https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–21
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Sample Range as Subjective Variability and Implications for Experience-Based Decision Making

Tianwei Liu,[a,*] **Benjamin Scheibehenne**[a]

[a] Department of Business and Economics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

*Corresponding author

tianweiliu.psi@gmail.com (T. Liu), scheibehenne@kit.edu (B. Scheibehenne)

https://orcid.org/0000-0001-6734-1297 (T. Liu), https://orcid.org/0000-0001-6097-9685 (B. Scheibehenne)

Received: February 25, 2024
Revised: April 8, 2025; September 21, 2025; March 20, 2026
Accepted: July 7, 2026
Published Online in Articles in Advance: November 19, 2026

https://doi.org/10.1287/mnsc.2024.05050

Copyright: © 2026 INFORMS

Abstract. Variance and standard deviation are common metrics of statistical variability in the management and decision-making literature. In everyday contexts, however, information about sample range prevails. Range is arguably easier to comprehend and compute than variance, yet it often yields equivalent judgments about variability. Hence, we reason that decision makers often rely on range to judge variability. We test this hypothesis across four behavioral experiments (total $n = 1{,}093$). When comparing the variability of two number sets, decision makers predominantly based their judgments on sample range rather than sample variance when the two metrics disagree (experiment 1). Sample range tends to increase with sample size, whereas sample variance does not. Consistent with the range as variability hypothesis, decision makers perceived distributions as more variable after observing larger samples (experiment 2). This perceived variability, in turn, influenced risky decisions, as larger sample sizes led to decreased valuations of uncertain rewards for risk-averse decision makers (experiment 3). In further support of the hypothesis, when asked to mentally simulate new samples from a previously experienced distribution, participants generated more numbers within the observed range and fewer numbers outside of it than the true distribution implies (experiment 4). To model the data, we propose a "cognitive fence" idea as a potential psychological explanation for how experience shapes subjective variability. We conclude by discussing implications for research on overconfidence, distributional learning, risk perception, and communication.

History: Accepted by Jack Soll, behavioral economics and decision analysis.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2024.05050.

Keywords: subjective variability • range • risk perception • sampling

Introduction

Understanding the variability or dispersion of outcomes is critical for successful decision making. Indeed, the whole body of literature termed "decision making under risk" explores how individuals make choices when outcomes are variable. Among the various measures of statistical variability (e.g., variance, range, mean absolute deviation), variance assumes a prominent position, as the concepts of risk and variability are often operationalized as variance or variance-related measures (e.g., coefficient of variation). This holds true for both prescriptive models (Markowitz 1952, Black and Scholes 1973, Kacelnik and Bateson 1996) and descriptive models (Kahneman and Tversky 1979, Weber et al. 2004).

The prominence of variance is particularly manifest in the field of economics and finance, where risk, volatility, and standard deviation are virtually interchangeable concepts (e.g., Jorion 1995, Blitz et al. 2013, but see Cox 2008). The near consensus on using variance as a measure of variability may stem from Harry Markowitz's impactful work on modern portfolio theory. Interestingly, however, Markowitz (1952, p. 89) himself regarded variance as merely a convenient proxy for risk:

> The concepts "yield" and "risk" appear frequently in financial writings. Usually if the term "yield" were replaced by "expected yield" or "expected return," and "risk" by "variance of return," little change of apparent meaning would result.
>
> Variance is a well-known measure of dispersion about the expected.

Compared with other common measures of data dispersion, such as (interquartile) range or mean absolute deviation, variance has desirable properties. It is defined as an expected value (EV) $\mathrm{E}[(X-\mu)^2]$ and therefore does not depend on the availability of

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

empirical data. In addition, for many common distributions, variance can be easily derived from parameters of the density/mass functions, making it mathematically tractable for theorists. However, when it comes to the subjective perception of variability by human decision makers, the universal use of variance demands justification. People in everyday situations are unlikely to calculate variance when trying to estimate the level of uncertainty in a quantity (Lovie 1978). As a result, approximating subjective perception of variability with variance or any variance-dependent statistics relies on an "as-if" rationale: decision makers might not explicitly compute variance, yet they behave *as if* they do (Peterson and Beach 1967). Thus, relying on variance in this context is tenable only if it predicts human behavior at least as accurately as other measures of dispersion. In the paper at hand, we shall show that this may not necessarily be the case.

Subjective Variability as Range

In contrast to its ubiquity in the academic world, variance is not often seen or used in everyday life. In fact, we had trouble conceiving a single case where variance or standard deviation (SD) is communicated to laypeople to describe variability. Instead, simpler measures of variability are used. Among the alternatives, sample range or, more generally, information about the minimum and maximum, appears to be quite common. For example, in weather forecasts, daily highs and lows are often displayed as central information. In comparison, the hour-to-hour fluctuations are less commonly reported. Likewise, some graduate programs, particularly MBA programs, conventionally publish "class profiles" as a reference for prospective students. These class profiles typically contain information about standard test scores of last year's student body. The range of the scores is often displayed alongside the mean and/or median, and whereas some programs report percentiles, we did not find any academic program that reports the standard deviation of the test scores. Even in finance, where variance/volatility is supposed to rule, range appears to be a more prevalent indicator of risk, at least for retail investors. On the stock quotes interfaces of financial data services and brokerage apps, the daily and 52-week ranges typically occupy salient positions, whereas information about volatility (e.g., beta) is either moved to secondary menus, buried at the bottom of the page, or completely absent.

Presumably, the way we design our choice environments reflects the relative importance we assign to different pieces of information (Simon 1990). Consistent with this perspective, in daily life, the minimum and maximum of a quantity are often more relevant to decision making than variance. For example, decisions such as what to wear today are typically driven by how hot/cold it can get, instead of how the temperature will fluctuate within the day. Similarly, for evaluating weather warnings and climate risks, we often focus on the extremes, such as the highest expected wind speed or flood level. For a prospective student hunting for suitable programs or a job candidate hunting for suitable jobs, the "minimum/maximum requirements" on GPA, test scores, or work experience arguably are among the most important information to consider. In the context of stock trading, whether a limit buy (sell) order can be filled within a specific period depends only on whether the lowest (highest) trading price of the asset will reach the limit price during that timeframe.

Another reason why range is more often used in everyday life than variance is because it is easier to grasp and computationally less demanding (Lovie and Lovie 1976). Conceptually, the "difference between the largest value and the smallest value" is more straightforward than the "expected value of the squared deviation from the mean." Practically, computing $\max(\boldsymbol{X}) - \min(\boldsymbol{X})$ is simpler than computing $\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2/(N-1)}$. Indeed, research on judgments of variability has shown that even when provided with accessible and comprehensive definitions of sample variance, university students struggle to estimate either the absolute (Lovie and Lovie 1976) or relative (Beach and Scopp 1968) magnitude of variance.

Whereas range is a simpler concept than variance, both are valid measures of statistical dispersion and hence are highly correlated under many conditions. Take the example of n observations from a continuous uniform distribution: $U(a, b)$. The expected value of sample SD is $(b-a)/\sqrt{12}$, whereas the expected value of sample range is $[(n-1)/(n+1)](b-a)$ (see Online Appendix A for a derivation). Therefore, for any given n, E[SD] and E[range] form a perfect linear relationship: $\mathrm{E}[range] = [\sqrt{12}(n-1)/(n+1)]\mathrm{E}[SD]$. Similar relationships also exist for other common continuous distributions such as the normal distribution and the exponential distribution (Ramirez and Cox 2012). In fact, because of the close relationship between range and SD in normal distributions, simple rules of thumb that estimate SD from sample range have been widely described in statistics textbooks (Johnson 2017, Triola 2004). Lovie (1978) explicitly tested the idea of using such rules of thumb to aid SD estimation.

Considering that range is more accessible in the environment and easier to compute than variance and given that range and variance often convey similar information about relative sample variability, it seems natural that humans, adapted to their environments, would often rely on range when judging the variability of continuous quantities. Indeed, when making numerical judgments, extreme values receive

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

disproportionate weighting (Spitzer et al. 2017). Relatedly, people overweight extreme outcomes in memory, which then affects their risk preferences (Lieder et al. 2018, Mason et al. 2022). Moreover, when asked whether a number is an element of a previously memorized set, people were more confident in their responses on distractors that lay just outside the range than those that lay just inside the range (Hamilos and Pitz 1977), indicating that the minimum and the maximum are salient features of a sample.

Relatedly, Konovalova and Pachur (2021) investigated factors that influence people's intuitive understanding and judgment of variance. Their findings suggest that whereas people's subjective perceptions of variance are influenced by the statistical sample variance, some other metrics such as range also have significant impacts. Based on our discussions above, however, it is doubtful that people in their everyday decisions would often seek to understand or estimate variance at all when the goal is to evaluate variability in their environments. Therefore, we propose that in many everyday tasks, range serves as the default metric that decision makers rely on to assess variability. Specifically, we hypothesize that range commonly predicts subjective perceptions of variability better than variance (hypothesis 1). This hypothesis leads to testable predictions when people learn about an outcome distribution by observing samples from it, such as in the decisions from experience paradigm (DfE) (Hertwig and Erev 2009). In this context, range tends to increase with sample size because the chance of observing extreme outcomes increases with sample size. As shown in Online Appendix A, the expectation of range is a concave, increasing function of sample size. In contrast, the expectation of the unbiased estimator for variance is, by definition, the variance of the underlying distribution, which does not vary with sample size. If decision makers judge variability based on sample variance, there should be no systematic relationship between perceived variability and sample size at a given variance. If they rely on sample range instead, they will perceive the same distribution as more variable after sampling more from it (hypothesis 2). Further along this line, perceived variability should also influence risky decisions. Because larger sample sizes lead to higher perceived variability, for risk-averse decision makers, valuation of a positive outcome should decrease as sample size increases (hypothesis 3).

Finally, when decision makers rely on range, we hypothesize that they will be less likely to imagine values that lie outside of what they have encountered in their previous experiences, compared with normative predictions of probability theory (hypothesis 4).

To test these hypotheses, we report four preregistered experiments. In experiment 1, when asked to compare the variability of two sets of numbers for which range and SD lead to opposite conclusions, participants' choices were predominantly driven by sample range rather than sample SD (hypothesis 1). In experiment 2, participants perceived a distribution as more variable after observing a larger sample from it (hypothesis 2). Moreover, the increase in perceived variability with sample size became more pronounced when the sample range was made more salient. This observation rules out the notable alternative hypothesis that people may have relied on uncorrected sample SD (i.e., divided by n instead of $n-1$), which does increase with sample size (Kareev et al. 2002). In experiment 3, as sample size increased, the valuation of a risky positive outcome decreased (hypothesis 3). In experiment 4, participants observed some samples from a distribution and were then asked to mentally generate new cases from the same distribution. The mean proportion of generated numbers that fell outside the experienced range was smaller than prescribed by probability theory. As a result, the mean variability of the generated samples was smaller than the objective variability of the underlying distribution and more so when observed sample size was smaller (hypothesis 4). We discuss these results in the context of related literature, we elaborate on possible cognitive explanations of the effect, and we highlight practical implications.

Transparent Reporting

Experiments 1, 2, and 4 were programmed and administered with Qualtrics.[1] Experiment 3 was programmed with laboratory.js[2] and administered through a JATOS server.[3] All experiments were run online on Prolific Academic.[4] We preregistered all experiments and determined all sample sizes before collecting the data. We report all manipulations, all measures, and all data exclusions.[5]

Experiment 1: Range vs. Variance in Variability Judgments

Experiment 1 tests the hypothesis that in a forced-choice task, decision makers rely on range rather than variance to judge variability. As discussed above, range and variance are often highly correlated, which makes it difficult to determine which measure people rely on. The design of experiment 1 aims to disentangle the two factors.

