

Journal of Experimental Psychology: General

Confidence in Evaluations and Value-Based Decisions Reflects Variation in Experienced Values

Julian Quandt, Bernd Figner, Rob W. Holland, and Harm Veling

Online First Publication, September 27, 2021. <http://dx.doi.org/10.1037/xge0001102>

CITATION

Quandt, J., Figner, B., Holland, R. W., & Veling, H. (2021, September 27). Confidence in Evaluations and Value-Based Decisions Reflects Variation in Experienced Values. *Journal of Experimental Psychology: General*. Advance online publication. <http://dx.doi.org/10.1037/xge0001102>

Confidence in Evaluations and Value-Based Decisions Reflects Variation in Experienced Values

Julian Quandt, Bernd Figner, Rob W. Holland, and Harm Veling
Behavioural Science Institute, Radboud University

Evaluations and value-based decisions are often accompanied by a feeling of confidence about whether or not the evaluation or decision is accurate. We argue that this feeling of confidence reflects the variation of an underlying value distribution and that this value distribution represents previously experienced values related to an object. Two preregistered experiments in which the variation of such value distributions was systematically varied provide causal evidence in favor of this hypothesis. A third preregistered experiment showed that, for natural food items with uncontrolled prior experiences, confidence in evaluations is again related to the variation of individuals' self-reported value distributions. Similarly, for choices between items, the variation of experienced values related to a choice pair influenced confidence in the perceived correctness of the choice. These findings converge with other domains of decision making showing that confidence tracks the variation of the underlying probability distribution of the evidence that a decision is based on, which in the case of value-based decisions, is informed by a value distribution reflecting priorly experienced values.

Keywords: confidence judgements, value variation, value-based decisions, evaluations


Supplemental materials: <https://doi.org/10.1037/xge0001102.supp>

Whether considering taking an umbrella to work, judging the credibility of a newspaper piece, or deciding what to order at a restaurant, everyday life is full of decisions. Some decisions are based on what we perceive in the outside world (e.g., whether there are dark clouds in the sky), some are based mainly on knowledge of facts (e.g., whether the news piece is inconsistent with what we know), and others are based mainly on our preferences (e.g., deciding what to eat). These three types of decisions, perceptual, factual, and value-based, are quite different from each other but all require the decision maker to evaluate evidence in order to select one of the available choice alternatives (Abelson, 1988; Dutilh & Rieskamp, 2016; Gold & Shadlen, 2007; Griffin & Tversky, 1992). For instance, the judgment of a newspaper piece's correctness might be based on other pieces about the same topic that the decision maker has read before. Depending on how clear the evidence is to us we may be more certain or confident that a decision is correct (Abelson, 1988; Boldt et al., 2017; Petrocelli et al.,

2007; Vickers & Packer, 1982). Thus, confidence is a judgment of the degree to which we think our decisions are correct and supported by evidence.

However, this description of confidence seems odd in the case of value-based decisions as, in contrast to perceptual and factual decisions, they do not have an objective criterion of correctness (Dutilh & Rieskamp, 2016). Nonetheless, people can readily report their confidence about their own evaluations of objects on any arbitrary scale (De et al., 2013; Folke et al., 2016; Lebreton et al., 2015; Tormala & Rucker, 2018) and confidence reports even reliably predict future choices (Folke et al., 2016; Petrocelli et al., 2007; Tormala & Rucker, 2018). This begs the question how we can collect and evaluate evidence about our own preferences, for instance how much we like apples, and what the nature of this evidence is.

Several theories converge on the idea that confidence reflects a summary of evidence quality (Kepecs & Mainen, 2012; Lebreton et al., 2015; Loeb & Fishel, 2014; Meyniel et al., 2015; Rolls et al., 2010; van den Berg et al., 2016). For instance, if a person judges the color of a traffic light in heavy rain, the evidence quality represents the degree of visibility of the traffic light's state. Specifically, the evidence quality can be described as a probability distribution where the possible states of the world (whether the traffic light is green, yellow, or red) are related to their probabilities indicated by the evidence (e.g., how much it looks like it is green, yellow or red). Confidence, in turn, reflects the belief that a decision or judgment is correct based on the variance of the probability distribution (Kepecs & Mainen, 2012; Pouget et al., 2016). The more the evidence is divided between the possible colors of the traffic light, the lower confidence in the judgment of the traffic light's color.

Julian Quandt  <https://orcid.org/0000-0002-3095-2710>

All online materials related to this article including data, analysis scripts and materials are available on the OSF at <https://osf.io/q72sm>.

This work has previously presented at Associatie van Sociaal Psychologische Onderzoekers conference in Wageningen in 2019 as part of a session about uncertainty in decision making. It has not been further disseminated in public.

Correspondence concerning this article should be addressed to Julian Quandt, Behavioural Science Institute, Radboud University, Radboud Montessorilaan 3, A 9.10, 6525 HR, Nijmegen, the Netherlands. Email: julian.quandt@ru.nl

What distinguishes value-based decisions from perceptual decisions is that the evidence to be used for value-based decisions cannot simply be observed in the environment (as opposed to the color of the traffic light) but is somehow represented internally. Even though one might argue that perceptual evidence is eventually evaluated based on an internal representation, it is clear what external evidence the internal representation is based on. However, what is the external evidence and the internal representation of evidence during value-based decisions? Research suggests that evidence during value-based decisions might be based on previous memories and experiences with the object of evaluation (Johnson et al., 2007; Weber et al., 2007). Specifically, evaluations during value-based decisions might be constructed through sequential sampling from memory (Bakkour et al., 2018, 2019; Shadlen & Shohamy, 2016; Vanunu et al., 2019). Sequential sampling models propose that, during the decision-making process, the evidence for different choice alternatives is accumulated until it supports one of the possible alternatives clearly enough to decide. Sequential sampling models have proven to be valuable models of decision making in perceptual decisions (Ratcliff et al., 2018; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998, 2000) as well as value-based decisions (Busemeyer et al., 2019; Busemeyer & Townsend, 1993; Krajbich, 2019; Krajbich et al., 2010; Krajbich & Rangel, 2011; van den Berg et al., 2016).

Even though we are not directly testing sequential sampling from memory here, it produces important predictions for the relation between the variation in experienced values and confidence. Specifically, in order to collect evidence to be used in the evaluation of objects (i.e., to construct the evaluation of objects), previous experience with these objects (within a certain context) is recalled and integrated into a sample of evidence. We propose that this sample of evidence forms a *value distribution* akin to the aforementioned probability distribution: a distribution that encodes probabilities of previously experienced object values. The variance of this value distribution describes the (inverse) quality of the underlying evidence with increasing variance resulting in lower confidence in an evaluation. Moreover, we propose that, during binary choice, two value distributions need to be compared. Specifically, when evaluating the value difference between two items (Lim, et al., 2011), the amount of conflicting information about the superiority of either option, should be reflected by postchoice confidence.

Interestingly, this idea is extending previous research on confidence in value-based decisions for binary choices that described confidence as a mere reflection of the value difference of choice alternatives (Folke et al., 2016; Lim et al., 2011) rather than how clear the evidence is for a given value difference. In other words, what matters for postchoice confidence should not mainly be the value difference, but rather the variance of possible value differences (Lebreton et al., 2015; Pouget et al., 2016). This is an important distinction, as two choice alternatives might have a small but very reliable value difference, or a rather large but very uncertain value difference.

Altogether, the goal of the current project was to investigate the nature of confidence in value-based decisions. Specifically, we examined three overarching questions.¹ Question 1 (Q1): How does a value distribution's variance relate to confidence in the evaluations of an object? That is, if the values that people have learned about an object have higher variance, will they then be less confident when evaluating the objects? Prediction 1 (P1): We predicted that

people would be less confident in the evaluations of an object when the object's value distribution has high variance.

Q2: As outlined above, we assume that people construct values through sequential sampling of experiences from memory. This raises the question whether repeated evaluation of the same object is more variable when an object's value distribution has high variance. P2a: We predicted that variance of value distributions would give rise to more diverse evaluations if the item is rated repeatedly. P2b: Moreover, we predicted that this higher evaluation variability would, again, be related to lower confidence.

Q3: Finally, we examined how the overlap of two value distributions influenced postchoice confidence. P3: We predicted that higher overlap between value distributions of two objects would result in lower postchoice confidence.²

An obvious challenge here is that value distributions cannot be controlled for most ecological stimuli as the underlying value distributions are learned throughout a person's lifetime. Therefore, we designed two experiments where participants first learned the value distributions of six novel fractal pictures (the *learning task*), which were manipulated in terms of average value and variance using a rapid serial visual presentation paradigm (Kunar et al., 2017; Tsetso et al., 2012). Next, an *evaluation and confidence-judgment task* assessed participants' evaluation and confidence judgements of the manipulated value distributions multiple times per fractal. Next, we administered a *distribution builder task* (Sharpe et al., 2000), in which participants were told to recreate previously learned value distributions by stacking points on a scale. Finally, we employed a binary-choice paradigm in which we systematically varied the overlap of value distributions assessing choices and the associated postchoice confidence. In Experiment 3, in which we used pictures of natural food items, we employed the latter three tasks to investigate whether we would find correlational support for the hypotheses using ecological objects.

As predicted, we find that value distribution width influences confidence in evaluations of novel stimuli and relates to confidence in evaluations of food items. Moreover, confidence in a choice between two items is predicted by overlap between the respective value distributions. However, exploratory analyses show that choices and postchoice confidence can better, and theoretically more parsimoniously, be predicted by a single distribution representing the difference between the two value distributions of the two choice alternatives.

Experiments 1 and 2

Method

All reported experiments were approved by the Ethics Committee of the Faculty of Social Sciences of Radboud University and

¹ These predictions are the predictions outlined in the preregistration of Experiment 2.

² Note that our original predictions included two predictions for Q3. The one reported here as P3 was originally prediction P3b, while we also predicted that the overlap of value distributions would increase the probability of choosing lower-value items (P3a). In response to reviewer comments recommending focusing the discussions mainly on confidence instead of choice outcomes, we excluded this prediction from the main text. However, note that this prediction was not confirmed and is now reported in the online supplemental material (Section 6).

all participants provided informed consent. As Experiment 2 was a direct replication of Experiment 1 with only small adaptations, we report both experimental methods together. The differences are mentioned in the descriptions below. The preregistrations of both experiments, the Python programs (Version 3.7) using pygame (Version 1.9.4; Shinnars, 2011) and PsychoPy (Version 3.1.4; Peirce, 2007) can be found at <https://osf.io/4sr3m> for Experiment 1 and <https://osf.io/haz3b> for Experiment 2.

