentally, ensemble perception has been demonstrated on low-level visual percepts such as people encoding the average motion, brightness, or orientation in a set of objects (Bauer, 2009; Watamaniuk & Duchon, 1992) and on higher level percepts such as encoding the average emotional state of a crowd of faces (Haberman et al., 2009).

Ensemble perception extends to evaluations of objects, too. For instance, ensemble perception has been demonstrated when people judge the attractiveness of a crowd of faces (Walker & Vul, 2014) or the average price in a set of consumer goods, even if participants did not explicitly remember the individual goods (Yamanashi Leib et al., 2020). Importantly, it has been shown that these ensemble evaluations constitute an integration of the average value of *all* items in the ensemble rather than relying on exemplar items or merely a subset of the items (Yamanashi Leib et al., 2020). Here, we leverage this phenomenon as a methodological tool to combine multiple objects into ensembles to investigate how the certainty of evaluations of these ensembles would differ for ensembles that are matched on some average characteristic of the objects in the ensemble, such as value positivity or value extremity, to investigate whether equalizing these factors would result in equal evaluation certainty.

Crucially, for individual objects, equal value positivity and extremity are merely observed quantities. By combining items into ensembles, however, we can systematically construct pairs of ensembles that are equal on, for example, value positivity or value extremity to experimentally investigate the impact of these factors on evaluation certainty. For instance, as Figure 1 (right) demonstrates, it is possible to create pairs of ensembles that are of equal value positivity (in terms of the average value positivity of the items contained in the ensemble), while the items in the ensemble have different evaluation certainty. This allows for testing whether, if people evaluate the ensemble, the differences in certainty between pairs (i.e., the dots connected by light-red lines in Figure 1) would remain or whether, because of equalizing value positivity, there would be no remaining differences in evaluation certainty of the ensemble. Hence, if there were no more differences in evaluation certainty between the ensemble pairs, this would strongly suggest that value positivity would be the main contributor to evaluation certainty.

In four experiments, we systematically construct ensembles of everyday objects (foods and consumer goods; henceforth referred to as the *components* of an ensemble) to investigate the factors contributing to evaluation certainty by systematically *varying* potential contributors to evaluation certainty, while experimentally *equalizing* other potential contributors. Specifically, we focus on disentangling

the effects of consistency within and between subjective values of components in an ensemble, value positivity and extremity, and direct effects of component evaluation certainty on ensemble evaluation certainty. Table 1 provides an overview of which factors were equalized and varied in the different experiments, and the *Ensemble Creation* part of the Methods section provides details about the algorithm used for creating the ensembles. To foreshadow the results, we find unique impacts of value positivity, value extremity, and the consistency between, but not within, components on ensemble evaluation certainty. Moreover, there are persistent effects of component evaluation certainty on ensemble evaluation certainty when equalizing both value positivity and value extremity. These findings indicate a complex and partially independent interplay of factors in shaping certainty in evaluations.

## Experiment 1

First, before applying the ensemble-matching approach outlined in the introduction and displayed in Figure 1, we aimed to provide an empirical justification for investigating evaluation certainty in ensembles. Specifically, we wanted to investigate whether certainty in component evaluations (henceforth component certainty) would predict certainty in ensemble evaluations (henceforth ensemble certainty), analogous to previous findings on ensemble evaluations (Whitney & Yamanashi Leib, 2018; Yamanashi Leib et al., 2020). Second, we examined whether ensemble certainty would be driven by the consistency of between-component evaluations in the ensemble (henceforth between-component consistency). The rationale behind this is that during evaluation, ensembles with lower between-component consistency might be more difficult to evaluate, leading to lower certainty in the evaluation. As we were not mainly interested in these difficulty effects, we wanted to test whether, beyond these effects, there would be a persistent effect of component certainty on ensemble certainty, as we would expect from integration of component certainty during ensemble evaluation.

To investigate this, we created ensembles of items that varied in component certainty and between-component consistency.

### Method

#### *Transparency and Openness*

We report all data exclusions, manipulations, and measures in the study that were specified in the respective experiments' preregistrations. To comply with Transparency and Openness Promotion guidelines, we share all data, materials, and analysis scripts on the Open Science Framework (OSF) at https://osf.io/9y38n/ (Quandt et al., 2025). The preregistrations for each specific study are available under the aforementioned link. The study protocol was approved by the Ethics Committee of the Faculty of Social Sciences of Radboud University, and all participants provided informed consent. The implemented experimental protocols are shared as jsPsych (de Leeuw, 2015) program code, including dependencies, in the OSF repository. The analysis plan was followed as specified in the preregistrations. All exceptions to this are explicitly pointed out in the article. The materials used in the experiments are also available in the program code folders on the OSF repository. No preexisting data were used in this study. All collected data were anonymized and shared in the OSF repository, both as anonymized unprocessed data and processed data, including codebooks and processing steps as R Markdown files. The article is written in R Markdown using the papaja package (Aust & Barth, 2024) and can be exactly reproduced by downloading the R Markdown file and the associated model fitting scripts from the OSF repository. The R environment, including all package versions, is shared in the repository as an renv (Ushey & Wickham, 2025) lock file that can be used to reproduce the analysis environment. Further details about used software and analysis packages are provided in the Data Analysis section.

#### *Participants*

We ran an a priori sensitivity analysis using the approach described in Westfall et al. (2014), taking 71 participants as the desired sample size based on budget constraints (preregistered), allowing us to detect effects of $d = .47$ and larger. We collected a sample of Dutch participants on Prolific (https://www.prolific.com). Due to technical problems, the data of one participant were not correctly saved. Nine additional participants were excluded according to preregistered exclusion criteria, resulting in a final sample size of 61 participants (21 female, 39 male, one nonspecified [free-response box]; $M_{age} = 27.20$, $SD_{age} = 7.91$).

#### *Materials and Procedure*

The stimulus material consisted of 60 pictures of fruits and vegetables that we considered to be well-known throughout various countries. An overview of all items was presented prior to the evaluation task.

**Component Evaluation Task.** In the first task, 60 food items were presented to participants in a random order, asking them to rate how much they would "like to receive this food." Even though the experiment was conducted online, and participants would not receive the vegetables or fruits for actual consumption, they were told to provide an answer based on their feeling of wanting to receive

**Table 1**
*Equalized and Varied Factors Across the Four Experiments*

| Experiment | Component equalized factor | Component varied factor | Matching method |
|---|---|---|---|
| Experiment 1 | | Evaluation certainty<br>Between-component consistency | Two-step sorting |
| Experiment 2 | Value positivity | Evaluation certainty | *K*-means clustering |
| Experiment 3 | Value positivity<br>Value extremity | Evaluation certainty<br>Evaluation certainty | *K*-means clustering |
| Experiment 4 | Value extremity<br>Within-component evaluation consistency | Within-component evaluation consistency<br>Value extremity | *K*-means clustering |