Method

Design and Procedure. We repeatedly presented participants two number sets, each consisting of 10 two-digit numbers next to each other on a computer screen. In each of the 30 trials, participants were instructed to judge which of the two sets was "more dispersed" or "more variable." They were informed that "the averages (arithmetic mean) of the two sets are about the same." In half of the trials, variance and

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

range led to the same choice; that is, the set with higher variance also had a wider range (consistent trials). In the other half of the trials, variance and range led to opposite choices: that is, the set with higher variance had a narrower range (inconsistent trials). In both conditions, the number sets were pseudorandom samples generated from uniform distributions that met the following criteria: For the inconsistent conditions, $SD_{set1} - SD_{set2} > 1$, and $Range_{set2} - Range_{set1} \geq 2$. As a result, set 1 in the inconsistent condition had higher variance but a narrower range than set 2. The stimuli in the consistent condition were created as mirror images of the stimuli in the inconsistent condition: for every inconsistent trial, there was a paired consistent trial whose two sets of numbers had roughly the same SDs, but the difference in range was reversed (see Table 1 for an example). We constructed 40 such pairs from which each participant saw a random subset of 15 pairs (15 consistent trials and 15 inconsistent trials, displayed in random order). The order in which the two sets were displayed on the screen was randomized. Table 1 provides examples of the stimuli.

If decision makers rely on range rather than variance to evaluate variability, in an inconsistent trial, they would judge the option with the wider range but smaller variance as more variable. To control for possible influences of wording or linguistic connotations, the exact term displayed to participants in instructions and questions to describe statistical variability ("dispersed" versus "variable") randomly varied between subjects.

We further hypothesized that people are more likely to rely on range to judge variability when sample range is made more salient. To test this, we randomly assigned participants to either a sorted or an unsorted condition. In the sorted condition, numbers presented on the screen were sorted from lowest to highest; in the unsorted condition, they were ordered randomly. Because the sample range is more salient (i.e., easier to eyeball) in the sorted condition than in the unsorted condition, we predicted that participants in the sorted condition are more likely to choose the set with the wider range than participants in the unsorted condition.

Participants. A total of 400 participants passed the comprehension check and the catch trial and finished the task (261 female, mean age = 42.4 years). Participants were compensated with a flat fee of £0.7. It took participants an average of six minutes to complete the task.

Results

We calculated the proportion choosing the wider-range option for each participant in each condition. The choice of terminology for variability ("variable" versus "dispersed") did not impact choices, as indicated by a Bayesian independent-sample *t*-test (using the default Jeffreys–Zellner–Siow (JZS) prior, $r = 0.707$ (Rouder et al. 2009)). The resulting Bayes factor (BF_{10}) that quantifies the evidence for the alternative hypothesis (predicting that the terminology has an influence) over the null hypothesis (predicting no such influence) was 0.86. As this value is below one, the data are more likely under the null hypothesis than the alternative hypothesis. Regression analyses further showed that the wording did not moderate the effects of other variables either. Hence, we excluded this factor from the following analyses.

Figure 1 displays the proportion of trials in which participants selected the wider-range option separately for trial type (consistent versus inconsistent) and condition (unsorted versus sorted). The figure shows that the mean proportions in each of the four conditions are reliably higher than 50%. These observations were confirmed with Bayesian one-sample *t*-tests against 0.5 (all $BF_{10}s > 10^9$). These results indicate that range serves as a strong cue for determining relative variability.

In the consistent trials, the mean proportion is 73.9%. This condition serves as a baseline and indicates that participants were generally sensitive to sample variability in the current setting. An error-free decision maker should choose the wider-range option, which is also the larger-variance option in this condition, 100% of the time. The considerable heterogeneity observed in Figure 1 for the responses in this condition shows that the task was somewhat challenging.[6]

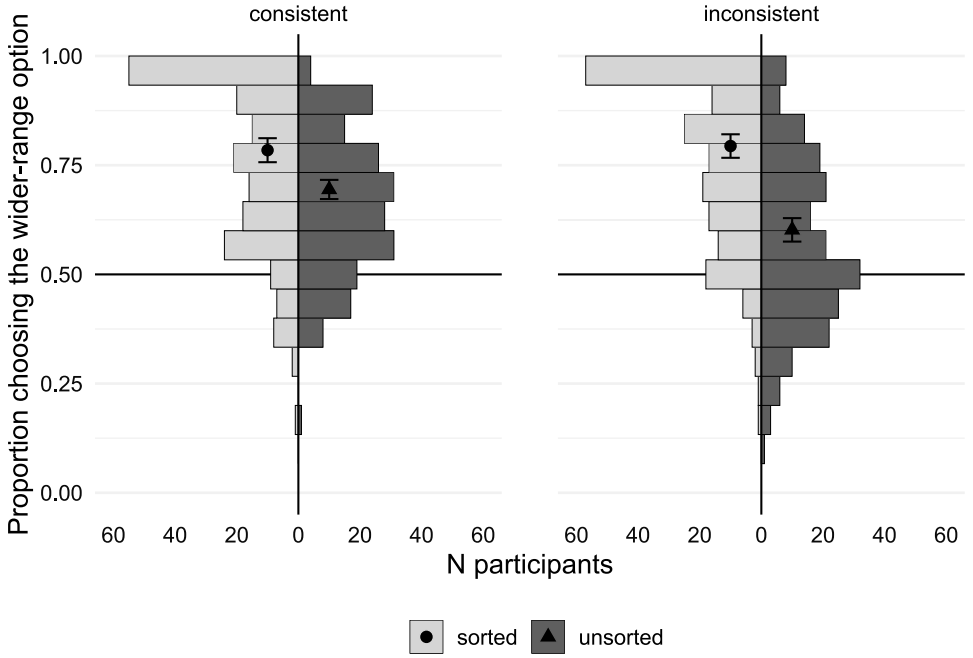
More importantly, in 69.6% of the inconsistent trials, participants perceived the option with a wider range but smaller variance as more variable, supporting hypothesis 1 that decision makers rely on range more than variance when judging variability. Figure 1 shows that there is significant heterogeneity in this condition as well, indicating that the majority of participants did not rely on one cue for 100% of the inconsistent trials.

Table 1. Example of a Mirrored Pair of Stimuli (Unsorted Condition) in Experiment 1

Trial type		Set	SD	Range	SD2 − SD1	Range2 − range1
Inconsistent	1	40 33 25 29 49 30 41 35 35 34	6.9	24	1.6	−3
	2	47 26 26 27 39 42 26 46 37 31	8.5	21		
Consistent	1	37 43 36 32 30 45 26 41 24 33	7	21	1.4	3
	2	31 35 36 48 28 24 47 30 29 44	8.4	24		

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Figure 1. Distribution of Participants' Proportions of Choice for the Wider-Range Option for Each Condition in Experiment 1



Note. The points with error bars indicate the mean choice proportions and their 95% confidence intervals.

As predicted, the proportion of choices of the wider-range option increased when the displayed numbers were sorted. This was true for the consistent trials (sorted versus unsorted: 78.4% versus 69.4%) and even more so for the inconsistent trials (sorted versus unsorted: 79.4% versus 60.2%). Bayesian independent *t*-tests confirmed these differences. The BF_{10} values were 1.9×10^4 and 6.6×10^{17}, respectively, indicating very strong evidence for the alternative hypothesis that participants were more likely to choose the wider-range option when the displayed numbers were sorted. Relatedly, it can be seen from Figure 1 that more participants rely on range across all trials in the sorted condition than in the unsorted condition.

The comparison between consistent and inconsistent trials further indicates that participants were not totally blind to variance: when the wider-range option also had higher variance, it was chosen 73.9% of the time. In contrast, when the wider-range option had lower variance, the proportion that it was chosen dropped to 69.6% (consistent versus inconsistent: BF_{10} = 3,833 with the Bayesian paired-sample *t*-test). However, this impact of variance was only observed when numbers were unsorted (consistent versus inconsistent: 69.4% versus 60.2%, $BF_{10} = 2.2 \times 10^8$). When the range cue was made more salient through sorting the numbers, variance was practically ignored (consistent versus inconsistent: 78.4% versus 79.4%, BF_{10} = 0.12). These differences suggest that participants did respond to relative differences in variance to some extent.

Discussion

Taken together, the results of experiment 1 support the hypothesis that participants' perception of relative variability depended on range more than variance. Alternative explanations are possible, though. Although the term we used to describe statistical variability ("dispersed" versus "variable") had no effect, one may still suspect that both terms may have hinted participants to focus on range rather than variance. To rule out this alternative explanation, we conducted a follow-up experiment on Prolific (n = 181, 96 females, mean age = 44.7 years) using identical stimuli and design as experiment 1, but instead of asking participants to compare levels of variability or dispersion, we explicitly asked: "Which of the two sets has larger variance?" To make this question meaningful to lay participants, the instructions included a brief explanation of the concept of variance. Because variance is precisely defined, the task had an objectively correct answer that allowed us to incentivize participants with £0.01 per correct response. To prevent participants from using software to calculate the sample variance, response in each trial was limited to 13 seconds. If no response was made within this time, the task automatically proceeded to the next trial, and no response was recorded for that trial.

Result patterns mirrored those of experiment 1. In particular, in 63.7% of the inconsistent trials, participants chose the option with a wider range but smaller variance as having larger variance, and this effect was

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

stronger for sorted than unsorted number sets (72.8% versus 54.5%, $BF_{10} = 6.8 \times 10^5$). Hence, the previous results were not driven by semantic connotations of the terms used to describe variability.

Comparing the two experiments, one difference emerges: for the inconsistent trials, the proportion of choices favoring the wider-range option was lower in the follow-up "variance" version than in the original "variability/dispersion" version (63.7% versus 69.6%, $BF_{10} = 8.13$). By contrast, no difference arose for the consistent trials (follow-up versus original: 71.8% versus 73.9%, $BF_{10} = 0.22$). These results further suggest that people can process variance to some extent: when explicitly asked to compare variances, they were more likely to rely on variance over range than when asked to compare "variability" or "dispersion." Nevertheless, these variance-driven choices remained below 50% overall.

In experiment 1, we attempted to disentangle range and variance by deliberately designing inconsistent trials with markedly different distribution shapes. As illustrated in Figure 2, the option with a narrower range but higher variance followed a bimodal distribution, whereas the option with a wider range but lower variance had longer, thinner tails. These contrasting shapes introduced a potential confound. Furthermore, because variance and range are intrinsically correlated, even such shape manipulations could hardly construct two samples where one option has substantially higher variance but a narrower range than the other option. As a result, the differences in variance and range were relatively small. In addition, whereas experiment 1 examines people's estimates of variability within an observed sample, a more theoretically and practically important question is how people infer the underlying distribution's variability from that sample. To address these limitations and reinforce our primary hypothesis, we conducted a second experiment that relied on a different design, outlined next.