Participants

For Experiment 1, we recruited 62 participants (44 female, 17 male, 1 other; $M_{age} = 24.40$ years, $SD_{age} = 7.02$), based on an a priori sensitivity analysis aiming for a power of 80% at an alpha level of .05 allowing us to detect any effect bigger than $f = .18$ for repeated-measures ANOVA and odds-ratios of at least 1.68 for logistic regression. We deemed this as sensitive enough to detect effects that were potentially interesting in this domain.³ For Experiment 2, we collected 60 participants (44 female, 16 male; $M_{age} = 26.92$, $SD_{age} = 9.02$). We used results from Experiment 1 for an a priori power simulation in R (R Core Team, 2019) using the simulation method described in Schad et al. (2019) as a template (the associated R-script is available on the OSF page of Experiment 2). The lower-bound of the Bayesian 95% credible interval for the effect of value distribution width on confidence (P1) from Experiment 1 was used as a conservative estimate for the effect size and the standard error of that effect was used as the sampling error for the effect size. We tested the null hypothesis that not more than 5% of the posterior density for the effect is opposite to the predicted direction in 80% of the simulations. This resulted in a sample size of 60 participants.

Materials

Six pictures of fractals from Mathôt et al. (2015) were chosen as novel objects and were randomly (across participants) matched to one of the six value distributions. Each fractal was coupled with one value distribution during the learning task described below. We chose the normal distribution as the shape for each value distribution for both methodological and theoretical considerations. First, normal distributions are conveniently described by the mean and SD , which are straightforward quantities to operationalize expected value and value uncertainty of a distribution. Second, they are theoretically compelling assuming that the experienced value of an object converges toward a specific mean from which different factors might cause random deviations (for a specific person within a specific context); a process that naturally results in a normal distribution (Frank, 2009). Third, they have been frequently used in related decision-making models such as risk-return models in which the risk is defined as the second moment (variance) of the probability distribution (Levy & Markowitz, 1979).

Following a procedure from Goldstein and Rothschild (2014), we constructed value distributions by simulating six large ($N = 1,000,000$) normal distributions. The random draws from each distribution were ordered and each 10,000th value was taken (starting at the 5,000th value) so that the resulting sample of 100 values of each distribution would closely resemble the shape of a normal distribution.⁴ The specific distributions that were sampled from are D1($\mu = 80$, $\sigma = 15$), D2($\mu = 100$, $\sigma = 20$), D3($\mu = 120$, $\sigma = 30$), D4($\mu = 130$, $\sigma = 40$), D5($\mu = 150$, $\sigma = 10$), and D6($\mu = 160$, $\sigma = 5$).

These distributions were chosen as their combination exhibited rather low correlation ($r = -.25$) between absolute value and spread of the distributions, which we intended to minimize but could not entirely eliminate due to other constraints, such as our desire to use various different distribution widths rather than for example, only two to demonstrate that the effect has some linearity to it rather than to demonstrate a mere difference between, for example, two groups. This minimized confounding influences of risk aversion on evaluations, and of absolute value on confidence. It has for example been shown that evaluations and confidence about evaluations are correlated for real-life objects such as food items (Polanía et al., 2015), which is one of the core elements that the present research strived to disentangle. Thus, each value distribution contained 100 sampled values from the defined normal distribution. Across distributions, the individual values ranged from 0 to 225 eurocents. Figure 1 provides an overview of the fractals and value distributions.

Procedure

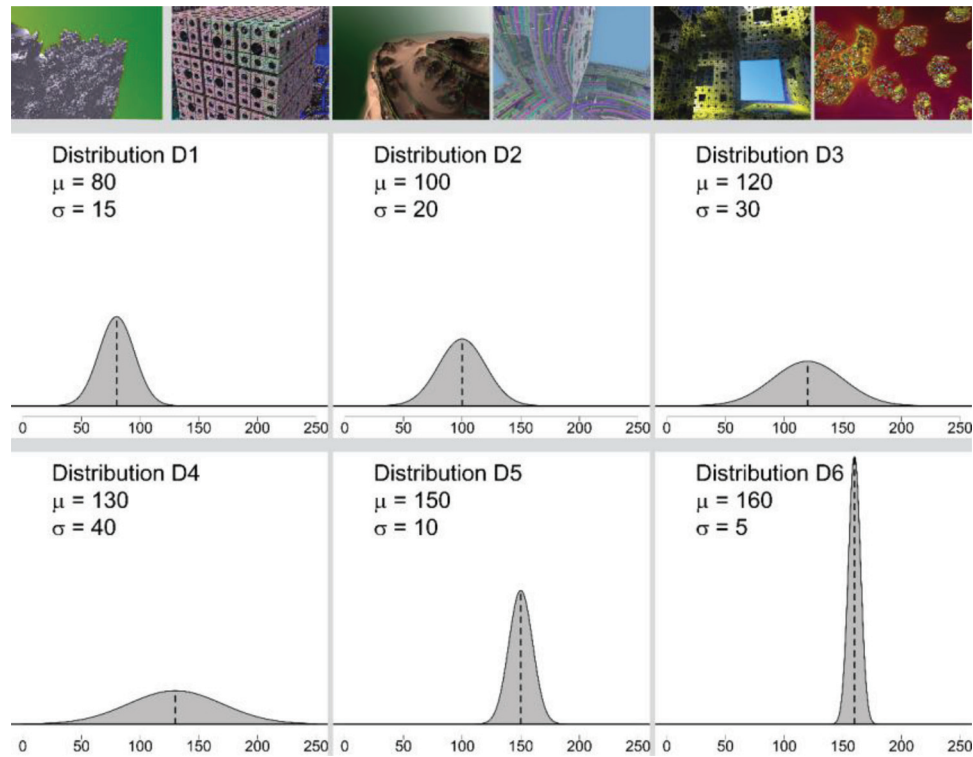
The described value distributions and fractals were used in four experimental tasks. Participants completed the first three tasks (a learning task, an evaluation and confidence-judgment task, and a distribution builder task) for the first fractal before starting with the second fractal and so forth. After completing the first three tasks for all fractals, participants engaged in a binary choice task. Figure 2 provides an overview of the experimental procedures. Before partaking in the actual experiment, participants saw detailed instructions for each task and completed a short practice round to familiarize themselves with the procedures, as explained below. They were given the chance to revisit the instructions and rehearse each task before starting the actual experiment. In total, the experimental procedure took about 45 minutes and participants were rewarded with €7.50 to €10 depending on their reward from the binary choice task.

Learning Task. The first task in Experiments 1 and 2 was a learning task in which participants saw a fractal and 100 values from one of the six value distributions (randomly assigned to the respective fractal). Before starting the task, we informed participants that the fractal that they were about to see identified a money bag that is attached to it. They were informed the money bag contains a large number of coins of different values of which 100 coins will be presented to them at random. We also informed them that it was important to learn the association between fractals and money bags as they would choose one fractal later and one random coin from the money bag would be paid out in addition to their compensation. We presented the values in a rapid serial visual presentation stream paradigm (Kunar et al., 2017; Tsetsos et al., 2012) in which the 100 values within the value distribution were

³ We conducted a frequentist power analysis for Experiment 1, because of the lack precise information about the to-be-expected estimates to conduct a more precise Bayesian simulation-based power analysis. After using the information gained from Experiment 1, we conducted a more appropriate Bayesian power analysis for Experiment 2. Note however, that the sample-sizes of both power analyses ended up being very similar suggesting that the initial approach was not a bad estimate.

⁴ As a reviewer pointed out, a more elegant solution to this problem is to directly utilize percentiles to construct the distribution without the need for sampling.

Figure 1
Overview of Materials in Experiments 1 and 2



Note. The fractals shown above were randomly matched with one of six distributions (the densities of the underlying sampling distribution is shown here) that were varied in their mean and *SD*. From each distribution, 100 values were sampled and presented to participants during the learning phase. See the online article for the color version of this figure.

flashed alongside one of the fractals in random order. Each value was shown for 600 milliseconds with 200 milliseconds breaks between values resulting in a learning phase of 80 seconds per fractal. This procedure was implemented to simulate learning by experience which we assume underlies the value learning for actual stimuli such as foods.

Evaluation and Confidence Judgments. In the subsequent task, participants were asked to rate the average value of the fractal on a circular rating scale comparable to the scale used by Kvam and Pleskac (2016). The circular shape was meant to enable us to assess whether wider value distributions result in longer evaluation times, as the distance between the starting position of the mouse cursor and each point on the circular line were identical. The scales had a precision of 200 points. For each trial, participants were asked to put the mouse-cursor in a red box that was at the center of the half circle that represented the rating scale. In total each item was rated five times in five different blocks consisting of four trials of two different trial types presented in random order:

(1) In one out of four trials, the *evaluation trials*, the previously learned fractal was presented to participants. Here, participants were asked to indicate the average value of the money bag. To give their evaluation, participants needed to cross the rating scale at the intended position that best reflected their assessment of the average value. Eleven tick marks were presented alongside the scale ranging from 0 to 200 in steps of 20. All evaluation trials