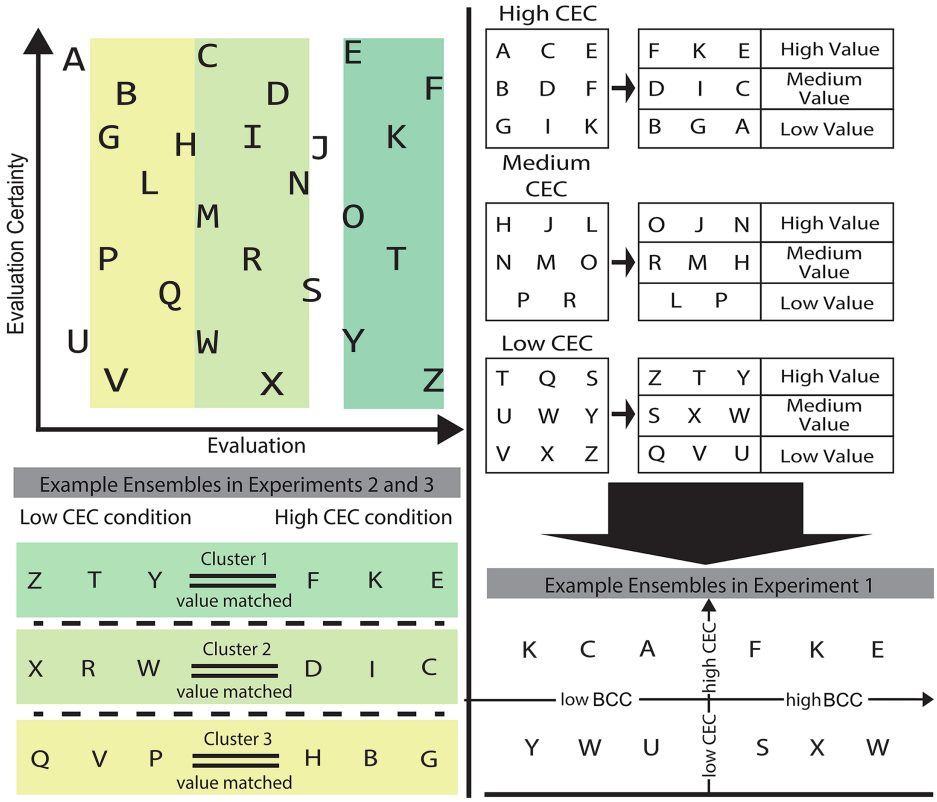
the food. After each evaluation, participants indicated their certainty in their evaluation. Both answers were assessed on a 0 (*not at all*) to 200 (*very much*) scale. Participants could skip evaluations for up to 10 items that they did not know. The number 10 here was chosen arbitrarily based on our assumption that people would know almost all the presented food items. In line with this, only a single person skipped 10 items, with the median participant skipping two items.

**Food Ensemble Creation.** Food ensembles were created using a two-step sorting algorithm (see Figure 2) and presented in the form of food baskets to participants. In Step 1, for each participant, the items were ranked based on component certainty. After ranking, the items were divided into a low-certainty, medium-certainty, and high-certainty category. In the second step, the items within each of the three categories were ranked based on evaluations from the component item evaluation task and again divided into three categories: a low-value category, a medium-value category, and a high-value category. Using these categories, five ensembles of four different types were created, each including three items:

1. *Low component certainty/low between-component consistency ensembles* including three items from the low-certainty item category from Step 1: one low-certainty/low-value item, one low-certainty/medium-value item, and one low-certainty/high-value item, resulting in a high standard deviation across evaluations, and hence low consistency.
2. *Low component certainty/high between-component consistency ensembles* including three items from the low-certainty category, but now including only low-, medium-, or high-value items in a respective ensemble, resulting in a low standard deviation across evaluations, and hence high consistency.
3. *High component certainty/low between-component consistency ensembles*, which are identical to the ensembles under 1, but including only the high-certainty items from the sorting in Step 1.

**Figure 2**

*Schematic of the Ensemble Creation Procedure Across Experiments*




*Note*. Top left: Objects (food items in Experiments 1 and 2, retail items in Experiments 3 and 4) here are depicted as letters, plotted by potential evaluations (*x*-axis) and evaluation certainty (*y*-axis). Right side, top to bottom: Experiment 1 involves ranking items by evaluation certainty and then creating ensembles based on certainty, followed by between-component consistency, with example ensembles shown in the lower right corner. Lower left: Example ensembles from Experiments 2–4 use a clustering algorithm to form ensembles by clustering items within narrow value ranges (shaded areas), selecting value-matched high- and low-certainty items to create pairs. CEC = component evaluation certainty; BCC = between-component consistency. See the online article for the color version of this figure.

4. *High component certainty/high between-component consistency ensembles* identical to 2, but only selecting high-certainty items.

**Ensemble Evaluation Task.** In the second task, the 20 created ensembles were presented to participants in random order, with three repeated ratings, with the presentation position of items rotated in each rating to account for presentation order effects. Participants were asked how much they would like to receive for each ensemble and how certain they were in this assessment.

### Data Analysis

The data were analyzed using Bayesian linear mixed-effects models in the probabilistic programming language Stan (Carpenter et al., 2017; Gabry & Češnovar, 2021), using the brms package (Bürkner, 2017) in R (Version 4.4.2, R Core Team, 2021). As certainty rating data are often skewed (see Quandt et al., 2022), we fitted a β-binomial response distribution model (see the Supplemental Materials). To investigate the hypotheses that between-component consistency and component certainty would predict ensemble certainty, we predicted certainty in ensemble evaluations on a 200-point scale by the created component certainty and between-component consistency conditions. Note that this is a deviation from the preregistration, where we planned to use component certainty and between-component consistency (as the standard deviation of evaluations of items in the ensemble) as numerical predictors instead of factorizing them. This deviation and others that are discussed in the Supplemental Materials were made for readability and conceptual clarity. The preregistered models were also run and did not differ in terms of conclusions. For formatting results and the entire article, the papaja R package (Aust & Barth, 2024) was used. As we report results of Bayesian models, instead of $p$ values, we report posterior proportions ($pp$). For effects predicted to be positive, we report $pp_-$, the posterior proportion that is negative (i.e., opposite to the prediction), while reporting $pp_+$, the positive proportion of the posterior, for effects predicted to be negative. This makes reading these values somewhat similar to $p$ values, though their interpretation is not entirely equivalent (see glossary in the Supplemental Materials). All reported experiments were approved by the Ethics Committee of the Faculty of Social Sciences of Radboud University, and all participants provided informed consent. The preregistrations of all experiments, their materials, and anonymized data, can be found on the OSF under https://osf.io/9y38n/.

## Results

### Main Analyses

We predicted (preregistered) that high (vs. low) component certainty would positively predict ensemble certainty and that high (vs. low) between-component consistency would also positively predict ensemble certainty. In line with this (see also Figure 3A), participants were more certain in their evaluations of ensembles with high-certainty components than low-certainty component ensembles (estimate = 0.15, 95% CI [0.10, 0.21], $pp_- < .001$). Similarly, high between-component consistency (estimate = 0.05, 95% CI [0.03, 0.07], $pp_- < .001$) resulted in higher ensemble certainty compared to low between-component consistency. A credible interaction between the predictors (estimate = 0.05, 95% CI [0.03, 0.07], $pp_- < .001$) indicated that the differences in ensemble certainty were stronger on the component certainty dimension than on the between-component consistency dimension (see Figure 3A). This result was corroborated by $k$-fold cross-validation (reported in the Supplemental Materials), showing that component certainty provided better predictions of ensemble certainty than between-component consistency.

### Exploratory Analyses: Including Value Positivity and Value Extremity

Next, we explored how the results would change if we included value positivity and value extremity as predictors of ensemble certainty to get a first glance at whether these factors would contribute to ensemble certainty independently of component certainty. When including value positivity and value extremity as predictors, the results significantly changed. There was no difference in ensemble certainty anymore between high versus low component certainty ensembles (estimate = 0.03, 95% CI [−0.18, 0.24], $pp_- = .408$) and no difference in ensemble certainty between high versus low between-component consistency ensembles (estimate = −0.02, 95% CI [−0.08, 0.03], $pp_- = .799$). Instead, there were credible differences in ensemble evaluation certainty for value positivity (estimate = 0.25, 95% CI [0.02, 0.47], $pp_- = .017$) and value extremity (estimate = 0.54, 95% CI [0.38, 0.70], $pp_- < .001$).