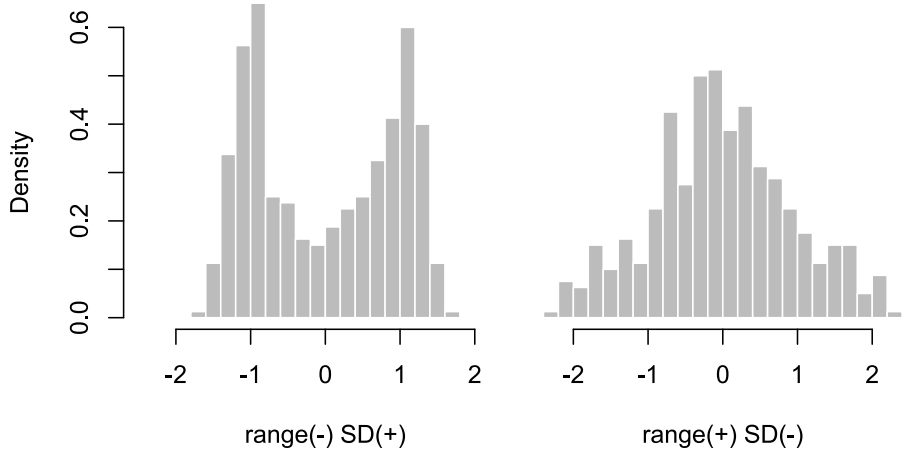
Experiment 2: Perceived Variability Increases with Sample Size

Experiment 2 tests the hypothesis that decision makers perceive a distribution as more variable after seeing a larger sample from it (hypothesis 2), which is consistent with range as variability but inconsistent with variance as variability. In addition, we hypothesized that the effect of sample size on perceived variability will be larger when sample range is made more salient (i.e., sorted versus unsorted, as in experiment 1). Recent research by Olschewski and Scheibehenne (2024, study 2) reported empirical evidence for an influence of sample size on perceived variability. Participants in their experiment repeatedly observed sequences of numbers from different distributions. As a measure of perceived variance, participants had to select one out of seven histograms with different SDs that best represented the distribution they had just experienced. The current experiment adopted a similar task but added the saliency of range as an additional factor, like in experiment 1.

Method

Design and Procedure. As depicted in Figure 3, in each trial, participants saw a set of numbers along with six different histograms presented on the computer screen. Each histogram visualized a normal distribution with an identical mean of 200 but varying SDs of 10, 15, 20, 30, 40, and 50. Participants answered

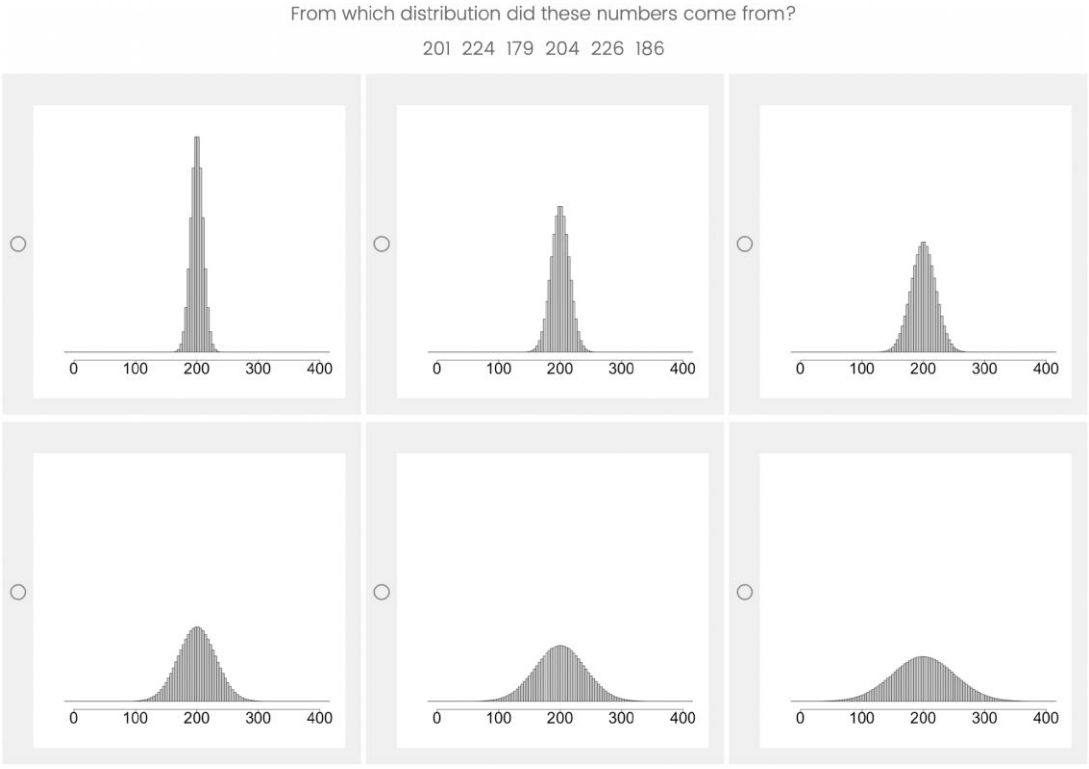
Figure 2. Aggregate Distributions of Stimuli Used in the Inconsistent Condition of Experiment 1



Notes. Distributions are *z*-scored separately for each set. Left: Options with narrower range and larger SD. Right: Options with wider range and smaller SD.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Figure 3. Screenshot of a Single Trial of Experiment 2



Note. In this trial, sample size = 6, SD = 20, and saliency of range = low.

the question "From which distribution did these numbers come from?" by selecting one of the histograms. Like in experiment 1 above, but unlike the experiment by Olschewski and Scheibehenne (2024), the numbers were displayed simultaneously as a list. This design was necessary for the sorted condition because a perfectly increasing sequence could misleadingly hint at a trend. The set of numbers presented in each trial was generated from normal distributions with an expected value of 200. The sample size (n) and SD of the numbers varied systematically across trials, with three levels of n (3, 6, or 10) and two levels of SD (20 or 40), resulting in six unique combinations of parameters. Each parameter combination was repeated five times throughout the task, resulting in 30 trials per participant in total. For each combination, we generated 12 sets of stimuli where the sample SD and mean closely tracked the designated EV and SD. Each participant was presented with a random subset of 5 out of these 12 stimuli for each combination. The first six trials, one for each $n \times$ SD combination, were designated as practice trials and were excluded from the analysis. The task was incentivized. Participants received £0.03 for each selection that matched the actual SD on the trial.

As in experiment 1, to manipulate the saliency of range, we randomly assigned participants to either a sorted or an unsorted condition. In the sorted condition, numbers were presented from lowest to highest on the screen; in the unsorted condition, they were presented in random order. If the potential effect of n on perceived variability was driven by the effect of n on range, then the effect of n on perceived variability should be larger in the sorted condition than in the unsorted condition. More precisely, we expect an interaction between n and *Saliency* of range such that the effect of n will be stronger in the sorted condition than in the unsorted condition.

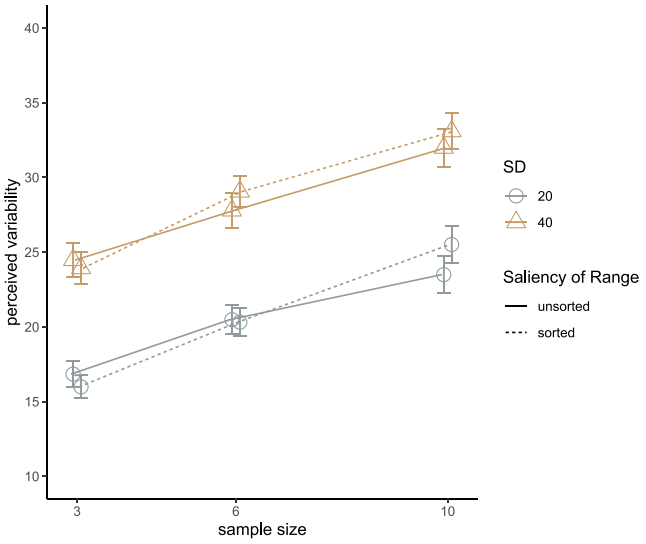
Participants. A total of 217 Prolific participants passed the comprehension check and finished the task (116 female, mean age = 41.7 years), of which 205 passed the preregistered exclusion criteria. Participants were compensated with a flat fee of £0.7, along with a performance-contingent bonus as described above. On average, it took participants about nine minutes to complete the task.

Results

Responses are coded as the standard deviation of the histogram participants chose in each trial.[7] Figure 4 provides a descriptive summary of participants' responses. The figure shows that perceived variability increases with SD, confirming the validity of the task,

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Figure 4. (Color online) Perceived Variability as a Function of SD, *n*, and Salience of Range in Experiment 2



Note. Error bars = 95% CI.

as participants were able to discern differences in variability. It further shows that perceived variability increases with *n*, which supports our hypothesis 2. To corroborate these descriptive findings, we fitted a Bayesian linear mixed-effects model to the data. For this and subsequent linear models, we used the brms package in R (Bürkner 2017), which fits Bayesian regression models via Markov chain Monte Carlo (MCMC) methods using the probabilistic programming language Stan (Stan Development Team 2023). Bayesian regression models are comparable to their frequentist counterparts but offer several advantages, including the avoidance of null-hypothesis significance testing pitfalls (Andraszewicz et al. 2015) and the ability to incorporate regularizing priors, which help stabilize parameter estimates and constrain them to theoretically plausible ranges (McElreath 2018). We employed the default weakly informative priors provided by brms, which are Student's *t*-distributions with three degrees of freedom, centered at zero. Specifically, regression coefficients were assigned a Student's $t(3, 0, 2.5)$ prior, which assumes that most plausible values lie within a reasonable range (e.g., 95% within approximately −10 to 10) while still allowing for larger estimates if strongly supported by the data. These priors are not very informative but, rather, serve to regularize the model and reduce overfitting. Below, we report the means of the posterior distributions along with 95% highest density intervals (HDIs), which represent the range within which 95% of the most credible values of the parameter lie, given the data and the model.

In this statistical model, participants' response in each trial (i.e., the standard deviation of the selected histogram) is the dependent variable, and SD, *n*, and *Saliency* are predictors (SD and *Saliency* are both coded as −0.5 and +0.5 for their respective two levels). We included the *Saliency* × *n* interaction to test the hypothesis that saliency of sample range moderates the effect of *n*. Finally, the sample range increases more sharply with *n* when SD is larger.[8] Therefore, we included the SD × *n* interaction in the model. The model also included a random intercept and separate random slopes for all the above factors and interactions by participant (Barr et al. 2013).

The results show that perceived variability increases with SD ($M = 7.80$, 95% HDI (6.47, 9.09)), and, more importantly with *n* ($M = 0.99$, 95% HDI (0.76, 1.25)). The latter effect is consistent with variability as range but inconsistent with variability as SD because range generally increases with *n*, but SD does not. Sorted sets tend to be perceived as less variable than unsorted sets, although this effect is not reliable ($M = -1.66$, 95% HDI (−3.75, 0.30)). This is consistent with past findings that mean absolute rate of change between successive elements in a sequence is positively related to perceived variability (Lathrop 1967). As predicted, saliency of range moderated the effect of *n* on perceived variability (*n* × *Saliency* interaction: $M = 0.32$, 95% HDI (0.00, 0.65)) such that the increase of perceived variability with *n* became steeper when sample range was more salient. Last, SD did not moderate the effect of *n* on perceived variability ($M = 0.03$, 95% HDI (−0.14, 0.19)). It is possible that our dependent measure was not sensitive enough to detect the different slopes of *n* on range at different SDs.