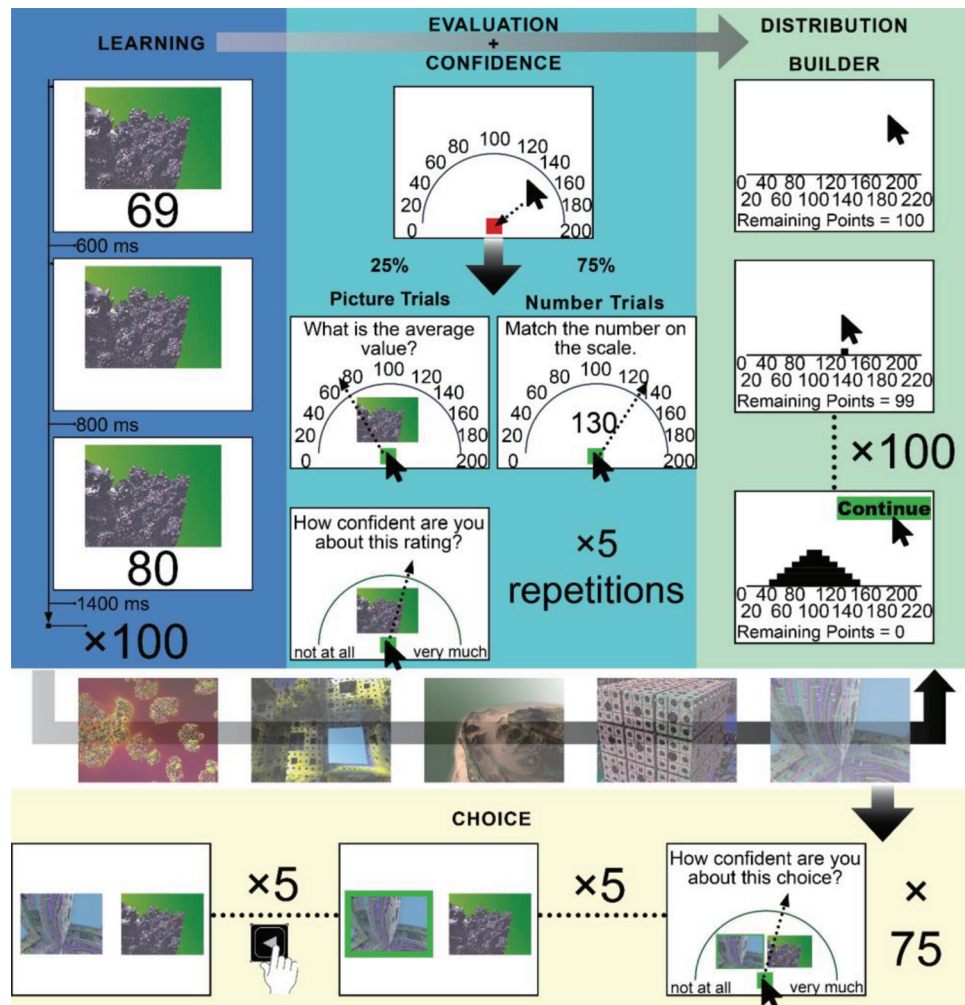
were immediately followed by a *confidence-judgment trial* in which participants were asked to indicate their confidence on the same circular scale with scale anchors at 0 = *not at all confident* and 200 = *very confident*.

(2) On the remaining three out of four trials, the *number-matching trials*, participants saw a number representing an arbitrary point on the scale and were asked to cross the scale at the position indicated by the number. These trials were used to check whether participants' motor accuracy was similar across the scale and to reduce the influence of previous ratings on the follow-up ratings. This is, when the ratings occurred directly after each other, participants might see the first trial as an actual evaluation and merely try to give the same motor response on subsequent trials. While this is not ruled out by intertwining the number-matching trials, the memory load of repeating the previous evaluations on a motor level is certainly increased. Thus, when giving subsequent ratings, this increases the chance that each trial represents a (new) evaluation.

Thus, each fractal was evaluated five times followed by five confidence judgements, which were randomly intertwined with three number-matching trials across five blocks.

Distribution Builder. Following the evaluation and confidence judgment task, participants engaged in an adapted version of the distribution builder task from Sharpe et al. (2000). In this task, participants were asked to imagine that they were to draw 100 new coins from the money bag (i.e., value distribution) represented by a

Figure 2
Procedure of Experiments 1 and 2



Note. Learning. First (blue [left-hand] panel), one by one, the individual values from each distribution were presented in the learning phase. Evaluation and Confidence. Second (turquoise [middle] panel), participants were asked to rate the average value of the picture based on the underlying value distribution and state their confidence in their ratings on a circular rating scale. On two thirds of the trials during the rating task, number-matching trials were presented in which numbers must be matched on the scale. Distribution Builder. Third (green [right-hand] panel), participants recreated the value distribution in the distribution builder task by distributing 100 points based on their expectation about the next 100 coins that would be presented for the specific distribution. Binary Choice. After completing these three tasks for all pictures and distributions, participants completed an incentivized choice task and postchoice confidence judgements (yellow [bottom] panel). Fractal pictures adapted from “Large pupils predict goal-driven eye movements.” by S. Mathôt, A. Siebold, M. Donk & F. Vitu, 2015, *Journal of Experimental Psychology: General*, 144(3), pp. 513–521. Copyright CC BY 3.0, 2015 by the authors. Adapted with permission. See the online article for the color version of this figure.

given fractal. Furthermore, they were asked to place each of these 100 imaginative coins on the scale, stacking coins that they would expect to draw multiple times, thereby creating a distribution that accurately resembled the value distribution that they learned earlier. Building a distribution was achieved using the computer mouse by clicking above the scale at the respective position on the scale where they wanted to add a coin. In case participants felt like having made a mistake, coins could be removed and redistributed by clicking below the scale at the respective position (these removed coins could then again be distributed). Once they were satisfied

with all 100 distributed coins, they could click on *continue*. This task was completed once for each fractal.

Binary Choice Task. After completing the previous three tasks for each fractal, participants completed a binary choice task in which they chose between all possible combinations of two fractals. Each pair was repeated five times while switching presentation sides of the choice options after each block. Participants were informed that they should try to always choose the fractal that they thought would yield a greater monetary reward. They knew that one choice trial would be randomly selected and from the money

bag associated with the fractal chosen in that trial, one coin would be randomly selected and the respective amount would be paid out to them at the end of the experiment. The earnings were rounded up to the closest 50 eurocents mark adding 50 to 250 eurocents to their compensation. For each choice, participants were asked to select the left or right image by pressing the corresponding arrow key on the keyboard followed by a green frame around the selected picture for 500 milliseconds. In Experiment 1, there was no time limit for choosing. In Experiment 2, choices had to be made within 3 seconds (to make choices more comparable to food choices for which people often decide rather quickly; Chen et al., 2019) and were repeated later if participants did not choose within this time window. Each choice was immediately followed by a postchoice confidence assessment requiring participants to indicate their *postchoice confidence*, that is, how confident they were that they chose the higher value fractal, on the same circular scale that was used for the evaluations and confidence judgements. During the confidence assessment of Experiment 1, participants were only presented with the picture that they selected in the previous choice. In Experiment 2, we changed this to a presentation of both choice options with a green frame around the chosen picture to put more emphasis on the comparison between the two items during confidence judgements.

Exclusion Criteria

We defined the following participant-based and trial-based a priori exclusion criteria in the preregistrations: Single trials were excluded when a response-time on a trial in any of the tasks was less than 50 milliseconds or more than 10 seconds. Participants were excluded from all analyses when we observed: (a) An average deviation of more than 50% of the scale width on at least half of the number-matching trials; that is, trials on which numbers are displayed indicating at what value the scale needs to be crossed in the rating task. This was done to ensure that participants could use the scale with reasonable precision. (b) An average difference of less than 10% of the scale on the picture ratings. That is, when despite the manipulated differences in value, all fractals were rated identically (the averaged ratings across the five trials per fractal were compared), the participant was excluded. (c) More than 50% of the response times were 50 ms or shorter.

The participant-based criteria did not apply in either experiment. However, one participant was posthoc excluded from the analyses in Experiment 1 as the person was not sufficiently fluent in English and needed translations from the experimenter to understand the task. This resulted in long breaks between the different parts of the experiment (e.g., learning and evaluation and confidence-judgment task), which is why the person was excluded from analyses. This resulted in final sample sizes of 61 and 60 participants, respectively. The trial-based criteria were applied separately for each analysis.

Data Analysis

All research questions and hypotheses were preregistered on the Open Science Framework. The preregistered hypotheses are exhaustively discussed in the article for Experiment 2 and Experiment 3, while for Experiment 1, some of the preregistered hypotheses are omitted from the article for the purpose of readability but are reported in the online supplemental materials (Section 2). The online supplementary materials also contain detailed information on model fitting and model selection (Section 1). It should also be

noted that in the preregistration we mentioned the use of Bayesian mixed-effects models, but without the model family. Different model families lead to similar results, and we opted for the model with the best fit, based on visual inspection and estimated log posterior density as implemented in the *loo* package (Vehtari et al., 2019). This is partly because some aspects of the data were difficult to predict and a model that captures these aspects (e.g., skewness) can make the estimates less prone to outliers and provide a better representation of the data-generating process as opposed to, for instance, Gaussian models that might be prone to overfitting the mean and variance of the response distribution while failing to capture other aspects of the response distribution such as its skewness. Thus, while the preregistration does not specifically mention some aspects of the models, it can be seen as a logbook that clearly outlines our knowledge about the project before data collection while the steps taken during model-evaluation are described in detail in the online supplemental materials (Section 1).

All data-analyses were conducted in the Statistical Software R (Version 3.6.1; R Core Team, 2019). Specifically, Bayesian mixed-effects models were estimated in the probabilistic programming language Stan (Stan Development Team, 2016) and interfaced via the *brms* package (Version 2.10.0; Bürkner, 2018). For all models, we followed a maximal model approach (Barr et al., 2013) by including crossed random intercepts for participants and either stimulus or choice pair. Random slopes on both random intercepts were added for all predictors in the model. Model families were chosen based on theoretical assumptions and model fit so that we ended up with three different model families for the three different dependent variables that were used:

(a) Evaluations and confidence judgements (P1, P2b, and P3) were modeled as Beta-Binomial distributions. For these analyses, each observation constitutes one of the five evaluations or confidence judgements given by a participant for a particular item. To account for the dependence of the five repeated measurements for the same item per participant, random slopes were added for all focal predictors across both, participant ID and item ID, which were added as random intercepts to the model. (b) *SDs* of evaluations in P2a were implemented as lognormal models where each observation is the *SD* of the five evaluations for each item per participant.

For the evaluation and *SD* models, the reported results were checked for robustness by fitting Gaussian and skew-normal models. For all models, we used weakly informative priors that are described in the online supplemental material (Section 1). We interpreted the results based on the proportion of posterior density (*pp*) of the focal predictor in the direction opposite of the prediction (*pp*₋ for expected positive estimates and *pp*₊ for expected negative estimates). Note that, if the reader desires to infer a more traditional statistical significance criterion based on this number, for example, $p < .05$, the criterion needs to be divided by 2 (e.g., $p < .025$ instead of $< .05$), as the posterior density statistics are one-sided. However, as we prefer more fine-grained labels based on the posterior distribution to binary decisions about the presence or absence of an effect, we use the labels *somewhat credible* for posterior probabilities (*pp*₊ or *pp*₋) between .025 and .01 (between 25 and 10 out of every 1,000 samples from the posterior are in the opposite direction to what we expected), *credible* between .01 and .001 (between 10 and 1 out of every 1,000 samples from the posterior are in the opposite direction to what we expected) and *highly credible* for anything $< .001$ (less than 1 out

of every 1,000 samples from the posterior are in the opposite direction to what we expected).

We analyzed the data using Bayesian mixed-effects models. For Experiments 1 and 2 we report the respective estimates as Exp1 and Exp2 in addition to a combined estimate (Combined) for the combined data of both experiments. The combined estimates are reported in text and figures, while the estimates for the main predictions of Experiment 1 and Experiment 2 are presented in Table 1.