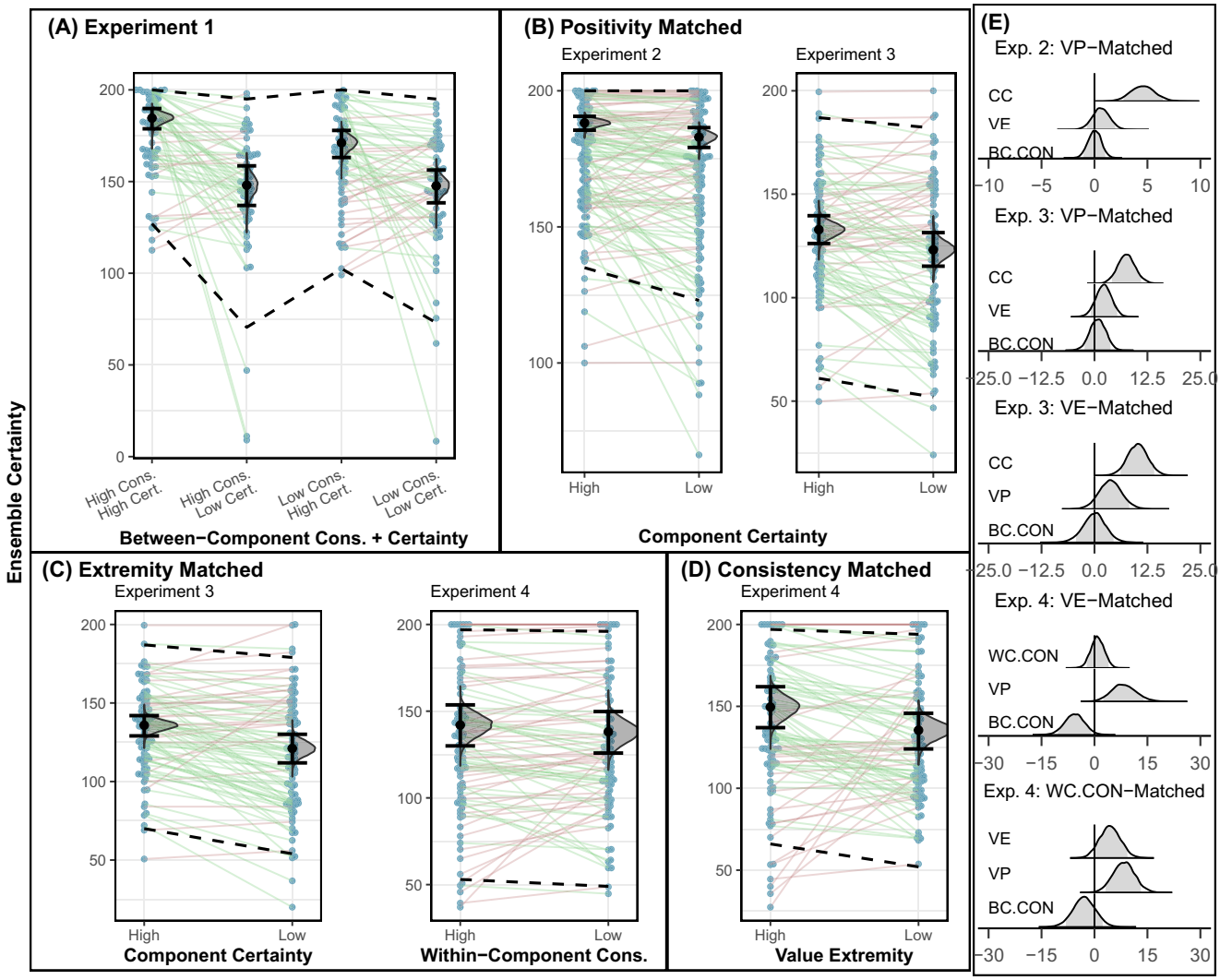
## Discussion

Experiment 1 demonstrates that ensemble perception effects (Whitney & Yamanashi Leib, 2018; Yamanashi Leib et al., 2020) generalize to the domain of evaluation certainty, complementing previous literature showing that certainty in specific attributes of objects independently contributes to overall certainty (Lee & Hare, 2023). Moreover, these effects extend beyond the integration difficulty of component values, as the effect of component certainty on ensemble certainty was stronger than the effect of between-component consistency. Together, these results suggest that ensemble certainty is integrated from component certainties.

However, the question remains what drives the effect of component certainty on ensemble certainty. Without experimentally equalizing value positivity and extremity, we cannot be sure that component certainty directly determines ensemble certainty or whether there is a confounder that determines this relation, such as value positivity or value extremity.In Experiment 1, when including value positivity and value extremity as predictors, there was no remaining effect of component certainty and between-component consistency, which could suggest that these effects are confounded. While we did not find statistical evidence for multicollinearity to support a confounded relation (all variance inflation factors < 2), making arguments about causal relations based on statistical changes when including additional predictors in a model is difficult and often problematic (Westfall & Yarkoni, 2016). This is exactly why it is important to experimentally disentangle the individual contribution of potential factors that could drive the effect of component certainty on ensemble certainty.

Hence, in Experiments 2 and 3, we equalized the value positivity and value extremity of the items in the ensembles but varied component certainty. This allows for directly probing the persistence of component certainty effects on ensemble certainty, even after experimentally removing any potential effect of value positivity and

**Figure 3**
*Main Results of Experiments 1–4*




*Note*. Main results of Experiments 1 (A) and 2–4 (B–D), presented by ensemble-matching factors. (B) Ensembles matched on value positivity. (C) Ensembles matched on value extremity. (D) Ensembles matched on within-component consistency. Blue dots represent ensemble certainty across ensembles pooled per participant. Black dots, error bars, and gray vertical densities represent model estimates of the condition mean, 95% credible interval, and posterior distribution, respectively. Solid-colored lines connect data points of individual participants and show whether the data are in line with predicted directions for a given participant (green: directionality in line with prediction; red: directionality not in line with prediction). Dashed lines indicate the 95% Bayesian prediction intervals (i.e., the interval in which 95% of out-of-sample observations are predicted to be observed). (E) Model estimates for predictors of models fitting the difference in ensemble certainty within matched pairs with densities representing posterior distributions with gray shaded areas showing the 95% credible intervals. CC = component certainty; VE = value extremity; VP = value positivity; BC.CON = between-component consistency; WC.CON = within-component consistency. See the online article for the color version of this figure.

value extremity. Additionally, we hypothesized that one important contributor to evaluation certainty could be the consistency of value-relevant evidence, which would be sampled during the evaluation process. Such a relation has, for example, been demonstrated by Lee and Hare (2023), who found that evaluation certainty was higher for objects that were rated on multiple dimensions when evaluations across the dimensions were more consistent. Hence, in Experiment 4, we investigated whether, similar to component certainty, the consistency of repeated evaluations *within* each component would directly predict ensemble certainty, even when equalizing value extremity.

## Experiment 2

In Experiment 2, we aimed to investigate whether there would be a difference in ensemble certainty between ensembles with high and low component certainty when experimentally equalizing value positivity.

## Method

### Participants

Sample size was determined through power simulation (see the Supplemental Materials for details). Excluding participants based on preregistered criteria and replacing the excluded participants, we arrived at a sample size of 131 participants (48 female, 82 male, one nonbinary [free-response box]; $M_{age}$ = 26.22, $SD_{age}$ = 7.95) to achieve 80% power. We recruited a sample of Dutch participants on Prolific (https://www.prolific.com).

### Materials

The stimulus material for Experiment 2 was identical to Experiment 1.

### Procedure

**Component Evaluation Task.** The procedure for the component evaluation task was identical to Experiment 1.

**Ensemble Creation.** We created food product ensembles, each of which contained three items. A clustering algorithm produced ensembles with matching value positivity, while exhibiting maximal within-cluster differences in component certainty. The algorithm ensured that the value positivity of the items within each ensemble was matched, while the component certainty was maximally different between the two ensembles. A detailed description of the algorithm is provided in the Supplemental Materials, and a schematic depiction is presented in Figure 2. Table 1 provides an overview of which aspects were matched and varied in each experiment. In short, the algorithm created several clusters of items based on component evaluations to ensure minimal differences in value positivity within each cluster. Then, three high-certainty and three low-certainty items from each cluster were selected to create a positivity-matched pair of two ensembles, one with high component certainty and one with low component certainty.