The estimated standard deviation of the slopes for *n* was 0.98, indicating that whereas most participants perceived larger samples as more variable, the strength of this effect varied considerably among them. Substantial participant-level heterogeneity also emerged in effects for *Saliency* × *n* (SD = 0.32).

One central assumption in the above analyses is that perceived variability increases with sample size because sample range tends to increase with sample size. However, an important alternative explanation for this sample-size effect comes from Kareev et al. (2002), who showed that people tend to underestimate variability, and such underestimation is more pronounced at smaller sample sizes. From a relative perspective, this means that an increase in sample size leads to higher estimates of variability. Kareev et al. (2002) interpreted their results by suggesting that people rely on the sample variance (i.e., the sum of squared deviations divided by *n*) without applying the usual correction for *n*−1.

To assess the relative explanatory power of the range account versus the uncorrected variance account, we fitted three separate Bayesian mixed models that each used only one predictor—sample range (*r*), corrected sample SD (*cs*), or uncorrected sample SD (*us*)—and

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

then compared their fit using leave-one-out cross validation (LOO) (Vehtari et al. 2017) via the brms R package (Bürkner 2017).[9] Each predictor reliably explained perceived variability (*r*: $M = 0.17$, 95% HDI (0.16, 0.18); *cs*: $M = 0.39$, 95% HDI (0.37, 0.42); *us*: $M = 0.49$, 95% HDI (0.46, 0.52)). Differences in predictive accuracy, as indexed by expected log predictive density (ELPD-LOO), indicated that the *r* model was the best, followed by the *us* model ($\Delta\text{ELPD} = -183.91 \pm 16.92$, $p < 0.001$) and the *cs* model ($\Delta\text{ELPD} = -306.26 \pm 22.63$, $p < 0.001$). Moreover, LOO-based model weights, which approximate the relative plausibility of each model, showed that the *r* model captured essentially all of the model weight, whereas the other two models had effectively zero weight.

Because the 95% HDI for the $n \times$ *Saliency* interaction included zero, we verified the reliability of this effect with a similar comparison between the models with and without the interaction term. The difference in predictive accuracy, as indexed by ELPD-LOO, suggests that the model including the $n \times$ *Saliency* term is superior to the model without the interaction ($\Delta\text{ELPD} = -4.42 \pm 3.75$). Also, LOO-based model weights show that the model with interaction captures 98.8% of the weight, whereas the model without interaction only captures 1.2%. Hence, the $n \times$ *Saliency* interaction meaningfully improves the explanatory power of the model.

For completeness, we also estimated a Bayesian linear mixed-effects model with both the sample SD and the sample range as predictors for participants' responses. This regression confirms that sample range predicts participants' choices above and beyond sample SD. Likewise, the model far outperforms a simpler regression with only the sample SD as a single predictor ($\Delta\text{ELPD} = -595.6 + -39.5$).

Discussion

In summary, experiment 2 confirmed the findings in experiment 1 that variability judgments align better with range than with variance in the given contexts. Whereas the observed effects of n on subjective variability could also be captured qualitatively by assuming that people rely on uncorrected sample variance, our model comparison shows that sample range is a better predictor of participants' choices. Furthermore, when the salience of range was manipulated, the model that included the interaction between salience and sample size showed better predictive performance than the model without the interaction—an outcome that the uncorrected sample variance account cannot explain. The results also work against a Bayesian account predicting that participants' answers gravitate toward the middle of the answer scale when sample size is small and then become less regressive as sample size increases: In the SD = 20 condition, we observe the opposite pattern.

Nonetheless, one may argue that the tasks in the first two experiments had limited ecological validity, as people rarely rate variability explicitly in their daily lives. To address this concern, we conducted a third experiment that focused on a more widely studied dependent measure—risk preference—in an established DfE paradigm where people observe samples from an underlying payoff distribution.

Experiment 3: Valuation of Risky Outcome Decreases with Sample Size

If larger sample sizes lead to higher perceived variability, a distribution should appear riskier as one samples more from it. Consequently, risk-averse decision makers should find a positive, uncertain outcome less attractive as they sample more from it. Experiment 3 tests this prediction.

The notion that an outcome can appear riskier as more information about it becomes available may seem counterintuitive because more information decreases uncertainty and hence reduces risk. Differentiating between aleatory uncertainty and epistemic uncertainty can resolve the seeming contradiction (Ülkümen et al. 2016). Variability, measured by either range or SD, arises from the irreducible randomness in the distribution, which is a form of aleatory uncertainty. At the same time, with larger sample sizes, it becomes possible to estimate the parameters of the underlying outcome distribution (e.g., its expected value) more precisely. This uncertainty in parameter estimation, because of a lack of information, constitutes a type of epistemic uncertainty, which can be reduced by acquiring more information. Thus, under the variability-as-range hypothesis, n has two opposing effects on subjective uncertainty. On the one hand, a larger n leads to a wider range, increasing perceived aleatory uncertainty. On the other hand, a larger n also leads to more precise mean estimates, reducing epistemic uncertainty. A prototypical decision maker may seek to reduce both types of uncertainty (Walters et al. 2023). As a result, the net effect of n on the attractiveness of an outcome is not easy to determine.

In experiment 3, we resolve this problem by informing participants about the underlying expected value of the outcomes, hence controlling for epistemic uncertainty. Participants in the experiment learned about the possible outcomes of payoff distributions by sampling from them. For each distribution, participants then stated their willingness to pay for a new random draw. The distributions systematically varied in its EV, SD, and n. Our key interest was the influence of n on participants' valuations. With epistemic uncertainty controlled for, we expected that risk-averse decision makers will value an uncertain outcome less as n increases. We anticipated that the majority of participants in our sample would be risk averse for positive outcomes (Tversky

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

and Kahneman 1992). Therefore, we expected that the negative relationship between n and valuation will hold true for the average participant sample.

As mentioned earlier, sample range increases more steeply with n as SD gets larger. Therefore, besides the main hypothesis that higher n yields lower valuations, we also expected the effect of n on valuation to be more pronounced for a larger SD, which implies a negative interaction between SD and n in the regression model.

Method

Design and Procedure. The payoff distributions were visualized as virtual boxes on the computer screen. Each box generated random samples from a normal distribution with a predefined EV and variance. The numbers were presented as values in U.S. dollars. By pressing the space bar, a randomly sampled value appeared inside the box and stayed visible until the participant pressed again for the next draw. After drawing a predefined number of samples, participants indicated the maximum price they would be willing to pay for *another* random draw from that same box (WTP). This answer was the main dependent variable. Three parameters varied across trials: expected value (EV: 200 or 400), standard deviation (SD: 20 or 40), and sample size (n: 3, 6, 10, or 24). This resulted in 2 (EV) × 2 (SD) × 4 (n) = 16 distinct trials. Each participant encountered all 16 trials once in random order. As noted above, in each trial, participants were informed of the EV of each box before sampling to reduce the epistemic uncertainty of EV. As observed sample means may deviate from the EV of the underlying distribution, we added in the instructions: "Sometimes you may feel that the average of the samples drawn is larger/smaller than the stated average. This is just the nature of random sampling. Please be assured that the stated average is true." Experiment 3 did not include the saliency of range factor (sorted sequence) as in experiment 2 because for sequential displays, a perfectly ascending/descending sequence may give participants strong yet misleading impressions about trends. The task was hypothetical, hence not incentive compatible.

Participants. A total of 188 participants passed the comprehension check and finished the entire task (139 female, mean age = 25.7 years[10]). Participants were compensated with a flat fee of £1.3. On average, it took participants about nine minutes to complete the task.

Results

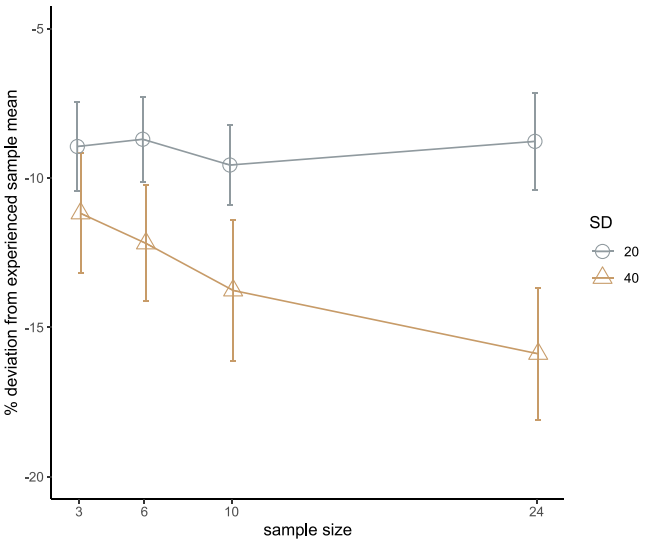
Following the rules for excluding observations reported in the preregistration, data from 47 participants were excluded from further analyses (> 20% responses fell outside EV ± 5 × SD of the relevant condition). Of the 2,256 responses from the remaining 141 participants, 26 additional responses were excluded based on the same criteria. Figure 5 shows the mean percent deviation of participants' WTP from the experienced sample mean as a function of SD and n, averaged across the two EVs.

As can be seen from the figure, valuations decreased for higher SD. This confirmed our expectation that participants were mostly risk averse in the current task. Figure 5 further shows that in the large SD condition, valuations decreased with n. To test our hypotheses on statistical grounds, we fitted a Bayesian linear mixed model with valuation as the dependent variable and sample mean, SD (entered as −0.5 and +0.5 for low and high SD, respectively), n, and SD × n as predictors. The model also included a random intercept and separate random slopes for all the above factors and interactions by participant. Parameter priors were set similarly as in the previous experiment.

Results confirmed the main effect of SD: high SD decreases valuation by an average of \$5.66 (95% HDI (−9.17, −2.18)). More importantly, there is also a main effect of n such that valuation decreases with n (M = −0.25, 95% HDI (−0.41, −0.09)), and there is a credible interaction between SD and n (M = −0.51, 95% HDI (−0.77, −0.24)). The main effect of n on valuations was driven by the high SD condition: at SD = 20, n does not have credible effect on valuation (M = 0.00, 95% CI (−0.14, 0.13)). In contrast, at SD = 40, every additional draw reduces valuation by \$0.50 (95% CI (−0.70, −0.26)). Finally, sample mean is a strong predictor for valuation (M = 0.97, 95% HDI (0.96, 0.99)).

The estimated standard deviation of the slopes for n was 0.67, indicating that whereas most participants valued random rewards after experiencing larger samples

Figure 5. (Color online) Mean Percent Deviation of Valuation as a Function of SD and n, Averaged Across Two EVs in Experiment 3



Note. Error bars = 95% CI.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

less, the strength of this effect varied considerably among participants. Substantial participant-level heterogeneity also emerged in effects for SD × n (SD = 0.68).