The rationale behind combining the data sets is that the from a Bayesian perspective, given that the two experiments were almost exact replications, the combined data provide the best posterior estimate for the effects of interest as it includes all collected data. We additionally estimated an effect for experiment number (i.e., Experiment 1 or Experiment 2) in the combined model and an interaction between the experiment number and the focal predictors. If this interaction is found to be credible in combination with a noncredible effect for the focal predictor in either experiment in isolation, this might suggest that the effect in the combined data is driven by one experiment alone. If the interaction is credible and in both experiments the effect of the focal predictor is also credible, then the interaction suggests a stronger effect in one of the experiments. If the interaction is not credible, the effects in both experiments do not differ from each other, meaning that the combined results are not driven by either experiment alone. For brevity, we only report significant interactions and what they mean for the interpretation of the results (this was only the case for Question 1; the full report of the experiment number effects and interactions can be found in the online supplemental materials, Section 5). All data, modeling code in R including prior specifications and MCMC settings, extensive analysis reports, programs and other files related to the project are available on the OSF at <https://osf.io/q72sm/>.

Results

Q1: Relation Between Variation in Experienced Values and Confidence

We tested whether the width of the manipulated value distributions in Experiments 1 and 2 influenced how confident people were in their evaluations (henceforth *postevaluation confidence*) of the fractals. As preregistered, to address the possibility that these results were a mere reflection of inaccurate learning of the value distribution, we controlled for participants' learning accuracy (i.e., the number of coins that participants distributed in the distribution builder task that matched the learned distribution).

Table 1
Results From Individual Analyses of Predictions 1 to 3 for Experiments 1 and 2

Prediction	Experiment	Estimate	CI	<i>pp</i> _{+/-}
1	1	-0.15	[-0.23, -0.06]	< .001
1	2	-0.29	[-0.40, -0.18]	< .001
2a	1	0.05	[-0.02, 0.12]	.082
2a	2	0.04	[-0.05, 0.12]	.159
2b	1	-0.03	[-0.05, -0.01]	.004
2b	2	-0.02	[-0.04, 0.00]	.038
3	1	-0.02	[-0.03, -0.01]	.002
3	2	-0.02	[-0.03, -0.01]	.002

Note. CI = credible interval; *pp*_{+/-} = proportion of posterior that is opposite to the predicted direction.

Learning accuracy was on average 67.19 out of 100 coins ($SD = 15.28$) in Experiment 1, and 66.81 coins ($SD = 16.16$) in Experiment 2 ($M = 67.00$, $SD = 15.72$ in the combined data). P1 was confirmed with highly credible effects of value-distribution width on postevaluation confidence (*Estimate* = $-.21$, 95% CI [-0.25, -0.17], *pp*₊ < .001, Figure 3). This effect was larger in Experiment 2 compared to Experiment 1 (see Table S7 in the online supplemental materials). Interestingly, there was no credible influence of learning accuracy on postevaluation confidence (*Estimate* = $.01$, 95% CI [-0.00, .02], *pp*₋ = .070.⁵ This suggests that postevaluation confidence is related directly to the variance of the value distribution instead of representing participants' ability to accurately represent the value distribution.

Q2: Relation Between Value Distributions, Rating Variability, and Confidence

To investigate whether more diverse experiences resulted in increased variability in evaluative ratings, we predicted the SD of the five evaluations that each participant provided for each fractal by the manipulated value distribution width in Experiments 1 and 2.⁶ In contrast to P2a, we found no credible influence of the manipulated value distribution width on the rating variability (*Estimate* = $.04$, 95% CI [-0.00, .08], *pp*₋ = .033; Figure 4A). The effect became somewhat credible when excluding extreme evaluation SD s (*Estimate* = $.04$, 95% CI [.00, .07], *pp*₋ = .015; Figure 4B).

For P2b (not preregistered for Experiment 1), the prediction that higher variance of evaluations would be related to lower confidence judgements was credible in the combined data (*Estimate* = $-.02$, 95% CI [-0.03, -0.01], *pp*₊ = .002; Figure 4C)⁷ and remained somewhat credible after excluding extreme evaluation SD s in the combined data (*Estimate* = $-.15$, 95% CI [-0.27, -0.03], *pp*₊ = .013; Figure 4D).

Q3: Relation Between Value Distribution Overlap and Postchoice Confidence

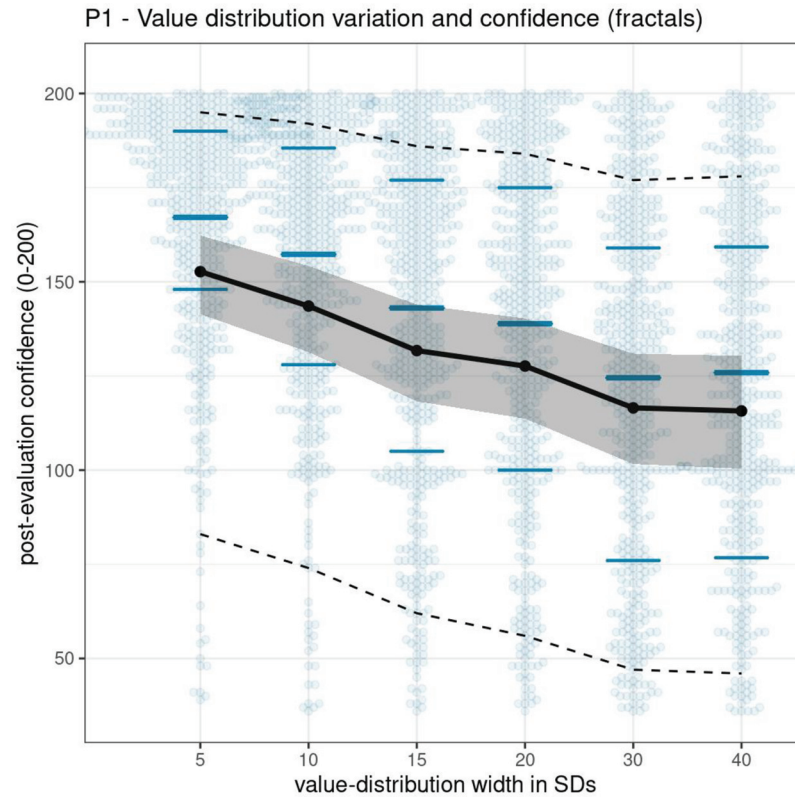
We tested the prediction that higher overlap between two value distributions in a choice pair (i.e., how many points the distributions of the choice alternatives share) would result in lower postchoice confidence. We added value difference as a predictor to the model, to exclude the possibility that the variance explained by overlap is inflated by the shared variance with value difference. This was not preregistered but makes the reported analyses less likely to yield

⁵ Note that the estimates seem numerically very small. This is because beta-binomial models are fitted on the log-scale. Therefore, the estimates must be exponentiated to be transformed back to the original scale. For instance, in this case the Estimate of 0.01 in Experiment 1, after transformation back to the original scale, can be interpreted as a difference of 50 points on the outcome as learning accuracy increases from 0 to 100. The online supplemental materials (Section 4) demonstrates this calculation. However, as effects can be non-linear on the response scale, the transformed estimates can easily be misinterpreted. Therefore, we report untransformed parameters and instead plot the results on the original response scale of the outcome variable.

⁶ As SD s can be unstable with only five observations, a reviewer suggested repeating the analyses for P2 with the range of ratings rather than the SD . The results converge with the ones reported in the paper and can be found in the online supplemental materials (Section 9).

⁷ Again, these estimates seem numerically small as they present log-odds changes in centered rather than standardized predictors. However, as presented in Figure 4C, the decrease in confidence as a function of rating variability is still substantial (from about 140 at a rating SD close to 0 to about 60 at a rating SD of 100).

Figure 3
Relation Between SD of Value Distributions and Evaluation Confidence for the Combined Data of Experiments 1 and 2



Note. Blue (grey) dots represent raw data, thick blue (dark grey) lines represent observed means and thin blue (dark grey) lines the 25th, and 75th percentile. Black solid lines and (light) grey shaded areas represent model estimated means and 95% credible interval. Dashed black lines represent the 95% predictive interval. See the online article for the color version of this figure.

results that are confounded by value difference (see online supplemental materials, Section 7). For the combined data of Experiments 1 and 2, we found credible evidence for an effect of overlap on post-choice confidence ($Estimate = -.02$, 95% CI $[-.02, -.01]$, $pp_+ < .001$; Figure 5A). There was also a credible negative effect of value difference, suggesting lower confidence for items that had a larger value difference, but this effect might be difficult to interpret due to possible collinearity concerns (see online supplemental materials, Section 7; $Estimate = -.01$, 95% CI $[-.02, -.01]$, $pp_+ < .001$).

Experiment 3

Method

In a third study, instead of manipulating distributions of novel fractals, we used familiar food items as stimuli of which we measured the distributions. The preregistration of the study and its implementation can be found at <https://osf.io/vdn2s>.

Participants

In the absence of a good estimate of the expected effect size based on previous data on food items and in line with Experiment

1, another 61 participants were invited to the lab (48 female, 13 male; $M_{age} = 23.28$ years, $SD_{age} = 6.24$).