**Ensemble Evaluation Task.** The procedure of evaluating the ensembles was identical to Experiment 1.

### Data Analysis

For the main data analysis, we used a β-binomial Bayesian mixed-effects model with ensemble certainty as the dependent variable and component certainty (factorial: low vs. high), value positivity (standardized), between-component consistency (standardized), and component value extremity (standardized) as the predictors (see model specification in the Supplemental Material). Moreover, to allow for more nuanced conclusions, a region of practical equivalence (ROPE; Kruschke & Liddell, 2018) was preregistered and defined as [−0.0125, 0.0125] (see the Supplemental Materials for details). Specifically, as our goal was to identify *major* determinants of ensemble certainty, we wanted to ensure that any such claim would only be made when the effect size was large enough to be considered meaningful in this context. While the definition of "meaningful" remains arbitrary, the size of the ROPE relates to the size of effects of evaluation variability on evaluation certainty found in previous experiments, fitting similar models, when directly manipulating the consistency of value-relevant evidence on evaluation certainty for artificial fractal items (Quandt et al., 2022).

If the 95% credible interval of an effect lies entirely inside the ROPE and the 95% credible interval includes zero, we would consider the effect precisely estimated and conclude that it did not, in an important way, impact ensemble certainty. If the 95% credible interval (partly) lies within the ROPE but does not include zero, we consider the effect credible but too small to be a major contributor to ensemble certainty. If the 95% credible interval falls outside of the ROPE, we consider the effect to be a major contributor to ensemble certainty, comparable with experimental findings where only a single manipulated factor was influencing evaluation certainty (Quandt et al., 2022).

## Results

### Manipulation Checks: Resulting Matched Ensembles

First, we wanted to check whether the matched ensembles would have the desired properties, that is, being close to identical on average component value positivity but differing in average component certainty. For this, we fitted a Gaussian family Bayesian mixed-effects model predicting the average component value positivity from the ensemble condition (low vs. high component certainty). We found that the matching algorithm succeeded in creating ensemble pairs in the two conditions that were indeed highly similar on average value positivity (average component value positivity in high component certainty condition = 113.14 vs. average component value positivity in low component certainty condition = 112.26), with no credible difference between the two conditions (estimate = 0.41, 95% CI [−1.79, 2.54], $pp_- = .350$; see also the high overlap of posterior distributions in Supplemental Figure S2 that strongly suggests no difference).

Moreover, the matching algorithm successfully created ensemble pairs that differed in average component certainty (average component certainty in high component certainty condition = 184.28 vs. average component certainty in low component certainty condition = 141.97), with a credible difference between the two conditions (estimate = 21.14, 95% CI [19.03, 23.27], $pp_- < .001$; see the Supplemental Material for visualization).

### Main Analysis: Matching Value Positivity

Next, we tested our prediction (preregistered) that higher component certainty would result in higher ensemble certainty after experimentally equalizing value positivity of ensembles. We found a credible difference in ensemble certainty between the high and low component certainty ensembles (estimate = 0.03, 95% CI [0.02, 0.04], $pp_- < .001$; high component certainty: 188.20, low component certainty: 182.97), but this difference was not credibly larger than the defined ROPE[2] (estimate = 0.00, 95% CI [−0.01, 0.01], $pp_- = .411$; see Figure 3B). Moreover, we found credible estimates for between-component consistency (estimate = −0.02, 95% CI [−0.03, −0.01], $pp_+ < .001$; estimate negative as operationalized as *SD* between component evaluations, i.e., negative consistency), component value positivity (estimate = 0.03, 95% CI [0.01, 0.05], $pp_- = .005$), and component value extremity (estimate = 0.12, 95% CI [0.09, 0.14], $pp_- < .001$). Importantly, the estimates of between-component consistency, value extremity, and value positivity

[2] The test against the ROPE indicates how much of the posterior falls inside of the ROPE.

represent the effect of these factors *across* ensemble pairs and hence cannot be interpreted in the same way as the experimentally varied factor of component certainty. In other words, while these estimates are credible, they can only be interpreted as a correlation between the respective factor and ensemble certainty across the range of ensemble values.

### *Exploratory Analyses*

While the effects across ensemble pairs are not surprising and in line with previous literature (Lee & Hare, 2023; Polanía et al., 2019), the question remains whether these factors similarly explain the differences in ensemble certainty within the value positivity matched pairs of ensembles. Hence, we conducted additional exploratory analyses to investigate the impact of between-component consistency and component value extremity within matched pairs of ensembles by calculating difference scores for each predictor within pairs of matched ensembles. That is, for each predictor, we calculated the difference score between the high and low component certainty ensembles in each pair (e.g., the difference in value extremity between a given high and low component certainty ensemble in a matched pair) and predicted the difference in ensemble certainty from these difference scores. This has the advantage that these predictors directly allow us to infer their impact on ensemble certainty within matched pairs instead of representing the correlation between the respective factor and ensemble certainty across the range of ensemble values.

We fitted a Student-T family Bayesian mixed-effects model (the Student-T model is considered an outlier-robust alternative to the standard Gaussian family model; Bürkner, 2017) predicting the difference in ensemble certainty from the difference in component certainty (i.e., using a standardized numerical predictor instead of the factorial indicator), ensemble value extremity (standardized), and between-component consistency (standardized). We found that the difference in component certainty was the only factor that predicted ensemble certainty (estimate = 4.58, 95% CI [2.40, 6.86], $pp_- < .001$), with no effect of value extremity (estimate = 0.62, 95% CI [−1.00, 2.21], $pp_- = .221$) and between-component consistency (estimate = 0.01, 95% CI [−1.22, 1.23], $pp_+ = .513$). Hence, the impact of component certainty on ensemble certainty was attributable to neither value extremity nor between-component consistency (see Figure 3E—Exp. 2: VP-Matched).

## Discussion

Experiment 2 demonstrated that value certainty in components significantly influenced ensemble certainty, independent of value positivity, which was equalized within the matched ensemble pairs. Exploratory analyses revealed that this component certainty effect is independent of value extremity and between-component consistency, both of which did not predict ensemble certainty in the value positivity matched ensembles. Especially, the absence of an effect of value extremity on ensemble certainty is surprising, as previous research has shown that value extremity is a major contributor to ensemble certainty (Lee & Hare, 2023; Polanía et al., 2019). One reason for this might be that the current set of ensembles of food items contains primarily positive-value ensembles (median value = 129.17), with primarily very high-certainty ratings (median certainty = 180). While differences in certainty within matched ensemble pairs (see the Manipulation Checks section) were still pronounced, this could still have deflated potential effects.

Hence, in Experiment 3, we aimed to decrease the average ensemble evaluation and ensemble certainty by using a different set of stimuli (retail goods) and a different evaluation task (the Becker–DeGroot–Marschak method; Becker et al., 1964). Specifically, we expected that people would be less opinionated about prices they would be willing to pay for a diverse set of retail goods, compared to their liking of common foods, and that this would lead to a larger variance in ensemble evaluation (i.e., willingness to pay) and ensemble certainty. Moreover, given the noteworthy absence of an effect of value extremity on ensemble certainty in Experiment 2, we also aimed to investigate this more closely by directly matching ensembles on value extremity.

# Experiment 3

Experiment 3 is partially a replication of Experiment 2, but with a different set of stimuli (retail goods instead of food items) and a different method of evaluation (the Becker–DeGroot–Marschak auction instead of evaluations). Moreover, we matched half the ensembles on component value positivity (as in Experiment 2) and the other half on component value extremity.

## Method

### *Participants*

The sample size was determined using power simulation based on the results of Experiment 2 (see the Supplemental Material for details). This resulted in a suggested sample size of 90 participants (38 female, 48 male, four nonbinary [free-response box]; $M_{age}$ = 29.81, $SD_{age}$ = 8.52) to arrive at 80% power, which we recruited from a Dutch sample on Prolific (https://www.prolific.com).