To determine whether uncorrected variance could equally account for the results of experiment 3, we conducted a model comparison similar to that in experiment 2. Specifically, we fitted three new Bayesian mixed models, each using the sample mean alongside one additional predictor—sample range (r), corrected sample SD (cs), or uncorrected sample SD (us)—and then compared their fits using the same procedure as in experiment 2. Each predictor reliably explained responses (r: $M = -0.17$, 95% HDI (-0.19, -0.15); cs: $M = -0.53$, 95% HDI (-0.61, -0.46); us: $M = -0.60$, 95% HDI (-0.69, -0.52)). According to ELPD-LOO, the r model provided the best fit, followed by the us model (ΔELPD = -6.19 ± 7.72, $p = 0.423$) and the cs model (ΔELPD = -13.99 ± 9.98, $p = 0.161$). The relatively small differences in ELPD, compared with their standard errors, may stem from the sample mean being a much stronger predictor of willingness to pay than variability. Although the ELPD differences do not reach conventional significance relative to their standard errors, the LOO model weights offer a complementary perspective, showing that the r model captured the vast majority of the model weight ($w = 0.998$), and the us model accounted for the remainder ($w = 0.002$). Because these weights allocate probability mass across the candidate set, even small but consistent predictive advantages can translate into high weight for one model. In this case, the range predictor consistently outperformed the others, yielding the large weight. Consistent with experiment 2, these results suggest that the uncorrected variance account does not explain the findings of experiment 3 as well as the range-based account.

For completeness, we also estimated a mixed regression with sample mean, sample SD, and sample range as predictors. Results of this regression indicate that the effect of sample range on valuations is credible above and beyond the effect of SD. Likewise, the model clearly outperforms a simpler regression that does not include sample range (ΔELPD = $-70.99 + -14.84$).

Discussion

In summary, valuation decreased as sample size increased, as predicted by hypothesis 3. This pattern was only credibly found when SD is relatively large, presumably because in this case, n had a higher leverage on range. In the small SD condition, the increase in range with n relative to the magnitude of the EV may not have been steep enough to influence participants' valuations. Whereas we informed participants about the true EV, observing more samples might still have decreased epistemic uncertainty by providing participants with a better "sense" of what to expect from an additional draw. As we mentioned above, this effect goes the opposite direction such that larger sample size makes an uncertain reward more attractive for risk-averse decision makers. This counter effect may have thwarted the effect of n in the small SD condition.

An alternative explanation for the interaction between sample size and SD is that participants formed a prior around an SD of 20, which larger samples could override. In tasks like the present one, individuals develop a sense of likely parameter ranges through repeated trials and feedback (Oaksford and Chater 2007, Hertwig and Erev 2009). To test this alternative explanation, we conducted a follow-up study of experiment 3 on Prolific using only SD = 40. Because SD = 20 no longer appeared in any trial, there was no reason for participants to hold a prior centered on 20. Results from 84 participants closely match the SD = 40 condition in experiment 3: once again, valuation decreased with sample size ($M = -0.21$, 95% HDI (-0.37, -0.04)). These results render the above alternative explanation for experiment 3 unlikely.

Sample size is a natural parameter of interest in DfE research. Past behavioral experiments on monetary gambles have reported no impact of sample size on valuation (Hoffart et al. 2022) and on choices (Hotaling et al. 2022). In contrast, we found that larger sample sizes lead to systematic changes in preference. Two differences in design can explain this inconsistency. First, unlike both previous studies, we controlled for epistemic uncertainty by informing participants of the expected value of each distribution. As discussed above, when epistemic uncertainty and aleatory uncertainty are both present, it becomes difficult to determine the effect of sample size on a gamble's appeal. Second, following the classical DfE paradigm, both prior studies used two-outcome gambles as stimuli. Once the two possible outcomes are experienced at least once, the range does not increase further with larger sample size. Thus, these previous null results appear consistent with our account.

Experiment 4: Mental Sampling Under Range Constraints

With limited cognitive resources, statistical inferences regarding the complete distributions are usually difficult to achieve. Therefore, people may approximate probabilistic inference by sampling (Griffiths et al. 2012, Vul et al. 2014). Mental simulation of potential outcomes is a form of sampling (Lieder et al. 2018). In the context of variability judgments, people may thus use mentally simulated samples to approximate the variability of the underlying distribution. Consistent with this perspective, subjective perceptions of variability can be elicited by asking people to mentally generate new samples themselves. For example, experiment 5 in Kareev et al. (2002) asked participants to predict what new

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

random draws from a given population would look like. Experiment 2 in the same paper elicited people's beliefs of the likely distance between two random draws from an underlying distribution. Konovalova and Le Mens (2020) reported two experiments that relied on a similar elicitation method. The success of these paradigms indicates that people are able to mentally simulate random samples based on experienced ones.

Our findings so far show that in many contexts, variability judgments depend on range. If people develop perceptions about variability by mentally sampling, it follows that experienced sample range influences the mental simulation of new samples. Experiment 4 aims to explore how range influences such mental simulation. Intuitively, after experiencing a sample from a continuous, unimodal distribution, a decision maker should be highly certain that any element within the experienced range *can* occur (i.e., has greater than zero density), but may be less certain about the possibility of elements outside this range. Hence, we hypothesize that the samples people mentally simulate are distorted such that values outside the experienced range (i.e., the minimum and maximum of an experienced sample) are less likely to occur than what would be expected from normative probability theory (hypothesis 4).

To test this hypothesis, in experiment 4, we presented participants with samples from normal distributions, informed them of the shape of the distribution, and asked them to mentally generate new random samples from the same underlying distribution. Similar tasks that ask participants to mentally generate new random samples based on observed samples have been adopted in empirical studies of distribution learning (Tran et al. 2017, Mason et al. 2024). In addition, we also manipulated the number of samples that participants observed to test the effect of sample size on the variability of generated samples. The fewer samples one experiences, the narrower the experienced range tends to be and the more confining the range is. Therefore, we predicted that the samples that participants generate themselves would become more concentrated—appearing more "compressed" and exhibiting lower variance—after viewing smaller-sized samples.

Method

Design and Procedure. We first introduced participants to the concept that the individual birthweight distributions of many animal species resemble a bell curve, illustrated through a distribution of real human birthweight data. Afterward, participants repeatedly observed number sequences (displayed for one second each with a one-second blank interval in between) on the computer screen. These numbers were described as birthweights, in grams, from randomly sampled specimens of an undisclosed animal species. The instructions further informed participants that the specific species would remain undisclosed and change from trial to trial. Upon viewing the sample, participants were asked to mentally generate a new random sample of size 10 from the same birthweight distribution of the species they just observed. They were told that the order of the samples which they observe and generate is unimportant.

As stimuli for the experiment, we constructed five unique normal distributions that differed in EV and SD, representing five species. For each distribution, we created two samples—one of size 3 and one of size 10—for participants to observe before generating new samples themselves. The mean and SD of the observed samples closely mirrored the EV and SD of their respective underlying distributions. With two trials per species and five species, each participant completed a total of 10 trials. An overview of the stimuli can be found in Table B.1 in the Online Appendix.

Participants. A total of 101 Prolific participants passed the comprehension check and completed the entire task (52 female, mean age = 38.5 years). Participants received a flat fee of £1.94. The task took an average of 13 minutes to complete.

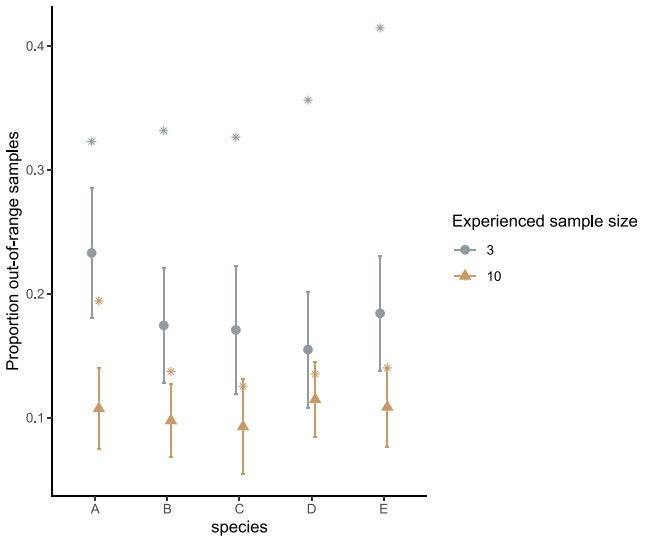
Results

Adhering to the preregistered rules for excluding observations, data from 19 participants were excluded from further analyses (>5 responses fell outside mean $\pm$ 5 $\times$ SD of the corresponding samples). Of the 8,200 responses from the remaining 82 participants, 10 additional responses were excluded based on the same criterion.

To test hypothesis 4, we calculated the proportion of generated samples that lay outside the experienced sample range for each trial and participant, as illustrated in Figure 6. The figure also visualizes the theoretical probability that a random sample falls outside the experienced range: $\Phi[(\min(\boldsymbol{X}_i) - \mu_i)/\sigma_i] + (1 - \Phi[(\max(\boldsymbol{X}_i) - \mu_i)/\sigma_i])$, where Φ is the cumulative distribution function (CDF) of the standard normal distribution, and $\boldsymbol{X}_i$ is the set of experienced samples in the ith trial. Figure 6 reveals that across all trials, the observed out-of-range proportions were lower than theoretically predicted. This observation is substantiated by one-sample t-tests: for the combined set of five trials with $n = 3$, an average of 9.16 out of 50, or 18%, of generated samples were out of range, which is credibly lower than the theoretical prediction of 35% ($BF_{10} > 10^{13}$). Similarly, for the five trials with $n = 10$, an average of 5.21, or 10%, of generated samples were out of range, which is also credibly lower than the theoretical prediction of 15% ($BF_{10} = 43$). These results confirm hypothesis 4, which predicts that people are less likely to generate samples outside of experienced range than predicted by the underlying distribution.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Figure 6. (Color online) Mean Proportion of Out-of-Range Samples Generated for Each Trial, Averaged Across Participants

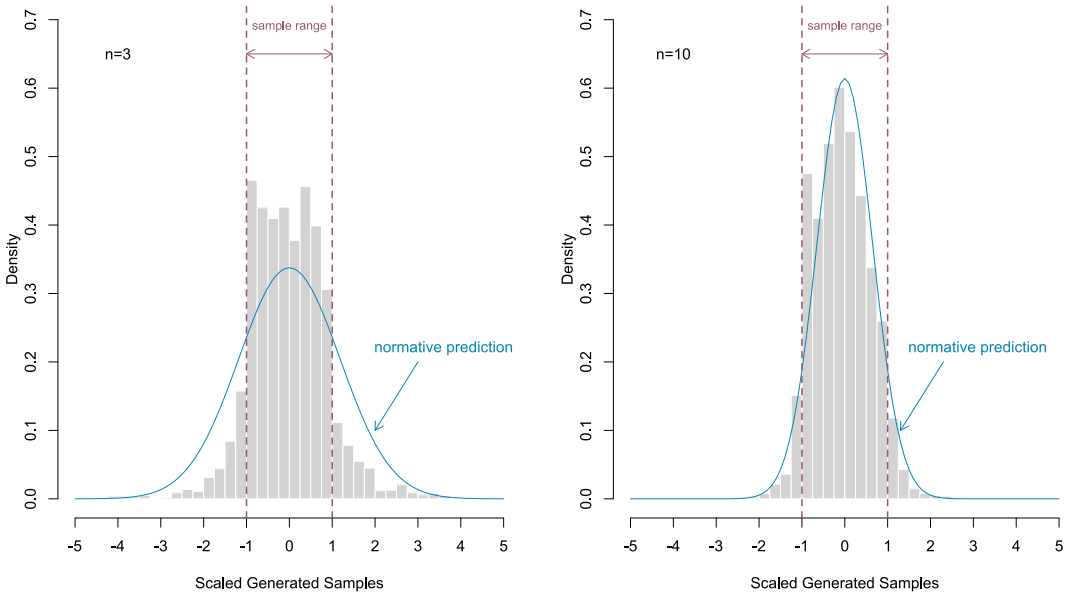


Notes. Error bar = 95% CI. The asterisks mark the probability of out-of-range values implied by the underlying density function.