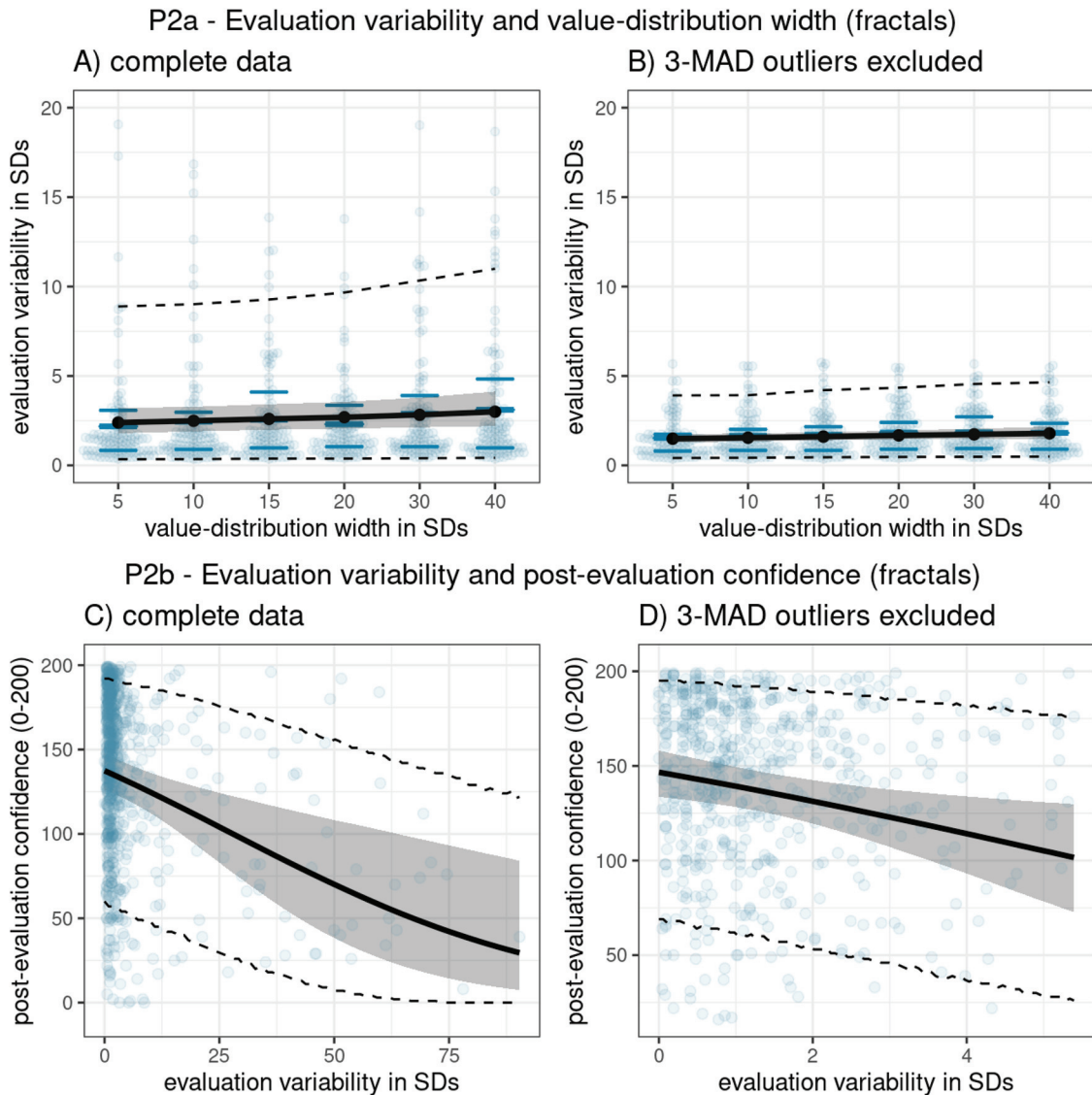
Materials

We used pictures of 24 food snacks by Veling et al. (2017) that were of about equal size and comparable prizes that were all available in common supermarkets in the Netherlands. The selected food snacks were taken from four different categories: six vegetables (e.g., baby carrots), six fruits (e.g., apple), six savory snacks (e.g., potato chips), and six sweet snacks (e.g., chocolate). All food pictures can be found in the materials provided with the pre-registration link.

Procedure

The procedures were highly similar to the ones of Experiments 1 and 2. Most important, however, Experiment 3 did not involve a learning phase as the food snacks that were chosen were very common in the Netherlands and likely well-known by the participants. Thus, the value of each food snack was based on previously learned experiences with these snacks. Moreover, the use of food items with preexisting values allowed us to present all items in a task before administering the next task, as participants could not

Figure 4
P2 for Combined Data of Experiments 1 and 2



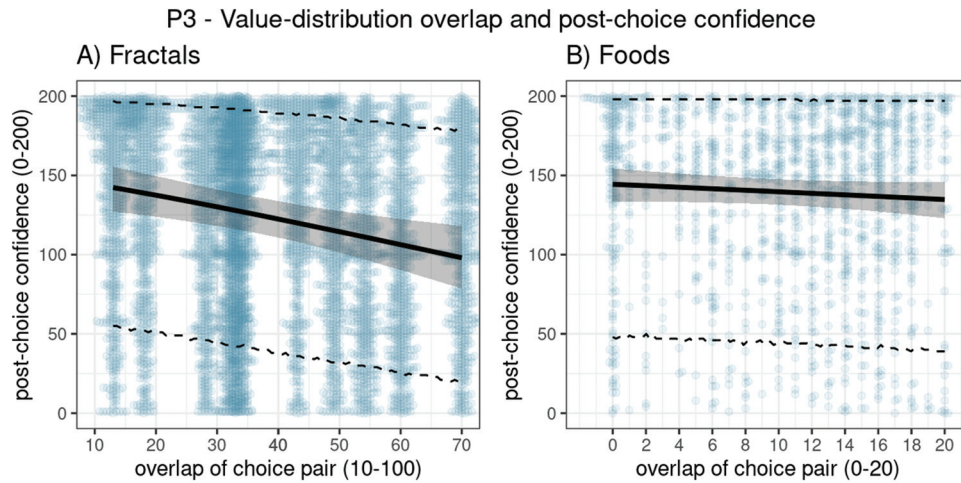
Note. (A–B) Prediction (P)2a: evaluation variability as a function of value distribution width with and without 3-MAD outliers on the outcome excluded. Blue (grey) dots represent raw data. Black solid lines and (light) grey ribbons represent model estimated means and 95% credible interval. Dashed black lines represent the 95% predictive interval C-D) Prediction (P)2b: confidence as a function of *SD* of evaluations with and without 3-MAD outliers on the predictor excluded. Black solid lines and (light) grey ribbons represent model estimated means and 95% credible interval. Dashed black lines represent the 95% predictive intervals. 3-MAD criterion = three Median-Absolute-Deviation criterion. See the online article for the color version of this figure.

get confused about the values that may be connected to the different items. There were a few additional differences that are discussed for each task separately.

Evaluations and Confidence Judgements. Each food picture was presented for an evaluation and confidence judgment four times in a block-wise randomized order. The rating task in Experiment 3 did not include any number-matching trials as the repeated evaluations for specific food items were separated by evaluations of other food items reducing the probability that participants would base their current evaluation on a previous evaluation of the same food item.

Distribution Builder. As we did not manipulate any values in Experiment 3, there was no objective criterion for a correct representation of the value distribution. Thus, the distribution builder task was conceptually different in that participants had to not recreate, but elicit from scratch, a distribution that they thought would represent the value distribution of the food items. We instructed participants to imagine consuming a food item multiple times and that across these consumptions, the liking of a food item might vary. To simulate this experience, we asked them to build a distribution from 20 hypothetical consumptions of each food item

Figure 5
P3: Relation Between Value Distribution Overlap and Postchoice Confidence



Note. A) Combined data of Experiments 1 and 2. Note that overlap starts at 10 points as lower-overlap choice pairs were excluded (preregistered). B) Experiment 3. Black solid lines and (dark) grey shaded areas represent model estimates of mean confidence and 95% credible interval. Dashed black lines represent 95% predictive intervals. See the online article for the color version of this figure.

on a 200-point scale from *not at all appealing* to *very appealing*. We reduced the number of points to be distributed to 20 (as opposed to 100 in Experiments 1 and 2) to reduce the duration of the task with 24 food pictures. The range of this scale was chosen in line with prior research on food evaluations (Chen et al., 2016, 2019; Quandt et al., 2019)

Binary Choice Task. Due to time constraints, we could not present participants with all possible combinations of the 24 food items. Instead, we matched pairs based on participants' individual ratings of the food items to create pairs with relatively small value differences. The matching process was restricted to other criteria for some of the choices that were important to answer additional research questions, namely healthiness and sweetness, but are not relevant for the present study. This resulted in 36 choice pairs per participant, of which 12 were constructed based on the smallest possible mean-value difference in the distribution builder task, 12 pairs based on the smallest possible value difference within each health category of the food pictures (either both healthy or both unhealthy), and 12 pairs that were matched on values across health categories (creating healthy vs. unhealthy pairs). Each choice pair was repeated once with counterbalanced positions of the choice alternatives. Before the task, participants were informed that choices would be consequential in that one of the choices would be paid out to them after the experiment in the form of one of their chosen food snacks that they could take home. To this end, we added two more filler choice pairs between four different food snacks that we had available in the lab and that were always used as the consequential choices to minimize the diversity of food that we had to have available in the lab and therefore to minimize food waste due to expiration of products.

Food Property Questions. For exploratory purposes, Experiment 3 included a questionnaire in which people were asked about different properties that they ascribe to the respective food. Participants were asked to rate their familiarity with the food ("How

often do you eat this food?") on a 10-point scale from *never* to *very often* and to rate their experienced similarity of the food's taste ("If you would eat this food multiple times, how similar would it taste each time?") on a 10-point scale from *always different* to *always the same*. Moreover, participants rated the foods complexity. They were presented with 11 different attributes (sweet, salty, sour, umami, bitter, watery, spicy, mild, bland, nutty, smoky) and indicate which of these attributes they ascribed to the food item by checking the respective boxes. The number of checked boxes were added up to a score from 0 to 11 (where higher scores are interpreted as higher complexity).

Exclusion Criteria

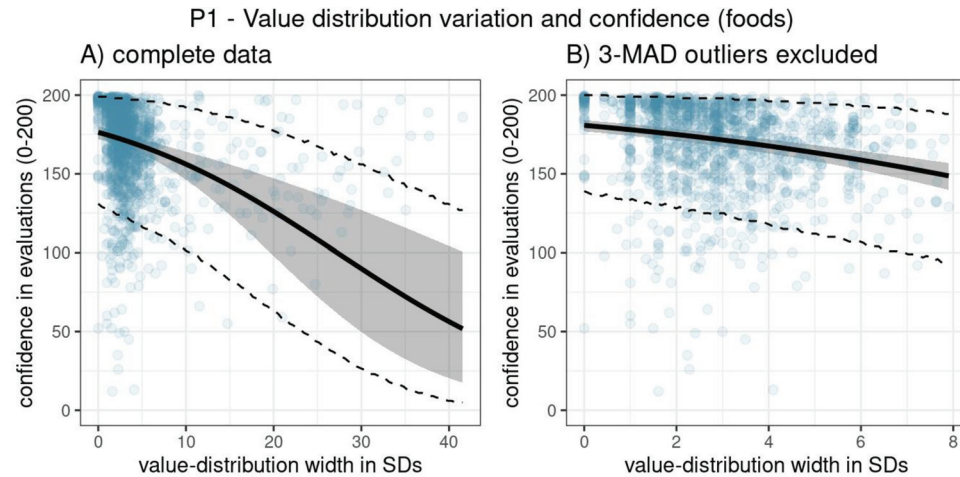
We preregistered that if two evaluative ratings within a food item within a participant differed by more than 100 points on the rating scale (50% of the scale width), this item was omitted from all analyses for that participant. The reasons for this were that on the one hand, extreme differences could point to bi- or multimodal value distributions which is a more general application of value distributions that we do not investigate here. On the other hand, extreme value differences could hint at other processes that might shadow or inflate the effects that we are interested in here.