### *Materials and Procedure*

Stimuli consisted of 60 retail products that were taken from a popular online retailer in the Netherlands. Examples of items are a bracelet, a cocktail set, a drinking bottle, or a photo album. All products were selected to cost up to €25.

**Component Evaluation Task.** The structure of the component evaluation task was mostly identical to Experiment 2. However, instead of indicating their liking for the retail goods, participants were asked to place bids on each good according to the rules of the Becker–DeGroot–Marschak auction method (Becker et al., 1964). These bids were hypothetical and nonincentivized.

**Ensemble Creation.** The ensemble creation procedure was identical to Experiment 2, with the exception that only half the ensembles were matched on value positivity, while the other half was matched on value extremity, meaning that products (components) with extremely low and extremely high value could be included in the same ensemble.

**Ensemble Evaluation Task.** In the ensemble evaluation task, participants placed bids on the ensembles, similar to the component evaluation task, with the potential bids ranging from €0 to €75, to reflect that in purchasing, the sum of the prices determines the ensemble price, rather than the average.

### *Data Analysis*

The data analysis procedure was identical to Experiment 2. That is, we predicted ensemble certainty from the difference in component certainty, component value positivity, component value extremity, and between-component consistency (for details, see model specification in the Supplemental Material).

## Results

### *Manipulation Checks: Resulting Matched Ensembles*

Again, we fitted a Gaussian family Bayesian mixed-effects model predicting the component value positivity from the ensemble condition (low vs. high component certainty) to check whether the matched positivity-matched ensembles would have the desired properties, that is, being close to identical on component value positivity but differing in average component certainty. We found that the matching algorithm succeeded in creating ensembles in the two conditions that were indeed highly similar on average value positivity (sum of value positivity [price] in high component certainty condition = 25.84 [out of 75] vs. sum of component value positivity in low component certainty condition = 25.82), with no credible difference between the two conditions (estimate = 0.01, 95% CI [−1.02, 1.07], $pp_- = .494$). As intended, these matched ensembles differed in average component certainty (average component certainty in high component certainty condition = 157.10 vs. average component certainty in low component certainty condition = 94.84), with a credible difference between the two conditions (estimate = 31.13, 95% CI [28.75, 33.63], $pp_- < .001$; see the Supplemental Material for visualization).

Similarly, for the extremity-matched ensembles, we found that matched ensembles were highly similar in average component value extremity (sum of component value extremity in high component certainty condition = 18.10 vs. sum of component value positivity in low component certainty condition = 18.09), with no credible difference between the two conditions (estimate = 0.00, 95% CI [−0.69, 0.72], $pp_- = .496$). As intended, these matched ensembles differed in average component certainty (average component certainty in high component certainty condition = 159.11 vs. average component certainty in low component certainty condition = 93.38), with a credible difference between the two conditions (estimate = 32.89, 95% CI [30.25, 35.62], $pp_- < .001$; see the Supplemental Material for visualization).

Together, this shows that the matching algorithm worked as intended for both matching methods and that the overall value positivity and value extremity decreased compared to Experiment 2.

### *Main Analyses*

**Positivity-Matched Ensembles.** Similar to Experiment 2, we predicted (preregistered) that higher component certainty would result in higher ensemble certainty after experimentally equalizing value positivity of ensembles. As in Experiment 2, we found a credible difference in ensemble certainty between the high and low component certainty ensembles (estimate = 0.05, 95% CI [0.02, 0.08], $pp_- < .001$; high component certainty: 132.88, low component certainty: 123.13). Again, this difference was not credibly larger than the defined ROPE (estimate = 0.02, 95% CI [−0.00, 0.05], $pp_- = .040$; see Figure 3B). Replicating findings in Experiment 2, component value positivity (estimate = 0.10, 95% CI [0.05, 0.16], $pp_- < .001$) and component value extremity (estimate = 0.11, 95% CI [0.06, 0.16], $pp_- < .001$) predicted ensemble evaluation certainty across matched ensemble pairs. Unlike in Experiment 2, there was no credible effect of between-component consistency on ensemble evaluation certainty (estimate = −0.02, 95% CI [−0.05, 0.01], $pp_+ = .079$).

**Extremity-Matched Ensembles.** For the ensembles matched on component value extremity, we similarly found a credible difference in ensemble certainty (preregistered) between the high and low component certainty ensembles (estimate = 0.07, 95% CI [0.04, 0.11], $pp_- < .001$; high component certainty: 135.83, low component certainty: 121.08), this time larger than the defined ROPE (estimate = 0.05, 95% CI [0.02, 0.08], $pp_- = .001$; see Figure 3C). This suggests that even when equalizing component value extremity, component certainty still strongly impacted ensemble certainty. Additionally, across extremity-matched pairs, we found that component value positivity (estimate = 0.07, 95% CI [0.02, 0.12], $pp_- = .004$), component value extremity (estimate = 0.08, 95% CI [0.04, 0.12], $pp_- < .001$), and between-component consistency (estimate = −0.05, 95% CI [−0.08, −0.01], $pp_+ = .012$) predicted ensemble evaluation certainty across matched ensemble pairs.

### *Exploratory Analyses*

**Comparing Predictors of Ensemble Certainty Within Matched Pairs.** Again, we explored whether the difference in ensemble certainty *within* matched pairs could be predicted by the difference in component certainty, value extremity, and between-component consistency. Similar to Experiment 2, for the positivity-matched ensembles, we found that the difference in component certainty was the only factor that predicted ensemble certainty (estimate = 7.56, 95% CI [3.50, 11.82], $pp_- < .001$), with no effect of value extremity (estimate = 2.19, 95% CI [−1.53, 5.83], $pp_- = .123$) and between-component consistency (estimate = 0.79, 95% CI [−2.93, 4.39], $pp_+ = .660$; see Figure 3E—Exp. 3: VP-Matched).

For the extremity-matched ensembles, we found similar results, with component certainty being the only credible predictor of ensemble certainty within matched pairs (estimate = 10.20, 95% CI [5.39, 15.02], $pp_- < .001$), while value positivity (estimate = 3.77, 95% CI [−1.49, 9.20], $pp_- = .075$) and between-component consistency (estimate = −0.02, 95% CI [−5.47, 5.20], $pp_+ = .504$) did not credibly predict ensemble certainty (see Figure 3E—Exp. 3: VE-Matched).

**Rating Variability and Ensemble Certainty.** According to our theoretical outline in the introduction and based on previous research (Quandt et al., 2022), we expected component certainty to tap into the consistency of value-relevant evidence that is sampled during evaluations of those items. One possible implication of this could be that the more consistent the evidence sampled during the component evaluations, the more similar the evidence sampled during the ensemble evaluation. Hence, we predicted that the difference between the average component evaluations and the respective ensemble evaluations would be smaller for high component certainty ensembles compared to low component certainty ensembles.

To account for zero differences between component evaluations and ensemble evaluations, we fitted a hurdle-lognormal model that estimates the chance of a difference being exactly zero per condition

separately from the chance of observing difference scores larger than zero in each condition. For severely right-skewed data, this avoids the problem of having to add a constant to difference scores of zero for log transformation. We performed this analysis on the combined data of Experiments 2 and 3, which was preregistered after data collection and after already being in close contact with the data, but before exploring or conducting this particular analysis or computing the required scores.

In the combined data, we found that there were credibly more average component evaluation versus ensemble evaluation deviations of exactly zero in the high component certainty ensembles (estimate of the percentage of deviations that are exactly zero = 7.06, 95% CI [4.63, 9.89]) compared to the low component certainty ensembles (estimate = 1.82, 95% CI [0.90, 3.07]). For deviations of component evaluations and ensemble evaluations that were larger than zero, we found no credible difference between the high and low component certainty ensembles in terms of average component evaluation versus average ensemble evaluation (estimate = 0.00, 95% CI [−0.05, 0.05], $pp_+ = .529$). Initially, this might suggest that high component certainty makes it more likely for ensemble evaluations to be exactly like the underlying average component evaluation. However, a credible interaction (estimate = 0.03, 95% CI [0.02, 0.06], $pp_- < .001$) between the chance of observing a zero deviation between conditions and the experiment identifier term in the model suggested that this difference might be more pronounced in Experiment 2 compared to Experiment 3.