Figure 7 provides a more intuitive illustration of the range effect. The figure shows the distribution of all responses, linearly transformed for each participant in each trial such that the minimum value of the corresponding *experienced* (i.e., observed) samples in the ith trial, $\min(\boldsymbol{X}_i)$, is set to -1, and $\max(\boldsymbol{X}_i)$ is set to $+1$. The (blue) solid curve in the figure shows the theoretical density function (i.e., the underlying distribution of the observed samples), also rescaled so that -1 and $+1$ represent the expected value of the minimum and the maximum for the respective n. As can be seen in the figure, at $n = 3$, the empirical density everywhere inside the sample range is above the theoretical curve, whereas the empirical density outside the sample range is mostly below the theoretical curve. Additionally, the empirical response distribution shows visible discontinuities around -1 and $+1$, indicating a sudden drop in frequency when values cross the boundaries of experienced samples. Such a range effect is not as visually pronounced at $n = 10$, which is not surprising for two reasons. First, as the experienced sample size grows, the sample range tends to expand, producing a wider "fence" and therefore exerting less of a constraining effect on subsequent responses. Second, identifying and memorizing the precise minimum and maximum in a sequential display becomes more cognitively challenging as n increases, which can further weaken the effect of range.

More concentrated responses should yield smaller reconstructed variance. To test this, we examined how experienced range influenced the variability of generated samples. We first standardize all responses by $z_{ijk} = (x_{ijk} - \mu_i)/\sigma_i$, for the ith trial, jth participant, and kth response, so we could average the variances of generated samples across trials and participants and compare them against one (the variance of the experienced distributions). At $n = 3$, the mean variance of standardized responses was 0.59, credibly lower than one ($BF_{10} > 10^{10}$). At $n = 10$, the mean variance was

Figure 7. (Color online) Distribution of Generated Samples for $n = 3$ and $n = 10$, Rescaled to $\min(\boldsymbol{X}_i) = -1$ and $\max(\boldsymbol{X}_i) = +1$, Where $\boldsymbol{X}_i$ Is the Set of Experienced Samples in the ith Trial



Notes. Left: $n = 3$. Right: $n = 10$. The vertical dotted lines represent the experienced sample range. The solid curves represent the density functions of the experienced distributions.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

0.86, also credibly lower than one ($BF_{10} = 7.34$). Moreover, the mean variance at $n = 3$ was smaller than that at $n = 10$ ($BF_{10} > 10^6$). Thus, reconstructed distributions exhibit reduced variance relative to experienced distributions, with the reduction stronger at smaller sample sizes. These results parallel the relationship between sample size and perceived variability observed in experiment 2.

Discussion

To summarize, participants' generated samples were less dispersed than the original distributions, particularly at smaller sample sizes. This result dovetails with experiments 2 and 3, where larger sample size leads to increased perceived variability and perceived risk, respectively. If mental simulations of random samples are constrained by the experienced range, the reconstructed distribution will be overly dense within that range and underrepresented outside it. The narrower the range, the more compressed the reconstructed density. Accordingly, distributions learned from larger samples are perceived as more variable.

A Cognitive Fence Hypothesis for the Range Effect. As a possible explanation of the observed effect of range in experiment 4, decision makers may have been influenced by two distinct factors when attempting to generate random samples. First, guided by experimental instructions and statistical intuition, a decision maker's perceived density function of animal birthweights, denoted as $f(x)$, was continuous and bell-shaped. Second, upon observing a sample $\boldsymbol{X}$ with $\min(\boldsymbol{X}) = l$ and $\max(\boldsymbol{X}) = h$, people were certain that all values within the sample range are possible; that is, $f(x) > 0$ for all $l \le x \le h$. In contrast, they remained uncertain whether any value outside the sample range is possible. Consequently, they may have mentally discounted out-of-range values relative to within-range values, leading to underrepresentation of out-of-range values in the posterior distribution. We refer to this as the *cognitive fence hypothesis.*

Discounting values outside the experienced range is not necessarily a reasoning flaw, as it can be ecologically valid in real-life contexts. Theoretically, the normal density $\varphi(x)$ is greater than zero for all $x \in \boldsymbol{R}$, but this usually does not hold true for real-life data that only approximate normal distributions. For example, the density function for birthweights of human babies certainly assumes a value of zero for any $x \le 0$ or $x \ge 50$ kilograms.

Because the mental discount is categorical (certainly possible within the experienced range and uncertain otherwise), it can lead to discontinuities of the subjective density at l and h. More formally, without the discount, for a continuous distribution, the subjective density should satisfy $\lim_{x\to l^-} f(x) = \lim_{x\to l^+} f(x) = f(l)$ and $\lim_{x\to h^-} f(x) = \lim_{x\to h^+} f(x) = f(h)$. However, with the discount, the subjective density has the property that $\lim_{x\to l^-} f(x) < \lim_{x\to l^+} f(x)$ and $\lim_{x\to h^-} f(x) > \lim_{x\to h^+} f(x)$. A direct corollary of these inequalities is that, immediately inside the boundaries l and h, the empirical density should exceed the normative density, whereas immediately outside these boundaries, the empirical density should fall below the normative density.

As shown in Figure 7, the response distributions exhibit apparent discontinuities at the boundaries of the experienced range, consistent with the cognitive fence hypothesis. As a first step to statistically test this discontinuity, we defined symmetric windows of width δ around each boundary. For the left (minimum) boundary, the inside and outside windows were $[l, l + \delta)$ and $[l - \delta, l)$, respectively. For the right (maximum) boundary, the inside and outside windows were $(h - \delta, h]$ and $(h, h + \delta]$. Consistent with the histogram visualizations in Figure 7, where $[l, h]$ is divided into eight bins, we set δ = sample range/8. We then compared, for each participant, the total number of responses within each window to the number expected if responses were randomly drawn from the underlying continuous distribution N(sample mean, sample SD). The discontinuity prediction is supported if counts within the inside windows exceed these normative expectations, whereas counts within the outside windows fall below them.

Normative expectations were estimated via a bootstrap simulation procedure in which we generated 10,000 simulated data sets, each consisting of 82 virtual participants (matching the empirical sample size). For each simulated data set, participants generated samples from the underlying distribution N(sample mean, sample SD). We then computed the mean number of responses falling within each window for each simulated data set, yielding a distribution of 10,000 simulated means. Table 2 reports the empirical mean number of responses within each window across all experimental conditions, alongside the corresponding expected means derived from the bootstrap simulation.

Table 2. Mean Number of Samples Observed in Experiment 4 Compared with a Bootstrap-Based Normative Expectation, Derived from the True Underlying Distribution, for Narrow Windows Immediately Inside and Outside of the Observed Range

	Left boundary		Right boundary	
Data source	Outside $[l-\delta, l)$	Inside $[l, l+\delta)$	Inside $(h-\delta, h]$	Outside $(h, h+\delta]$
Empirical	3.85 (2.82)	11.74 (5.18)	7.04 (4.03)	2.89 (2.40)
Simulated	4.64 (2.09)	7.58 (2.64)	6.00 (2.36)	3.74 (1.88)

Note. Values in parentheses indicate standard deviations across experiment 4 participants or standard errors of means for the simulated participants.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

For both boundaries, the observed pattern is consistent with the predicted discontinuity: empirical counts substantially exceed normative expectations within the inside windows and fall below them in the outside windows. The observed means are more extreme than virtually all bootstrap means, corresponding to $p < 0.0001$ in each case. These results are robust across a range of reasonable values of δ.

Toward a more comprehensive definition and test of the cognitive fence hypothesis, we proceed by specifying it as a mathematical model. The model assumes only two components in how people form subjective distributions. First, individuals can to some extent learn the underlying distribution, which is supported by prior research (Goldstein and Rothschild 2014, Sun et al. 2019, Lindskog et al. 2021, Mason et al. 2024). Second, consistent with the fence hypothesis, individuals disproportionately weight values within the experienced range relative to those outside it. In our test, this overweighting and underweighting is captured in a categorical way—values inside the observed range receive a uniform boost, whereas values outside do not. These two components are not intended to provide a complete account of how people build subjective distributions. The actual cognitive process is almost certainly more complex, and there may be many other factors that influence the observed behavior. Our goal here is to demonstrate that even under these highly simplified assumptions, introducing the discontinuity predicted by the fence hypothesis captures the observed data in a parsimonious way.

We formalize these two components using a statistical mixture model. In this framework, responses are represented as a combination of (i) the true underlying density, capturing the first component, and (ii) a uniform boost applied to values within the observed sample range, capturing the second component. We refer to this statistical device as the *fence model*. However, it is not intended as a literal model of cognitive process. Rather, it serves as a parsimonious statistical device to test the central predictions of the cognitive fence: that values inside the observed range are overweighted and those outside it are underweighted and that discontinuities emerge at the boundaries of the experienced range.

Specifically, this test assumes that participants' subjective density $f(x)$ is a mixture of the true underlying density—in the context of experiment 4, a normal distribution $N(\mu, \sigma^2)$ with density $g(x)$, where μ and σ^2 truthfully reflect the observed distribution—and a uniform distribution $U(l, h)$ with density $h(x)$, where l and h are experienced sample minimum and maximum, respectively,

$$f(x) = (1 - w) \cdot g(x) + w \cdot h(x).$$

In this specification, $g(x)$ corresponds to the first cognitive component mentioned above and $h(x)$ to the second. The parameter w controls the relative weight of the two components and is the sole free parameter estimated from observed data. For any $w > 0$, the model predicts a discontinuity in $f(x)$ right at the edges of the observed range. If the relative discounting of out-of-range values meaningfully shapes participants' responses, we should observe aggregate estimates of w that are substantially greater than zero. Additionally, because l and h are supposedly more difficult for participants to identify with larger sample sizes, the influence of range should diminish with larger samples. Accordingly, we expect the estimate of w at $n = 10$ to be smaller than that at $n = 3$.

To test these predictions, we estimated w in a hierarchical Bayesian framework separately for each sample size, using the statistical software Stan via the RStan interface (Stan Development Team 2023), with default uninformative priors. The parameter w was modeled as varying across participants with a distribution $N(\mu_w, {\sigma_w}^2)$. Results show that, as predicted, at $n = 3$, μ_w is credibly larger than zero ($M = 0.52$, 95% HDI (0.44, 0.59)). At $n = 10$, μ_w is also credibly larger than zero ($M = 0.39$, 95% HDI (0.32, 0.46)). Moreover, as predicted, μ_w is credibly larger at $n = 3$ than at $n = 10$ ($M = 0.12$, 95% HDI (0.03, 0.22)). The hierarchical standard deviation σ_w reveals considerable individual differences in the estimates of w: at $n = 3$, $M = 0.30$, 95% HDI (0.25, 0.35), and at $n = 10$, $M = 0.24$, 95% HDI (0.19, 0.29).

As a sanity check, the model should yield $w \approx 0$ when data are generated purely from $g(x)$, and $w \approx 1$ when data are generated purely from $h(x)$. To validate this, we simulated random samples exclusively from $g(x)$ or $h(x)$ and reestimated w with these simulated data. As expected, when the data are generated purely from $g(x)$, the estimated μ_w is indeed very close to zero (95% HDI (0.01, 0.04)). When the data are generated purely from $h(x)$, the estimated μ_w is almost exactly one (95% HDI (1.00, 1.00)).