Results

Q1: Relation Between Variation in Experienced Values and Confidence

We predicted that measured variation in experienced values, as assessed by the distribution-builder task, was related to confidence in evaluations of natural food items. This prediction was confirmed (*Estimate* = $-.07$, 95% CI [$-.11$, $-.05$], $pp_- < .001$; Figure 6A). As Figure 6A shows, there were a few distributions with very extreme *SDs* leading to inaccurate predictions for those

Figure 6
Relation Between SDs of Value Distributions Created by Participants in the Distribution Builder Task and Confidence in Their Previous Evaluations of Food Items in the Evaluation Task



Note. Blue (grey) dots represent raw data. Black solid lines and (light) grey ribbons represent model estimated means and 95% credible interval. Dashed black lines represent the 95% predictive interval. Panel A shows the complete data. Note that due to only very few observations with high *SD* the model does not represent those observations well. Panel B shows data with extreme values for the predictor being excluded, based on a 3-MAD criterion. Please note the different x-axis ranges and that the fit curves are not linear on the response scale as the beta-binomial models that are used are fitted on the log scale. 3-MAD criterion = three Median-Absolute-Deviation criterion. See the online article for the color version of this figure.

extreme values. We therefore exploratively repeated the analysis without these extreme observations based on a three Median-Absolute-Deviation (MAD) criterion (Leys et al., 2013). This did not change the results (*Estimate* = $-.15$, 95% CI [$-.18$, $-.11$], $pp_- < .001$; Figure 6B).

Q2: Relation Between Value Distributions, Rating Variability, and Confidence

We found highly credible support for a relation between measured distributions and evaluation variability in line with P2a (not preregistered; *Estimate* = $.07$, 95% CI [$.04$, $.12$], $pp_- < .001$; Figure 7A). Excluding outliers did not change the results (*Estimate* = $.15$, 95% CI [$.09$, $.20$], $pp_- < .001$; Figure 7B). Moreover, in line with P2b, there was a highly credible relation between this rating variability and confidence (*Estimate* = $-.03$, 95% CI [$-.03$, $-.02$], $pp_+ < .001$, Figure 7C), which remained highly credible after excluding outliers (*Estimate* = $-.03$, 95% CI [$-.04$, $-.02$], $pp_+ < .001$; Figure 7D).

Q3: Relation Between Value Distribution Overlap and Postchoice Confidence

There was no effect of overlap (*Estimate* = $-.01$, 95% CI [$-.03$, $.01$], $pp_+ = .107$; Figure 5B) nor of value difference (*Estimate* = $-.01$, 95% CI [$-.03$, $.01$], $pp_+ = .224$) on postchoice confidence.

Alternative Implementation of Choice Models

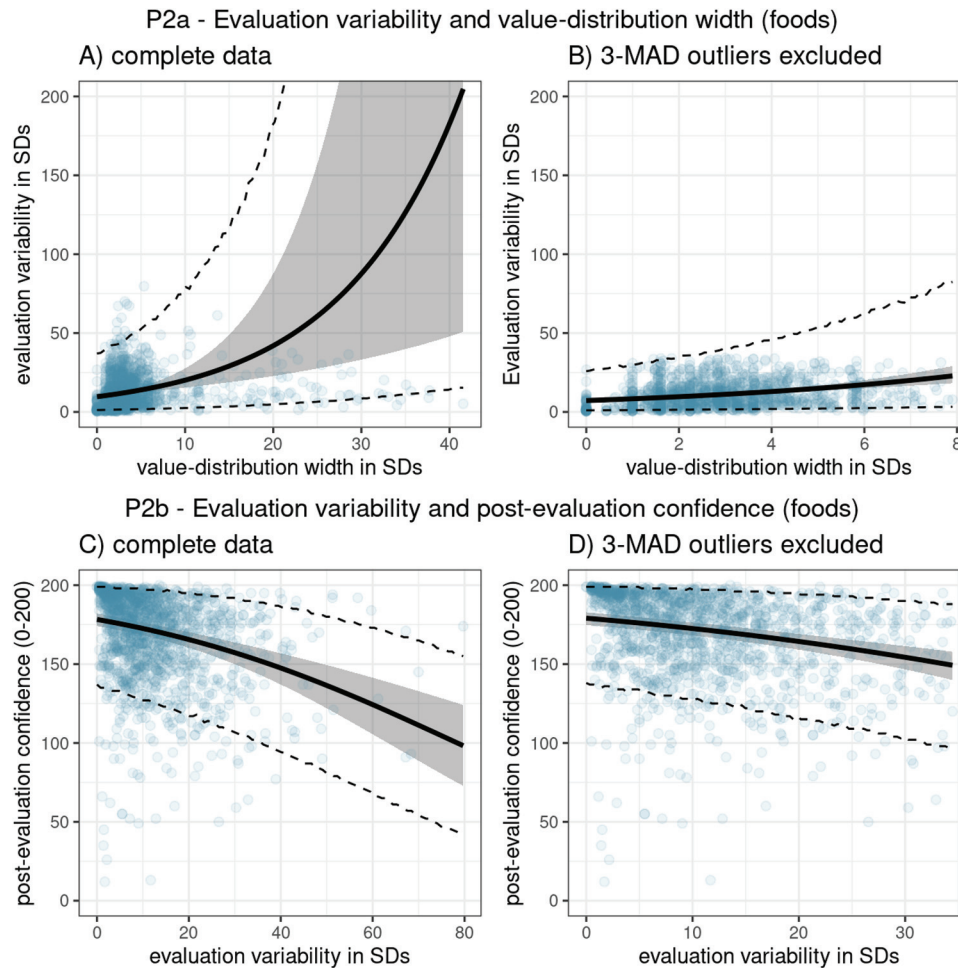
To further explore how variance of value distributions relates to postchoice confidence, we implemented alternative models of how combinations of value distributions might influence choices and postchoice confidence. Following reasoning that value-based decisions

might be inherently comparative in nature (Lim et al., 2011); we tested whether postchoice confidence could be predicted from a value-difference distribution. Thus, instead of predicting postchoice confidence from the average value difference and overlap between two distributions, we combined the two distributions into one value-difference distribution that represents a collection of possible value differences that might be sampled during decision-making. The mean of the resulting value-difference distribution encodes the expected value difference between the choice alternatives; the *SD* encodes the variance of possible value differences. For Experiments 1 and 2, the manipulated distributions, that is, the objective distributions that were presented in the learning phase, were combined and for Experiment 3, value-difference distributions were obtained by combining elicited distributions from the distribution builder. For details on the combination of distributions see the online supplemental materials (Section 3). For Experiment 3, we used the data containing all choice pairs (see the online supplemental materials; Section 8).

Interestingly, we found no clear association between postchoice confidence and the mean value of the value-difference distribution of the fractals (*Estimate* = $.00$, 95% CI [$-.00$, $.01$], $pp_- = .0520$). Thus, the average value difference between two item's value distributions was not related to postchoice confidence. Instead, there was a credible association with its *SD* (*Estimate* = $-.01$, 95% CI [$-.01$, $-.00$], $pp_- = .002$). Thus, postchoice confidence varies independently from value difference and is determined mainly by the variance of the value-difference distribution.

For food choices in Experiment 3, the mean (*Estimate* = $.01$, 95% CI [$.01$, $.01$], $pp_- < .001$) but not the *SD* (*Estimate* = $-.01$, 95% CI [$-.02$, $.00$], $pp_+ = .062$) of the value-difference distribution was credibly related to postchoice confidence. To check

Figure 7
P2 for the Food Items in Experiment 3



Note. A–B) Prediction (P)2a: evaluation variability as a function of value distribution width with and without 3-MAD outliers on both variables excluded. C–D) Prediction (P)2b: postevaluation confidence as a function of evaluation variability with and without 3-MAD outliers on the predictor excluded. Blue (grey) dots represent raw data. Black solid lines and (dark) grey shaded areas represent model estimates and 95% credible interval. Dashed black lines represent the 95% predictive interval. Black solid lines and (light) grey ribbons represent model estimated means and 95% credible interval. Dashed black lines represent the 95% predictive intervals. 3-MAD criterion = three Median-Absolute-Deviation criterion. See the online article for the color version of this figure.

whether these results were robust, we conducted another exploratory analysis with the mean difference and pooled *SD* of the ratings to serve as a measure of mean and *SD*. Both predictors were related to postchoice confidence (mean: *Estimate* = .01, 95% CI [.00, .01], $pp_- < .001$; *SD*: *Estimate* = $-.00$, 95% CI [$-.01$, $-.00$], $pp_+ = .012$).

Additional Factors in Confidence for Foods

Additional exploratory analyses were conducted to enrich our understanding of confidence in food items. As pointed out by a reviewer, the food items, as opposed to fractals, have more dimensions on which they can be evaluated (e.g., different tastes). Thus, it is conceivable that wider distributions in the distribution builder

task, and higher variability of evaluations, would reflect the complexity of the food items. Similarly, the distribution builder *SD* and evaluation variability could reflect the familiarity with a food rather than the similarity in taste. To investigate these possibilities, we used the food property questionnaire, that we included for exploratory purposes in Experiment 3 (see Method section), to investigate whether and how food complexity, familiarity and similarity in taste would be related to value distribution variation. We found that, of these three predictors, only similarity had a credible relationship with the *SDs* of the distributions from the distribution builder task (complexity: *Estimate* = .03, 95% CI [$-.02$, .08], $pp_- = .112$; familiarity: *Estimate* = $-.04$, 95% CI [$-.10$, .02], $pp_- = .084$, similarity: *Estimate* = $-.12$, 95% CI [$-.18$, $-.06$], $pp_- < .001$).

This suggests that the distribution builder task *SD* does indeed tap into the similarity in taste rather than food complexity. However, an obvious limitation in the interpretation of this finding is that, by design, the question asked in the distribution builder task was also most alike to the similarity in taste question from the questionnaire. Therefore, this credible relationship here might simply suggest that both measures tap into the same underlying construct, and hence serves to provide additional support for the validity of the distribution builder task. Moreover, the familiarity item that was used (“How often do you eat this food?”) is likely confounded with liking of food items as people are unlikely to consume often what they do not like, raising caution about the interpretation of self-reported familiarity. For rating variability, we did not find a relationship with any of the food properties that we measured (complexity: *Estimate* = .04, 95% CI [−.02, .11], *pp*_− = .088; familiarity: *Estimate* = −.02, 95% CI [−.11, .08], *pp*_− = .355, similarity: *Estimate* = −.06, 95% CI [−.15, .03], *pp*_− = .092).