To follow up on this, we explored the data separately for Experiment 2 and Experiment 3. For Experiment 2, we found that there were credibly more deviations of exactly zero between component and ensemble evaluations in the high component certainty ensembles (estimate = 5.95, 95% CI [3.37, 9.15]) compared to the low component certainty ensembles (estimate = 0.89, 95% CI [0.27, 2.04]). For deviations that were larger than zero, we found no credible difference between the high and low component certainty ensembles in terms of average component evaluation versus average ensemble evaluation (estimate = −0.00, 95% CI [−0.06, 0.06], $pp_+ = .484$).

For Experiment 3 (using both the positivity-matched and extremity-matched ensembles in the analysis), we found no difference in the chance of having a deviation of exactly zero between the high and low component certainty ensembles (estimate of the percentage of deviations that are exactly zero = 5.61, 95% CI [3.60, 7.97], vs. estimate = 4.97, 95% CI [3.21, 7.14]). Similarly, we found no effect on the average nonzero deviation between component evaluations and ensemble evaluations (estimate = −0.00, 95% CI [−0.02, 0.02], $pp_+ = .458$).

This suggests that the effect of observing more identical average component evaluations versus ensemble evaluations in the high component certainty ensembles compared to the low component certainty ensembles is driven by Experiment 2 and that this effect is not present in Experiment 3. A possible reason for this might be that the cases where the component evaluation versus ensemble evaluation deviation was exactly zero in Experiment 2 were almost exclusively those at the boundary of the scale (i.e., average component evaluations of exactly zero or exactly 200; about 98% of these cases), while this was only about 58% in Experiment 3. As these boundary cases were overrepresented in the high component certainty ensembles (141 cases in high component certainty ensembles vs. 75 cases in low component certainty ensembles), the observed difference in the chance of a zero difference between conditions might just be a statistical artifact and not in fact indicate that higher component certainty is indicative of higher evidence sampling consistency.

## Discussion

Experiment 3 demonstrated that value certainty in components significantly influenced ensemble certainty, independent of value positivity (replicating findings of Experiment 2) and value extremity. However, this effect was substantial (i.e., larger than the defined ROPE) only for the ensembles matched on value extremity. The stronger effect in extremity-matched ensembles suggests a closer relationship between value positivity and value certainty, aligning with prior research indicating a strong link between these factors (Lebreton et al., 2015; Lee & Hare, 2023).

The question remains as to what accounts for the persistent effect of value certainty. We assumed component certainty to tap into the consistency of value-relevant evidence that is sampled during evaluations of those items (Quandt et al., 2022). Exploratory analyses did not support this idea, as we found no credible difference in the average deviation between component evaluations and ensemble evaluations between high and low component certainty ensembles. This was surprising in light of previous work (Quandt et al., 2022). Therefore, in Experiment 4, we examined the idea more directly. Specifically, we matched ensembles on the consistency of repeated evaluations of the same components. Based on work suggesting that more consistent evaluations of components will be indicative of a more consistent underlying evidence sampling process (Quandt et al., 2022), we expected higher ensemble certainty in ensembles with more consistent component evaluations.

# Experiment 4

Experiment 4 used the same stimuli and procedure as Experiment 3, but with a different matching design. Specifically, to identify whether the effect of component certainty on ensemble certainty is driven by the consistency of the evidence *within* components, we created ensembles matched on within-component consistency but differing in average component value extremity (to isolate the effect of value extremity when equalizing within-component consistency), and vice versa (i.e., isolating the effect of within-component consistency when equalizing value extremity).

## Method

### Participants

Running a power simulation based on the data from Experiment 3 (see the Supplemental Material for details), we again collected 90 participants (36 female, 53 male, one genderfluid [free-response box]; $M_{age} = 30.61$, $SD_{age} = 9.19$) from a Dutch online sample on Prolific (https://www.prolific.com).

### Materials and Procedure

We used the same stimuli and evaluation procedure (the Becker–DeGroot–Marschak bidding task) as in Experiment 3.

**Component Evaluation Task.** The component evaluation task was identical to Experiment 3.

**Ensemble Creation.** To create the ensembles, we used the same algorithm as in Experiments 2 and 3, but with different matching criteria. Half the ensembles were created to be matched on component value extremity (as in Experiment 3), but instead of maximizing differences in average component certainty within matched pairs, we maximized differences within these ensemble pairs in terms of *within*-component consistency (operationalized as a smaller standard deviation of multiple evaluations for the same component). For the other half of ensembles, this matching criterion was reversed, creating ensemble pairs matched on within-component consistency but differing in average component value extremity.

**Ensemble Evaluation Task.** The ensemble evaluation task was identical to Experiment 3. However, while bids were hypothetical in Experiment 3, Experiment 4 included a raffle where one participant was randomly selected to receive one of the goods they had bid on.

### Data Analysis

The general approach to data analysis was identical to Experiment 3. However, we did not preregister a ROPE in Experiment 4. This is because we see any small effect of value extremity on ensemble certainty when within-component consistency is equalized as contradicting the idea that within-component consistency is the causal mechanism behind the effect of value extremity on ensemble certainty (hence conducting a more conservative test of our prediction). The β-binomial model in Experiment 4 predicted ensemble certainty from the matching condition (low vs. high within-component consistency or low vs. high component value extremity), the matching method (matched on value extremity vs. matched on within-component consistency), and the interaction of matching method by matching condition.

## Results

### Manipulation Checks: Resulting Matched Ensembles

The matching procedure employed in Experiment 4 again succeeded in creating the desired ensembles. For the ensembles matched on value extremity, there was no credible difference in component value extremity between the high and low within-component consistency conditions (component value extremity in high component consistency condition = 19.20 vs. component value extremity in low component consistency condition = 19.01), with no credible difference between the two conditions (estimate = 0.10, 95% CI [−0.49, 0.69], $pp_-$ = .359). Meanwhile, the average within-component consistency was credibly higher in the high within-component consistency condition than in the low within-component consistency condition (average standard deviation within repeated ratings in high component consistency condition = 1.06 vs. average standard deviation within repeated ratings in low component consistency condition = 4.27), with a credible difference between the two conditions (estimate = −1.60, 95% CI [−1.76, −1.44], $pp_+ < .001$).

For the ensembles matched on within-component consistency, there was no credible difference in average within-component consistency between the high and low value extremity conditions (average within-component consistency in high extremity condition = 1.98 vs. average within-component consistency in low extremity condition = 2.03), with no credible difference between the two conditions (estimate = −0.02, 95% CI [−0.14, 0.10], $pp_+ = .358$). Meanwhile, the component value extremity was credibly higher in the high value extremity condition than in the low value extremity condition (component value extremity in high extremity condition = 28.15 vs. component value extremity in low extremity condition = 10.79), with a credible difference between the two conditions (estimate = 8.69, 95% CI [8.20, 9.17], $pp_- < .001$).

### Main Analyses

We predicted (preregistered) that higher within-component consistency would give rise to higher ensemble certainty (Lee & Coricelli, 2020) even when value extremity is equalized. We also predicted that, assuming that ensemble certainty would mainly be driven by within-component consistency (Quandt et al., 2022), the difference in ensemble certainty would be smaller for ensembles matched on within-component consistency (preregistered) even if evaluation extremity would differ, which would provide evidence for the confounding idea outlined in Figure 1.