The model also allows for a rigorous statistical test of the predicted discontinuities at the boundaries of the experienced range. Simply finding $w > 0$ is insufficient to establish this, because an evenly compressed smooth distribution could also yield a positive w. To directly assess the discontinuity prediction, we compared the above fence model with a baseline model.

In this baseline model, y_{ijk}, the kth generated sample on the ith trial by the jth participant, follows $N(\mu_i, (\lambda_j \sigma_i)^2)$, where μ_i is the experienced sample mean, and σ_i is the experienced sample standard deviation. The parameter λ quantifies the extent to which the variability of the reconstructed distribution deviates from the variability in the observed distribution. Thus, the baseline model can capture compression but not discontinuities in distributional shape. If the fence model outperforms this baseline model in model fit, the

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

improvement can be attributed specifically to its ability to capture discontinuities.

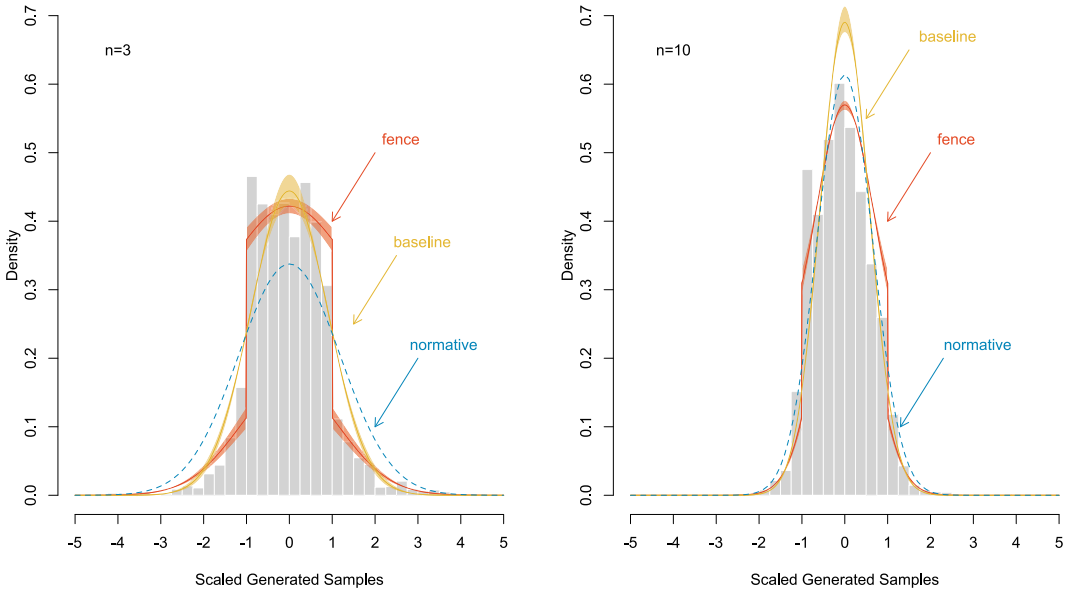
Analogous to w in the fence model, λ in the baseline model varies across participants according to $N(\mu_\lambda, \sigma_\lambda{}^2)$. Because the variance of $f(x)$ in the fence model can only fall between the variance of $h(x)$ and the variance of $g(x)$, we constrained the baseline model for comparability. Specifically, $\lambda_j\sigma_i$ was limited to the interval $[\sigma^h, \sigma^g]$, which corresponds to bounding all λ_j within $[\min_i\{\sigma_i^h/\sigma_i\}, 1]$. Without this constraint, the baseline would trivially outperform the fence model by flexibly accommodating outlier responses (e.g., a participant generating the number 23 after observing 10 samples in the 60s).

The estimated μ_λ of the baseline model again confirms that the generated samples were less dispersed at $n = 3$ ($M = 0.76$, 95% HDI (0.72, 0.80)) than at $n = 10$ ($M = 0.89$, 95% HDI (0.86, 0.91)). We quantified the predictive accuracies of both models using expected log pointwise predictive densities (ELPD) and estimated them using the leave-one-out cross validation (Vehtari et al. 2017). At $n = 3$, the fence model demonstrated higher expected predictive accuracy than the baseline model (ΔELPD $= 72.97 \pm 22.68$, $p = 0.001$). Similarly, at $n = 10$, the fence model also outperformed the baseline model in terms of expected predictive accuracy (ΔELPD $= 70.06 \pm 20.94$, $p < 0.001$). Given that both models are designed to be equivalent in their capacity to capture the first and second moments of the response distributions, the superior predictive accuracies of the fence model suggest its enhanced ability to account for the higher-order shapes of the response distributions, namely, the discontinuities mentioned earlier. The result of this model comparison was backed up by visualizations of the model predictions in Figure 8. As can be seen in the figure, at $n = 3$, the fence model represents a visible improvement over the baseline in capturing the discontinuities at l and h in the aggregate data. For the same reason mentioned earlier, the improvement is less noticeable at $n = 10$.

General Discussion

What metrics do people rely on to judge the variability of data? The four experiments reported here provide converging evidence that subjective perceptions of variability depend on observed range rather than variance. Experiment 1 showed that comparison of variability predominantly relied on sample range, not sample variance, when the two metrics produced conflicting conclusions. Experiments 2 and 3 reported an increase in perceived variability and, presumably because of risk aversion, a decrease in the valuation of positive outcomes as sample size increased. Finally, experiment 4 demonstrated that an experienced sample range confined mental simulation of new random samples, offering a cognitive explanation by which range may shape the perception of variability. In the following, we will discuss these findings in the light of related work, followed by a review of limitations and an outlook on future research.

Figure 8. (Color online) Distribution of Generated Samples and Predicted Density by Different Models



Notes. The histogram represents the distribution of generated samples, rescaled to $\min(\boldsymbol{X}_i) = -1$ and $\max(\boldsymbol{X}_i) = +1$, where $\boldsymbol{X}_i$ is the set of experienced samples in ith trial. The curves labeled "normative" represent the density functions of the experienced distributions. The curves labeled "fence" and "baseline" represent the predictions of the fence model and the baseline model, respectively. The error regions indicate the 95% HDIs. Left: $n = 3$. Right: $n = 10$.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Sample Size, Perceived Variability, and Distributional Learning

In experiment 2, we showed that perceived variability increased with sample size. This finding aligns with earlier research by Kareev et al. (2002), who found that people tend to underestimate variability and that such underestimation is more pronounced for smaller sample sizes. Whereas Kareev's uncorrected sample variance account can explain the results of experiment 2 qualitatively, our model comparison showed that the sample range better predicts subjective variability than the uncorrected sample variance. Furthermore, the "uncorrected sample variance" cannot explain why the effect of sample size on perceived variability becomes stronger when elements within a sample are sorted. This moderation effect does make sense, though, if variability is mentally represented by sample range instead of sample variance. The "uncorrected sample variance" account also cannot explain the distributions of generated samples in experiment 4: the compression of the distribution is not uniform over its support but contingent on experienced range. In contrast, the manipulations used in Kareev et al. (2002), such as out-of-view versus in view, sample size, and working memory capacity, essentially all had the effect of altering the sample size available to decision makers when evaluating variability and are thus consistent with the range hypothesis.

Relatedly, Konovalova and Le Mens (2020) noted that unbiased estimators of sample variance are more prone to underestimating than overestimating the population variance ($P(\hat{\sigma}^2 < \sigma^2) > 0.5$). Furthermore, as the sample size increases, the likelihood of this underestimation diminishes. Consequently, it's more likely than not that the sample variance of a larger sample will be greater than that of a smaller sample ($P(V_N^2 > V_n^2) > 0.5$). Konovalova and Le Mens (2020) tested the mathematical prediction empirically in two experiments by asking participants to identify the more variable option between two samples of different sizes. Results showed that more than half of the participants perceived the large sample as more variable than the small sample, thereby confirming the main prediction. Interestingly, when the larger sample exhibited higher variance, 63% and 65% of participants in experiments 1 and 2, respectively, chose it as the more variable option. Conversely, when the larger sample had a smaller sample variance, it was still selected as the more variable option by 50% and 56% of participants, respectively. Whereas the results confirmed the main theoretical prediction, it also showed a tendency toward perceiving larger samples as more variable even when they had lower variance, which cannot be explained by the comparison of actual sample variances. This pattern, however, is consistent with the "variability as range" hypothesis because the larger sample may have a wider sample range even when its variance is lower (e.g., the inconsistent samples in experiment 1).

The perception of variability is closely related to how people mentally represent and reconstruct probability distributions from observed samples. In experiments by Tran et al. (2017) and a follow-up by Mason et al. (2024) on that question, many participants mentally generated samples that tended to flatten into a more uniform shape across the observed range, even though the true distributions they observed were distinctively bimodal. This pattern of mental representation can be readily understood from the perspective of the fence model. Whereas a thorough reanalysis of these data is beyond the scope of this manuscript, the fence model predicts that mental representations of distributions in these tasks gravitate toward uniform shapes. However, this does not imply that people generally fail to mentally represent probability distributions. Indeed, previous research using different presentation and elicitation formats found that people can produce more precise representations when cues or scaffolding are sufficiently salient (e.g., Goldstein and Rothschild 2014). Some participants also achieved high accuracy in the experiments by Mason et al. discussed above.

Cognitive Fence, Sample Size, and Overprecision

To the extent that people rely on mentally simulated samples to approximate the variability of an underlying distribution, the cognitive fence hypothesis provides a mechanistic explanation for the observed effect of sample size on perceived variability in experiment 2. As outlined next, it can also help understand other behavioral phenomena. For example, research on overconfidence shows that when people give interval estimates for a quantity with a specified confidence level (X% sure that the quantity is within a lower bound and an upper bound), the correct answers often fall outside their estimated intervals more frequently than their confidence level would suggest (Soll and Klayman 2004, McKenzie et al. 2008). Juslin et al. (2007) attribute this overprecision to the "naive intuitive statistician" model, where people mistakenly apply observed sample characteristics directly to the broader population. Our cognitive fence hypothesis implies that overprecision can persist even if estimates of population properties are not limited to the sample properties. Even if people actively attempt to consider many possible cases beyond the observed samples, because of the inherent difficulty in imagining out-of-range cases, people's subjective distributions will still be too concentrated and their subjective confidence intervals too narrow, based on the given information. The cognitive fence account also helps explain why interval estimates often lead to higher degrees of overconfidence than other mathematically equivalent elicitation methods of subjective probability distribution (e.g.,

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Is the population of Germany larger than 70 million? Yes/No and *How confident are you?*). This is because the cognitive fence induced by experienced range is more compatible with the response mode of interval production than other formats.

Relatedly, Camilleri and Newell (2019) investigated how individuals generate predictions based on observed samples. They found that participants who observed samples from a distribution tended to generate new samples that fell within the observed range, rather than extrapolating beyond it. Additionally, participants' predictions of future events closely mirrored their recollections of past observations. These findings support our fence account, indicating a theoretical connection to exemplar learning, as participants demonstrated difficulty in extending beyond the examples they had encountered. In line with this, Lejarraga and Lejarraga (2020) found that decision makers who sampled more data were less confident about the observed information and the choice thereafter as compared with situations in which they sampled less. The authors explain this negative correlation between experience and confidence based on the theoretical insight that smaller samples amplify the differences between options, thus making choices easier (Hertwig and Pleskac 2010). This phenomenon also resembles the Dunning–Kruger effect (Kruger and Dunning 1999), where individuals with limited knowledge in a domain overestimate their competence. The variability as range hypothesis presents a complementary theoretical account for these phenomena. It suggests that, because of a lack of exposure to the broader spectrum of potential outcomes, individuals with restricted information might perceive a given estimate as less uncertain and more precise than it truly is, resulting in overconfidence. This overconfidence can be reduced as individuals gain more experience and encounter a more extensive range of possibilities.