When directly investigating the relationship of the self-reported food properties with confidence, only familiarity with the food was a credible predictor (complexity: *Estimate* = −.02, 95% CI [−.07, .04], *pp*_− = .294; familiarity: *Estimate* = .16, 95% CI [.07, .26], *pp*_− < .001, similarity: *Estimate* = .06, 95% CI [−.01, .13], *pp*_− = .047).

Previous research has shown that lower confidence in an evaluation or choice can be related to longer response times (Holland et al., 2003; Kiani et al., 2014; Zylberberg et al., 2016). Thus, we predicted to find a relationship between value distribution variation and rating time of evaluations in Experiment 1. As we did not find the predicted relationship between value distribution variation and response time for evaluations of the fractals (see online supplemental materials, Section 2) we only investigated this prediction in an exploratory fashion in Experiment 3. In line with previous research, and in contrast to the findings for Experiment 1, we found a credible effect when predicting response times of postevaluation from value distribution *SDs* of food items (*Estimate* = .06, 95% CI [.02, .10], *pp*_− = .001) and when predicting response times from postevaluation confidence directly (*Estimate* = −.08, 95% CI [−.11, −.04], *pp*₊ < .001). Interestingly, response times were unrelated to evaluations (*Estimate* = −.02, 95% CI [−.05, .02], *pp*_− = .812). Thus, response time seems to decrease with higher confidence for evaluations of naturalistic food items. However, these results are difficult to interpret as we do not observe the same relations for fractals with a controlled learning history, making a causal claim difficult.

Discussion

First, the present research shows that confidence in evaluations of items can be causally influenced by manipulating the width of the underlying value distribution of these items. People reported less confidence in their item evaluations for items with wider distributions, and this effect was independent of the accuracy with which the distribution was learned. Remarkably, this association between value distribution width and confidence was also found for natural food items for which the underlying value distribution could not be manipulated due to preexisting individually acquired experiences. This latter finding provides further evidence that, also in the absence of an external correctness criterion, confidence is related to the quality of the evidence that an evaluation is based on.

Second, we investigated how variation in experienced values would be related to the variability of evaluations. In line with the idea of sequential sampling from memory (Johnson et al., 2007; Shadlen & Shohamy, 2016; Weber et al., 2007), we predicted that during item evaluation, people would draw samples from the value distribution. Wider value distributions should therefore result in more variable evaluations. For the novel fractals, we found that the evidence did not clearly support this hypothesis. For the food items in Experiment 3, however, there was a clear relation between the reported variation in experienced values and the evaluation variability of the same food item. In addition, evaluation variability for food items predicted evaluation confidence, while for the fractals, the evidence was less convincing (i.e., not credible in Experiment 2).

An explanation for the different results for fractals and food items may be differences in the reliance on episodic memory. Specifically, when repeatedly evaluating fractals, the first evaluation might simply be recalled, and repeated during subsequent evaluations. For food items on the other hand, it is more plausible that each evaluation is constructed during each rating, because repeated evaluations of the same food item were separated by evaluations of different food items, and therefore generating new evaluations may be relatively effortless for these familiar objects compared to remembering previous evaluations. The larger rating variability for food items compared to fractals further supports this argument. Furthermore, the rating variability in food items might not only depend on the value distribution variability but also by other factors such as the complexity in taste. While the exploratory analyses did not suggest any relation between rating variability and food complexity, similarity in taste, and familiarity it cannot be excluded that there are other factors aside from the value distribution variability that influence rating variability in foods.

Third, we predicted that postchoice confidence would be higher for choice pairs with relatively low overlap. We found clear support for this prediction for fractals, but no support in the preregistered analyses of food items. To investigate these mixed results, we examined an alternative theoretical model for how value distributions could be combined during choice. We originally assumed that, during binary choice, individuals would sample independently from both value distributions and that overlap in the distributions would result in the overall lower distributions sometimes producing higher-value samples compared to the overall higher-value distribution. However, it has been argued that the value signal during decision processes is inherently comparative (Lim et al., 2011). Thus, in an additional exploratory analysis, we assumed that individuals would directly sample value differences from a value-difference distribution. Importantly, the higher the mean of such a value-difference distribution, the stronger the evidence that one choice alternative is higher than the other. Furthermore, the wider this distribution, the lower the quality of the evidence for how much higher one alternative is than the other. Interestingly, and in line with the reasoning above, choices were mainly predicted by how different the values were from each other (the strength of the evidence), while postchoice confidence mainly depended on how narrow the value-difference distribution was (the quality of the evidence).

The presented research provides new insights regarding confidence in evaluations and value-based decisions. Based on the confirmatory as well as exploratory findings reported here, we

conclude that we provide compelling evidence for the fact that confidence in evaluations and value-based decisions is causally influenced by the quality of evidence on which the respective evaluations or decisions were based. We show that confidence in evaluations can be manipulated independently of the overall value of an item and is in fact unrelated to the mean value. This is an important new empirical finding because for natural items, confidence and overall value are often confounded, which typically makes it difficult to draw conclusions about the extent to which confidence is related to overall value (Lebreton et al., 2015; Polania et al., 2015).

We show that confidence in evaluation and value-based choices is, in principle, unrelated to value and is instead based on the variation of experienced values, which is in line with research on perceptual and factual decision making (Boldt et al., 2017; Lebreton et al., 2015; Meyniel et al., 2015; Rolls et al., 2010; Vickers & Packer, 1982). Thus, the current work provides important new insights for the field of decision making by showing that explicit confidence representations reflect the quality of evidence not only in perceptual and factual decisions, but also in value-based decisions.

That learning accuracy (how accurately people were able to reconstruct value distributions) was unrelated to confidence shows that even though the objective correctness of an evaluation might vary, the confidence still reflects the inherent quality of the evidence. This seems to be at odds with previous literature suggesting that the perceived correctness of a judgment does indeed influence confidence (Petrocelli et al., 2007; Tormala & Rucker, 2018) and that confidence is mainly informed by the clarity of an evaluation (i.e., how clear is it to me that this is my evaluation) as well as the perceived correctness of an evaluation (how correct does my evaluation seem to be given external evidence; Petrocelli et al., 2007).

However, we think that this apparent discrepancy is consistent with the confidence-as-evidence-quality framework suggested here and in other decision-making domains (Kepecs & Mainen, 2012; Pouget et al., 2016). Specifically, earlier research proposes that people infer the correctness of an attitude or evaluation through social cues such as apparent agreement with a majority group (Petrocelli et al., 2007). As we do not explicitly provide correctness information in the form of social cues or otherwise in the present studies, there is no such information that participants can use. Furthermore, as we argued here, and exemplified in Experiment 3, objective correctness information is often not available in evaluations and value-based decisions. This is different from political and personal judgements that earlier research mostly investigated. Thus, for value-based decisions, evaluation clarity seems to be the more relevant indicator of confidence. In the current situation, for example, evaluations of, and decisions between economical values, social information might be less relevant. However, future research needs to address whether external information, such as social norms would influence confidence independently of the evidence quality, as it has frequently been shown that such external information can have a strong impact on judgements (Clarkson et al., 2013; Petrocelli et al., 2007; Tormala & Rucker, 2018; Visser & Mirabile, 2004).

The fact that the reported variability in previous experiences with food was related to the variability of evaluations sheds light on the question why evaluations in psychological experiments often show variability even on very short time scales and without

contextual changes (Chen et al., 2016; Folke et al., 2016; Quandt et al., 2019; Schonberg et al., 2014). This has important theoretical and practical implications. First, research investigating preference reversals (the phenomenon of people changing their mind about a preferred option) does often rely on a broad set of different explanations for the observed switch in preference (Lichtenstein & Slovic, 2006). However, all preference reversals might eventually result from an external influence acting on the value sampling process. For instance, research has shown that manipulating the salience of specific values causes preference reversals by acting on the sequential sampling process during value-based decisions (Tsetsos et al., 2012).

Research on evaluation and value-based decisions would benefit from addressing this issue by taking into account the underlying value distributions (Izuma & Murayama, 2013). There have been repeated calls in the literature to assess evaluation confidence when measuring attitudes or judgements to infer their stability (Abelson, 1988; Holland et al., 2003; Petrocelli et al., 2007; Tormala & Rucker, 2018). The present research supports these calls and suggests that, in cases where it might be undesirable to assess confidence directly, evaluations can instead be assessed repeatedly as both, confidence and evaluation variability, likely tap into the same underlying construct, namely the evidence quality. Moreover, exploratory analyses show that food complexity, similarity in taste, and familiarity with a food item, three factors that one might intuitively expect to influence rating variability, were not clearly related to rating variability in the present research. Thus, for food items, the distribution builder task and confidence might be more useful measures of an evaluation's stability than questionnaires assessing food properties.

Limitations of the presented research should be considered. As much as it is a strength of Experiments 1 and 2 to directly manipulate the variation of experienced values with novel objects, it is unclear whether the brief presentation of monetary values that we employed accurately reflects learning in everyday-life situations. We provided some evidence for a similar mechanism by showing that many of the findings hold for the natural food items in Experiment 3. However, we do not know for certain whether the similarity in findings between the manipulated variation in experienced values for fractals and the reported variation in experienced values for food is indicative of the same causal link. Specifically, confidence in Experiments 1 and 2 could, at least partly, reflect the fact that averaging an array of numbers is more difficult if those numbers are more dispersed, thus reflecting a metacognitive judgment of difficulty rather than variation. We believe that the missing link between learning accuracy and confidence, as well as the converging findings for food items where the difficulty explanation does not apply, speak against this point, but that it cannot be ruled out.

In similar vein, postconfidence judgements were not incentivized and are not value-based in isolation (i.e., without applying them to evaluations). Thus, the relation between confidence and distribution variance in Experiments 1 and 2 might be trivial, as mathematically, confidence is defined as the inverse of variance (Parr et al., 2018). Like our previous point, this boils down to the question whether participants were experiencing the task as an evaluative or merely numerical rating. Again, we think that, while this remains a weakness of the first two experiments that cannot easily be ruled out, the comparable results for food items render it

reasonable to assume that confidence judgements are evaluative in nature in Experiments 1 and 2.