Surprisingly, we found that ensembles matched on value extremity but differing in within-component consistency did not credibly differ in ensemble certainty (estimate = −0.02, 95% CI [−0.05, 0.01], $pp_+$ = .080; high within-component consistency: 144.38, low within-component consistency: 139.85; see Figure 3C). Moreover, and in stark contrast to the prediction and the argument presented in Figure 1, ensembles matched on within-component consistency and differing on evaluation extremity exhibited a credible difference in ensemble certainty (estimate = 0.07, 95% CI [0.03, 0.11], $pp_-$ = .001; high component value extremity: 150.69, low component value extremity: 136.44; see Figure 3D). This effect of average component value extremity on ensemble certainty in the within-component consistency matched ensembles was statistically more pronounced than the effect of within-component consistency in the extremity-matched ensembles, as indicated by a credible interaction between the matching condition and matching method (estimate = −0.02, 95% CI [−0.05, −0.00], $pp_+$ = .024).

### Exploratory Analyses

As in Experiments 2 and 3, we explored whether the differences in ensemble certainty within matched pairs could be predicted by the differences in component value extremity, value positivity, and between- and within-component consistency. We found that differences in ensemble certainty within extremity-matched pairs were predicted by differences in value positivity (estimate = 8.34, 95% CI [1.93, 15.46], $pp_-$ = .005) and between-component consistency (estimate = −5.66, 95% CI [−11.24, −0.19], $pp_+$ = .021), but not within-component consistency (estimate = 0.86, 95% CI [−2.84, 4.56], $pp_-$ = .323; see Figure 3E—Exp. 4: VE-Matched). For the ensembles matched on within-component consistency, we found that only value positivity was a credible predictor of ensemble certainty (estimate = 8.32, 95% CI [2.38, 14.43], $pp_-$ = .002), with no effect of between-component consistency (estimate = −2.92, 95% CI [−9.25, 3.27], $pp_+$ = .180) or value extremity (estimate = 4.42, 95% CI [−1.21, 10.24], $pp_-$ = .066; see Figure 3E—Exp. 4: WC.CON-Matched).

## General Discussion

The presented set of experiments investigated the impact of different factors on evaluation certainty. Using a novel approach that utilized a matched ensemble evaluation design, we found that evaluation certainty is a multifaceted construct that is influenced by multiple, independent factors.

First, certainty is influenced by the variation in values between components in an ensemble, but this effect could not account for all variance in evaluation certainty. Similarly, when experimentally equalizing value positivity and value extremity, there was a persistent effect of component evaluation certainty on ensemble evaluation certainty that was not explained by between-component value consistency, value extremity, or value positivity.

These results substantiate earlier findings (Lebreton et al., 2015; Lee & Hare, 2023) by showing that value positivity is a prominent factor that consistently impacted evaluation certainty. Specifically, when equalizing value positivity, the relation between component evaluation certainty and ensemble evaluation certainty was weak, and including value positivity as a predictor of ensemble value certainty in Experiment 1 rendered the relation between component certainty and ensemble certainty noncredible. These results suggest that value certainty and value positivity are inherently related.

Most importantly, we predicted that the consistency of value-relevant evidence would be the most likely underlying mechanism behind evaluation certainty in general, but did not find support for this hypothesis. First, we argued that more consistent evidence should lead to higher similarity between component evaluations and the resulting ensemble evaluations in high component certainty ensembles compared to low component certainty ensembles.

We did not find support for this idea. Moreover, within-component value consistency, which we expected to tap into evidence consistency, did not predict ensemble evaluation certainty. In fact, in line with previous research (Bobadilla-Suarez et al., 2020; Polanía et al., 2019), value extremity appears to be a stronger predictor of certainty than the consistency of value-relevant evidence. This directly contradicts the idea outlined in the introduction (and illustrated in Figure 1) and is inconsistent with the idea that consistency is the underlying factor driving the relation between evaluation extremity and certainty. These results are surprising, given the idea that during the sampling of value-relevant evidence from memory (Bakkour et al., 2019; Shadlen & Shohamy, 2016), more consistent evidence is sampled for high-certainty items, and given that previous research directly showed that manipulating evidence consistency predicts evaluation certainty (Quandt et al., 2022).

One reason for this contradiction could be that our operationalization of evidence consistency as the standard deviation of repeated evaluations of components is not adequately capturing the consistency of value-relevant evidence. Specifically, it is possible that evidence sampled *within* a single evaluation is consistent, but that for a repeated separate evaluation of the same item, other, potentially also consistent, evidence is sampled, resulting in two *inconsistent* evaluations, which can still both be based on consistent evidence *within* each individual evaluation.

In this case, as a reviewer pointed out, the absence of a relation between variability in component evaluations and ensemble certainty might not be surprising. While previous research suggests that evaluation variability does relate to certainty (Lee & Coricelli, 2020) and other potential indicators of consistency of past value-relevant experiences, it was also found that the relation between evaluation variability and manipulated evidence consistency might not be very strong (Quandt et al., 2022). Hence, it remains possible that more sensitive measures of evidence consistency, such as a distribution builder task (Quandt et al., 2022; Sharpe et al., 2000), might provide support for a relation between evidence consistency and certainty. The present results, however, do not suggest that the consistency of value-relevant evidence is, contrary to our prediction, the most prominent predictor of certainty.

The small yet consistent influence of component evaluation certainty on ensemble value certainty hence raises questions about its underlying factors, especially since none of the measured and equalized variables in the present design account for this effect. Previous studies suggest that attentional mechanisms (Brus et al., 2021) or the reliability of evidence for making predictions (Boundy-Singer et al., 2022; Koriat, 2024) might play significant roles in determining certainty in value-based decisions. However, it is unclear to what extent these factors might be relevant in the present research, where we investigate evaluation certainty rather than decision certainty, which might be different constructs driven by different value representations (Brus et al., 2021; Peters & Büchel, 2010; Pouget et al., 2016).

Another factor that was not investigated in the present design and that might explain the persistent effect of component certainty on ensemble certainty is the amount of available evidence (Kvam & Pleskac, 2016). For example, it is possible that the amount of evidence available for each component is not equal, which could lead to differences in certainty. Specifically, this could underlie the relation between value positivity and certainty (but not value extremity and certainty), as positive items might be more frequently consumed, leading to a larger amount of evidence available for these items. Future research might investigate this idea to complement the present findings and provide an even more complete picture of factors impacting evaluation certainty.

Several shortcomings of the present research should be considered. First, measuring certainty in component evaluations only in Experiments 1–3 where evidence consistency was not directly measured, and then measuring evidence consistency in Experiment 4 without assessing certainty, prevents us from drawing stronger conclusions about the relationship between these two factors. While previous research has shown that consistency in evaluations is related to certainty (Lee & Coricelli, 2020; Quandt et al., 2022), investigating this in an ensemble-matching design where the two variables could be experimentally equalized could provide a more comprehensive understanding of the relationship between certainty and consistency. Second, despite evidence showing that ensembles are perceived as unified entities (Whitney & Yamanashi Leib, 2018; Yamanashi Leib et al., 2020), the influence of unique characteristics of components on ensemble evaluation certainty, beyond simple integration, remains a possibility. For example, our focus was solely on evaluations, whereas objects are often judged on multiple dimensions, such as tastiness and healthiness for food, which are known to contribute to overall certainty as a multidimensional construct (Lee & Hare, 2023). The ensemble-matching design presented here might provide a valuable tool to construct items that can be experimentally controlled on one dimension (e.g., healthiness) while varying other dimensions (e.g., tastiness) to increase our understanding of not only certainty but also the importance of those attributes in choice processes.

Relatedly, previous research showed that value certainty is a strong predictor of choice accuracy, choice consistency, and choice confidence (De Martino et al., 2013; Lee & Daunizeau, 2020, 2021). As we did not assess choices between objects in this work, future research could investigate how the different factors that impact evaluation certainty identified in this article translate to decisions between ensembles, especially as considering all relevant attributes of an object (e.g., healthiness and tastiness of foods) is more likely to play a role in decision making, when attributes are directly goal-relevant (Peters & Büchel, 2010).