Moreover, there are some obvious differences between the learning procedure that we applied for fractals and the acquiring of values in real-life, such as the limited stimulus set in Experiments 1 and 2, the timescale of the learning process (throughout minutes vs. throughout life) and the richness of the experiences (seeing numbers on a screen vs. experiencing eating a food). All of these factors could cause the underlying processes to differ between fractals and foods. Unfortunately, especially the timescale challenge seems nontrivial and overcoming it would ask for extensive longitudinal designs in which the exact learning history of several real-life objects such as foods would need to be controlled. This seems practically and ethically challenging. One possible way forward could be to combine multiple real-life objects into ensembles (Yamanashi Leib et al., 2020) for which the value distributions might be manipulated. However, only little is known about how values are combined in ensembles and it is unclear how combining items will influence confidence.

The presented research provides a starting point for studying confidence in evaluations and value-based decisions. It shows that confidence in evaluations can be understood as a judgment of the width of a value distribution that encodes the variation of experienced values. By varying the width of the value distribution confidence can be manipulated. Moreover, eliciting value distributions for well-known items with unknown learning histories, such as food items, provides insights into the variation of experienced values and the variability of evaluations. This research offers the first causal empirical evidence that explicit confidence in value-based decisions reflects the variability of an evidence probability distribution, integrating it with other decision-making domains such as perceptual and factual decision making.

Context of Research

This research was conducted as part of Julian Quandt's PhD project about investigating how a feeling of confidence arises for value-based decisions and how it influences the stability of preferences and choice behavior. This project was inspired by the idea that individuals sample experiences from memory to construct a value during evaluations and decisions that serves as evidence, which is evaluated and expressed in a confidence judgment, and was performed to increase understanding of the instability of people's preferences, such as what they choose to eat. We hope to use the insights together with interventions that promote healthy eating, such as the go/no-go training in the future to see whether it can help to make effects more durable.

References

- Abelson, R. (1988). Conviction. *American Psychologist*, 43(4), 267–275. <https://doi.org/10.1037/0003-066X.43.4.267>
- 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, e46080. <https://doi.org/10.7554/eLife.46080>
- Bakkour, A., Zylberberg, A., Shadlen, M. N., & Shohamy, D. (2018). Value-based decisions involve sequential sampling from memory. *BioRxiv*, 269290. <https://doi.org/10.1101/269290>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 43(8), 1520–1531. <https://doi.org/10.1037/xhp0000404>
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>
- Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences*, 23(3), 251–263. <https://doi.org/10.1016/j.tics.2018.12.003>
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3), 432–459. <https://doi.org/10.1037/0033-295x.100.3.432>
- Chen, Z., Holland, R. W., Quandt, J., Dijksterhuis, A., & Veling, H. (2019). When mere action versus inaction leads to robust preference change. *Journal of Personality and Social Psychology*, 117(4), 721–740. <https://doi.org/10.1037/pspa0000158>
- Chen, Z., Veling, H., Dijksterhuis, A., & Holland, R. W. (2016). How does not responding to appetitive stimuli cause devaluation: Evaluative conditioning or response inhibition? *Journal of Experimental Psychology: General*, 145(12), 1687–1701. <https://doi.org/10.1037/xge0000236>
- Clarkson, J. J., Tormala, Z. L., Rucker, D. D., & Dugan, R. G. (2013). The malleable influence of social consensus on attitude certainty. *Journal of Experimental Social Psychology*, 49(6), 1019–1022. <https://doi.org/10.1016/j.jesp.2013.07.001>
- De, M. B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, 16(1), 4–6. <https://doi.org/10.1038/nn.3279>
- Dutilh, G., & Rieskamp, J. (2016). Comparing perceptual and preferential decision making. *Psychonomic Bulletin & Review*, 23(3), 723–737. <https://doi.org/10.3758/s13423-015-0941-1>
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1, Article 0002. <https://doi.org/10.1038/s41562-016-0002>
- Frank, S. A. (2009). The common patterns of nature. *Journal of Evolutionary Biology*, 22(8), 1563–1585. <https://doi.org/10.1111/j.1420-9101.2009.01775.x>
- Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535–574. <https://doi.org/10.1146/annurev.neuro.29.051605.113038>
- Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, 9(1), 1–14.
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3), 411–435. [https://doi.org/10.1016/0010-0285\(92\)90013-R](https://doi.org/10.1016/0010-0285(92)90013-R)
- Holland, R. W., Verplanken, B., & van Knippenberg, A. (2003). From repetition to conviction: Attitude accessibility as a determinant of attitude certainty. *Journal of Experimental Social Psychology*, 39(6), 594–601. [https://doi.org/10.1016/S0022-1031\(03\)00038-6](https://doi.org/10.1016/S0022-1031(03)00038-6)
- Izuma, K., & Murayama, K. (2013). Choice-induced preference change in the free-choice paradigm: A critical methodological review. *Frontiers in Psychology*, 4, 41. <https://doi.org/10.3389/fpsyg.2013.00041>
- Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3), 461–474. <https://doi.org/10.1037/0278-7393.33.3.461>

- Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 367(1594), 1322–1337. <https://doi.org/10.1098/rstb.2012.0037>
- 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>
- Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, 29, 6–11. <https://doi.org/10.1016/j.copsyc.2018.10.008>
- 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>
- Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 108(33), 13852–13857. <https://doi.org/10.1073/pnas.1101328108>
- Kunar, M. A., Watson, D. G., Tsetsos, K., & Chater, N. (2017). The influence of attention on value integration. *Attention, Perception & Psychophysics*, 79(6), 1615–1627. <https://doi.org/10.3758/s13414-017-1340-7>
- 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>
- Levy, H., & Markowitz, H. M. (1979). Approximating expected utility by a function of mean and variance. *The American Economic Review*, 69(3), 308–317. <https://doi.org/10.2307/1807366>
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764–766. <https://doi.org/10.1016/j.jesp.2013.03.013>
- Lichtenstein, S., & Slovic, P. (2006). *The construction of preference*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511618031>
- Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *The Journal of Neuroscience*, 31(37), 13214–13223. <https://doi.org/10.1523/JNEUROSCI.1246-11.2011>
- Loeb, G. E., & Fishel, J. A. (2014). Bayesian action & perception: Representing the world in the brain. *Frontiers in Neuroscience*, 8, 341. <https://doi.org/10.3389/fnins.2014.00341>
- Mathôt, S., Siebold, A., Donk, M., & Vitu, F. (2015). Large pupils predict goal-driven eye movements. *Journal of Experimental Psychology: General*, 144(3), 513–521. <https://doi.org/10.1037/a0039168>
- Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, 88(1), 78–92. <https://doi.org/10.1016/j.neuron.2015.09.039>
- Parr, T., Benrimoh, D. A., Vincent, P., & Friston, K. J. (2018). Precision and false perceptual inference. *Frontiers in Integrative Neuroscience*, 12, 39. <https://doi.org/10.3389/fnint.2018.00039>
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1–2), 8–13. <https://doi.org/10.1016/j.jneumeth.2006.11.017>
- Petrocilli, 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., Moisa, M., Opitz, A., Grueschow, M., & Ruff, C. C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nature Communications*, 6, 8090. <https://doi.org/10.1038/ncomms9090>
- 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., Holland, R. W., Chen, Z., & Veling, H. (2019). The role of attention in explaining the no-go devaluation effect: Effects on appetitive food items. *Journal of Experimental Psychology: Human Perception and Performance*, 45(8), 1119–1133. <https://doi.org/10.1037/xhp0000659>
- R Core Team. (2019). *R: A language and environment for statistical computing*. <https://www.R-project.org/>
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9(5), 347–356. <https://doi.org/10.1111/1467-9280.00067>
- Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 127–140. <https://doi.org/10.1037/0096-1523.26.1.127>
- Ratcliff, R., Voskuilen, C., & McKoon, G. (2018). Internal and external sources of variability in perceptual decision-making. *Psychological Review*, 125(1), 33–46. <https://doi.org/10.1037/rev0000080>
- Rolls, E. T., Grabenhorst, F., & Deco, G. (2010). Choice, difficulty, and confidence in the brain. *NeuroImage*, 53(2), 694–706. <https://doi.org/10.1016/j.neuroimage.2010.06.073>
- Schad, D. J., Betancourt, M., & Vasishth, S. (2019). Toward a principled Bayesian workflow in cognitive science. *ArXiv*. <https://arxiv.org/abs/1904.12765>
- Schonberg, T., Bakkour, A., Hover, A. M., Mumford, J. A., Nagar, L., Perez, J., & Poldrack, R. A. (2014). Changing value through cued approach: An automatic mechanism of behavior change. *Nature Neuroscience*, 17(4), 625–630. <https://doi.org/10.1038/nn.3673>
- 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*.
- Shinners, P. (2011). *PyGame*. <http://pygame.org/>
- Stan Development Team. (2016). *{RStan}: The {R} interface to {Stan}*. <http://mc-stan.org/>
- Tormala, Z. L., & Rucker, D. D. (2018). Attitude certainty: Antecedents, consequences, and new directions. *Counselling Psychology Review*, 1(1), 72–89. <https://doi.org/10.1002/arc.1004>
- Tsetsos, K., Chater, N., & Usher, M. (2012). Saliency driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24), 9659–9664. <https://doi.org/10.1073/pnas.1119569109>
- van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, 5, e12192. <https://doi.org/10.7554/eLife.12192>
- Vanunu, Y., Pachur, T., & Usher, M. (2019). Constructing preference from sequential samples: The impact of evaluation format on risk attitudes. *Decision*, 6(3), 223–236. <https://doi.org/10.1037/dec0000098>
- Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. <https://CRAN.R-project.org/package=loo>
- Veling, H., Chen, Z., Tombrock, M. C., Verpaalen, I. A. M., Schmitz, L. I., Dijksterhuis, A., & Holland, R. W. (2017). Training impulsive choices for healthy and sustainable food. *Journal of Experimental Psychology: Applied*, 23(2), 204–215. <https://doi.org/10.1037/xap0000112>
- Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, 50(2), 179–197. [https://doi.org/10.1016/0001-6918\(82\)90006-3](https://doi.org/10.1016/0001-6918(82)90006-3)

- Visser, P. S., & Mirabile, R. R. (2004). Attitudes in the social context: The impact of social network composition on individual-level attitude strength. *Journal of Personality and Social Psychology, 87*(6), 779–795. <https://doi.org/10.1037/0022-3514.87.6.779>
- Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science, 18*(6), 516–523. <https://doi.org/10.1111/j.1467-9280.2007.01932.x>
- 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>
- Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife, 5*, e17688. <https://doi.org/10.7554/eLife.17688>

Received October 15, 2020

Revision received March 30, 2021

Accepted May 14, 2021 ■