Overall, the present work provides experimental evidence for evaluation certainty being *independently* impacted by value positivity and value extremity but provides no evidence for a relation between consistency of value-relevant evidence and certainty in an experimentally controlled setting. This represents an incremental but important advancement in understanding the complex factors affecting evaluation certainty and suggests an ensemble-matching design as a valuable tool that could inform various areas within psychological and decision sciences, where understanding certainty in evaluations or attitudes is crucial.

## Constraints on Generality

The present work was conducted on an online sample of participants from the Netherlands and Germany to make sure that they were familiar with the items used in the experiments. Hence, it is not clear how the results would generalize to other populations, such as participants from other countries or cultures, especially non-Western, educated, industrialized, rich, and democratic countries, where perception of value-based decisions might differ (Anlló et al., 2024). Another constraint on generality is the use of a specific set of items (food and retail goods) in the present work. While these items were chosen to be exemplary of common value-based decisions, it is possible that the results would not generalize to other domains, such as social evaluations or beliefs, that might also inherently be value-based (Kim et al., 2020). Finally, the present work focused on evaluations of items in isolation and did not investigate comparative evaluations that are often used in real-world decision making. It is important to consider these constraints on generality when interpreting the results and their implications.

## References

Anlló, H., Bavard, S., Benmarrakchi, F., Bonagura, D., Cerrotti, F., Cicue, M., Gueguen, M., Guzmán, E. J., Kadieva, D., Kobayashi, M., Lukumon, G., Sartorio, M., Yang, J., Zinchenko, O., Bahrami, B., Silva Concha, J., Hertz, U., Konova, A. B., Li, J., … Palminteri, S. (2024). Comparing experience- and description-based economic preferences across 11 countries. *Nature Human Behaviour*, *8*(8), 1554–1567. https://doi.org/10.1038/s41562-024-01894-9

Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R markdown* (Version 0.1.3) [Computer software]. https://cran.r-project.org/web/packages/papaja/index.html

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., Shadlen, M. N., & Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *eLife*, *8*, Article e46080. https://doi.org/10.7554/eLife.46080

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record*, *59*(2), 171–185. https://doi.org/10.1007/BF03395657

Becker, G. M., Degroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232. https://doi.org/10.1002/bs.3830090304

Bobadilla-Suarez, S., Guest, O., & Love, B. C. (2020). Subjective value and decision entropy are jointly encoded by aligned gradients across the human brain. *Communications Biology*, *3*(1), Article 597. https://doi.org/10.1038/s42003-020-01315-3

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2022). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142–154. https://doi.org/10.1038/s41562-022-01464-x

Brus, J., Aebersold, H., Grueschow, M., & Polania, R. (2021). Sources of confidence in value-based choice. *Nature Communications*, *12*(1), Article 7337. https://doi.org/10.1038/s41467-021-27618-5

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

DeMarree, K. G., Petty, R. E., Briñol, P., & Xia, J. (2020). Documenting individual differences in the propensity to hold attitudes with certainty. *Journal of Personality and Social Psychology*, *119*(6), 1239–1265. https://doi.org/10.1037/pspa0000241

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), Article 0002. https://doi.org/10.1038/s41562-016-0002

Gabry, J., & Češnovar, R. (2021). *Cmdstanr: R interface to 'CmdStan'* (Version 0.9.0) [Computer software]. https://mc-stan.org/cmdstanr/

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(11), Article 1. https://doi.org/10.1167/9.11.1

Hayden, B. Y., & Niv, Y. (2021). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behavioral Neuroscience*, *135*(2), 192–201. https://doi.org/10.1037/bne0000448

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences*, *24*(2), 101–111. https://doi.org/10.1016/j.tics.2019.12.001

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. https://doi.org/10.1037/a0025648

Koriat, A. (2024). Subjective confidence as a monitor of the replicability of the response. *Perspectives on Psychological Science*, *20*(4), 744–761. https://doi.org/10.1177/17456916231224387

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a

Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180. https://doi.org/10.1016/j.cognition.2016.04.008

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167. https://doi.org/10.1038/nn.4064

Lee, D. G., & Coricelli, G. (2020). An empirical test of the role of value certainty in decision making. *Frontiers in Psychology*, *11*, Article 574473. https://doi.org/10.3389/fpsyg.2020.574473

Lee, D. G., & Daunizeau, J. (2020). Choosing what we like vs liking what we choose: How choice-induced preference change might actually be instrumental to decision-making. *PLOS ONE*, *15*(5), Article e0231081. https://doi.org/10.1371/journal.pone.0231081

Lee, D. G., & Daunizeau, J. (2021). Trading mental effort for confidence in the metacognitive control of value-based decision-making. *eLife*, *10*, Article e63282. https://doi.org/10.7554/eLife.63282

Lee, D. G., & Hare, T. A. (2023). Value certainty and choice confidence are multidimensional constructs that guide decision-making. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(3), 503–521. https://doi.org/10.3758/s13415-022-01054-4

Martino, B. D., & Cortese, A. (2023). Goals, usefulness and abstraction in value-based choice. *Trends in Cognitive Sciences*, *27*(1), 65–80. https://doi.org/10.1016/j.tics.2022.11.001

Peters, J., & Büchel, C. (2010). Neural representations of subjective reward value. *Behavioural Brain Research*, *213*(2), 135–141. https://doi.org/10.1016/j.bbr.2010.04.031

Petrocelli, J. V., Tormala, Z. L., & Rucker, D. D. (2007). Unpacking attitude certainty: Attitude clarity and attitude correctness. *Journal of Personality and Social Psychology*, *92*(1), 30–41. https://doi.org/10.1037/0022-3514.92.1.30

Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, *22*(1), 134–142. https://doi.org/10.1038/s41593-018-0292-0

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Quandt, J., Figner, B., Holland, R. W., Carere, M. T., Eversdijk, M., & Veling, H. (2025). *Disentangling the multifaceted nature of certainty in evaluations*. https://osf.io/9y38n

Quandt, J., Figner, B., Holland, R. W., & Veling, H. (2022). Confidence in evaluations and value-based decisions reflects variation in experienced values. *Journal of Experimental Psychology: General*, *151*(4), 820–836. https://doi.org/10.1037/xge0001102

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 0.1.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939. https://doi.org/10.1016/j.neuron.2016.04.036

Sharpe, W. F., Goldstein, D. G., & Blythe, P. W. (2000). *The distribution builder: A tool for inferring investor preferences* [Working paper]. https://web.stanford.edu/~wfsharpe/art/qpaper/qpaper.pdf

Smith, S. M., & Krajbich, I. (2021). Mental representations distinguish value-based decisions from perceptual decisions. *Psychonomic Bulletin & Review*, *28*(4), 1413–1422. https://doi.org/10.3758/s13423-021-01911-2

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Tormala, Z. L., & Rucker, D. D. (2018). Attitude certainty: Antecedents, consequences, and new directions. *Consumer Psychology Review*, *1*(1), 72–89. https://doi.org/10.1002/arcp.1004

Ushey, K., & Wickham, H. (2025). *Renv: Project environments* (Version 1.1.4) [Computer Software]. https://CRAN.R-project.org/package=renv

Walker, D., & Vul, E. (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science*, *25*(1), 230–235. https://doi.org/10.1177/0956797613497969

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*(5), 931–941. https://doi.org/10.1016/0042-6989(92)90036-i

Weber, E. U., & Johnson, E. J. (2006). Constructing preferences from memory. In P. Slovic & S. Lichtenstein (Eds.), *The construction of preference* (pp. 397–410). Cambridge University Press. https://doi.org/10.1017/CBO9780511618031.022

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*(3), Article e0152719. https://doi.org/10.1371/journal.pone.0152719

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Yamanashi Leib, A., Chang, K., Xia, Y., Peng, A., & Whitney, D. (2020). Fleeting impressions of economic value via summary statistical representations. *Journal of Experimental Psychology: General*, *149*(10), 1811–1822. https://doi.org/10.1037/xge0000745

Received August 28, 2024
Revision received August 11, 2025
Accepted August 23, 2025 ■