Overweighting and Underweighting of Extreme Events and Risk Perception

A sizable literature on decisions from experience has shown that people sometimes underweight rare outcomes when making repeated choices (Barron and Erev 2003, Hertwig et al. 2004). It has been proposed that such underweighting arises because, in these tasks, people typically rely on small samples that often omit rare outcomes altogether or because rare outcomes are psychologically underrepresented because of recency effects when larger samples are available (Hertwig and Erev 2009). This stands in apparent contrast to rich evidence of overweighting of rare extreme outcomes in certain estimation or memory-based tasks (e.g., Spitzer et al. 2017, Lieder et al. 2018, Appelhoff et al. 2023). The cognitive fence hypothesis helps reconcile these gaps. In our statistical test of the hypothesis, over- or underrepresentation of outcomes can be captured as the difference between the subjective density designated by the mixture model and the objective density function of the outcome: $d(x) = f(x) - g(x) = w[h(x) - g(x)]$. For any $0 < w < 1$ and any unimodal class $g(x)$, $d(x)$ peaks at the extreme values within the experienced range, which implies overweighting of previously experienced extreme values. Conversely, the fence model predicts underrepresentation for outcomes beyond one's experience, as $h(x)$ drops to zero outside this range. Therefore, according to the model, whether an extreme event is overrepresented or underrepresented depends on whether it lies within the experienced range. If someone just experienced a new record high of 41°C in their city, they will perceive another 41°C day as more likely than before, as $f(41)$ increased because $h(41)$ jumped up from zero after the experience. But that person may not necessarily perceive an even higher temperature like 42°C as more likely than before because of lack of prior experience, as $f(42)$ remains unchanged because $h(42)$ remains at zero. Hence, the model is compatible with findings that extremely rare events are often underweighted in DfE paradigms, as these outcomes are frequently unencountered in limited samples or psychologically discounted in experience because of recency effect. At the same time, the model explains how a single experience of a rare event can loom disproportionately large in subjective risk perception—particularly for real-life "black swan" events, which tend to be memorable and resistant to recency effects.

This insight also leads to the inference that human risk perception may not be "forward looking" and that we may be ill-prepared for extreme events we never experienced, such as extreme weather events or black swan geopolitical risks. Understanding this process is hence crucial for effective risk communication. Strategies of risk communication should provide more comprehensive information about the probability of various outcomes, not just those within common experience. It also underscores the importance of widening the range of experiences or examples when educating people about risks to help them form a more accurate understanding of less familiar, yet possible, outcomes. This approach could lead to more informed decision making and better risk management strategies (see also Sundh 2024). Human societies may have implicitly relied on this intuition in the past. For example, to remind citizens of local flood risks, historic flood and drought marks engraved in stone can be found across medieval cities in Europe (Brázdil et al. 2006). These range indicators outlast centuries and hence exceed the experience that could be made within a single lifespan.

Limitations and Future Research

Throughout our four experiments, we employed continuously distributed stimuli, where sample range is a common measure of statistical dispersion. Although

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

range may also serve as an important cue in many discrete contexts, we do not claim that it is a definitive indicator of subjective variability in all situations. For example, in two-outcome gambles commonly used in decisions-from-experience studies (Wulff et al. 2018), the range can only take on values of zero or the difference between the two outcomes, which may inadequately capture perceived variability. Consider a gamble with two outcomes ($0, 0.9; $100, 0.1): after the first $100 win following many $0 outcomes, the range is fixed, yet subsequent samples are still informative of the variability of the distribution. Even within continuous domains, we do not propose that range always outperforms variance. There are likely contexts where variance will provide a better account of subjective variability. For example, Jeong and Chong (2021) found that in ensemble perception tasks with orientation arrays, participants' variability judgments were more consistent with a weighted standard deviation—capturing internal dispersion among items—than with range, especially when different sets shared the same range but differed in how values were distributed. There may also be other measures of dispersion that people intuitively use and understand depending on the context, such as percentiles. Likewise, although the fence model provides a better explanation for the behavioral data in experiment 4 compared with the normal benchmark, it still leaves some variance unexplained. This suggests the presence of individual differences and additional cognitive processes—for instance, variations in how people subjectively perceive range within a given context, potentially influenced by selective attention to observed values and memory-related processes. Given these considerations, our contribution is not to elevate range as a universal replacement but to demonstrate that range can, under identifiable conditions, better predict subjective variability than variance—challenging the default status of variance-based measures as the sole or primary benchmark.

To underscore the foundational nature of our hypothesis, we purposefully employed experimental paradigms that are rather abstract and elemental. Whereas these approaches are crucial for establishing a theoretical basis, future research could further test the robustness of our accounts by extending investigations into more context-rich settings, ideally incorporating real-life data (see Levitt and List (2007) for discussion of the necessity of field experiments to assess whether laboratory findings generalize to natural settings). For instance, the application of our hypothesis to the domain of finance presents a compelling avenue for exploration. If retail investors assess the variability of a security based more on range than on variance, existing variance-based models (e.g., the option pricing model) may benefit from corresponding revisions.

In the scope of this paper, we also did not discuss how the relative magnitudes of mean and range may affect perceived variability. Given the substantial evidence that estimates of variability are affected by the magnitude of the mean (Pollard 1984, Weber et al. 2004), it is plausible that a relative range measure (e.g., range/mean) could offer a more veracious representation of perceived variability than the absolute range alone. For instance, in experiment 3, although we did not find the effect of n on valuation with a mean of 200 and SD of 20, the effect could be pronounced when the mean assumes a significantly smaller value, such as 50.

The cognitive fence hypothesis predicts that the probability of generating or imagining values drops sharply outside the experienced range. We evaluated this prediction using a statistical model that mixes the true underlying density with a uniform boost within the observed range. Whereas this specification is grounded in theoretical considerations and captures key qualitative patterns in the data, it should not be interpreted as a genuine cognitive process model. The underlying cognitive process that governs perceptions of variability is likely more multifaceted. This provides an opportunity for future research to further develop and refine the framework. For example, Shepard's law of generalization states that the perceived similarity between two objects is an exponentially decaying function of their distance on relevant dimensions (Shepard 1987). From this perspective, one could assume a "soft fence" in which the probability of generating a number beyond the observed range is a continuous probability function that exponentially decays with the distance from the experienced minimum and maximum.

Conclusions

This paper presents converging evidence that people's subjective perceptions of variability are strongly influenced by sample range. These findings support the hypothesis that range—by virtue of its simplicity and salience in daily life—shapes the perceptions and estimates of variability. This challenges the predominant reliance on variance in decision-making theories and underscores the need to consider alternative, more psychologically plausible measures of subjective variability in models of human judgment and decision making. The range as variability hypothesis not only helps bridge existing theoretical perspectives but also carries practical implications for understanding risk assessment and decision making in real-world contexts. Finally, it opens promising avenues for future research on how range-based perceptions of variability guide human behavior.

Endnotes

[1] See https://www.qualtrics.com.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

[2] See https://lab.js.org/.

[3] See https://mindprobe.eu.

[4] See https://www.prolific.com.

[5] Preregistrations, experimental programs, data, and codes for analyses are available at https://researchbox.org/2628.

[6] In the preregistration of this study, we intended to exclude all data of participants who fail to reach 2/3 accuracy in the *Consistent* condition. The actual data indicates that we underestimated the difficulties of the task. Obeying the criterion would have excluded 127, or 32%, of all participants. Therefore, we decided to remove this data exclusion criterion completely. No participant was excluded in the analyses. Note that this deviation from the preregistration works *against* our primary hypothesis: excluding these 127 participants would increase the average proportion of wider-range choices.

[7] Treating the dependent variable as ordinal (assuming that participants can only distinguish the rankings of variability represented by the histograms) leads to the same pattern of results.

[8] This relationship can be derived analytically when the expectation of the range takes a closed analytical form. For instance, as mentioned in the Introduction, for uniform distributions, $\mathrm{E}[range] = [\sqrt{12}(n-1)/(n+1)]\mathrm{E}[SD]$. From this, we can derive $\frac{\partial^2 \mathrm{E}(range)}{\partial \mathrm{E}(SD)\partial n} = \frac{4\sqrt{3}}{(n+1)^2} > 0$. This means precisely that E(*range*) increases more steeply with n as E(*SD*) gets larger. Although the expected range of a normal distribution has no simple closed-form in elementary functions, it can be closely approximated by $\mathrm{E}(range) \approx (3\sqrt{\ln n} - 1.5)\sigma$ (Ramirez and Cox 2012). It is easy to see that $\frac{\partial^2 \mathrm{E}(range)}{\partial \sigma \partial n} > 0$ is true for any $n \geq 2$.

[9] We conducted these analyses with sample SD, not sample variance, for the sake of consistent units of measurement. The results remain unchanged should we use sample variance instead.

[10] Demographic information is unavailable for five participants. Shortly before the implementation of experiment 3, Prolific experienced a large wave of signups because of a viral trend on social media, "skewed heavily towards female participants in their 20s" (Charalambides 2021). Our sample was clearly skewed as well.

References

Andraszewicz S, Scheibehenne B, Rieskamp J, Grasman R, Verhagen J, Wagenmakers EJ (2015) An introduction to Bayesian hypothesis testing for management research. *J Management* 41(2): 521–543.

Appelhoff S, Hertwig R, Spitzer B (2023) Control over sampling boosts numerical evidence processing in human decisions from experience. *Cerebral Cortex* 33(1):207–221.

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Memory Language* 68(3):255–278.

Barron G, Erev I (2003) Small feedback-based decisions and their limited correspondence to description-based decisions. *J. Behav. Decision Making* 16(3):215–233.

Beach LR, Scopp TS (1968) Intuitive statistical inferences about variances. *Organ. Behav. Human Performance* 3(2):109–123.

Black F, Scholes M (1973) The pricing of options and corporate liabilities. *J Political Econom.* 81(3):637–657.

Blitz D, Pang J, van Vliet P (2013) The volatility effect in emerging markets. *Emerging Markets Rev.* 16(1):31–45.

Brázdil R, Kundzewicz ZW, Benito G (2006) Historical hydrology for studying flood risk in Europe. *Hydrological Sci. J.* 51(5): 739–764.

Bürkner PC (2017) brms: An R package for Bayesian multilevel models using Stan. *J Statist. Software* 80(1):1–28.

Camilleri AR, Newell BR (2019) Better calibration when predicting from experience (rather than description). *Organ. Behav. Human Decision Processes* 150:62–82.

Charalambides N (2021) We recently went viral on TikTok—Here's what we learned. *Prolific* (August 24), https://www.prolific.com/resources/we-recently-went-viral-on-tiktok-heres-what-we-learned.

Cox LA (2008) Why risk is not variance: An expository note. *Risk Anal.* 28(4):925–928.

Goldstein DG, Rothschild D (2014) Lay understanding of probability distributions. *Judgment Decision Making* 9(1):1–14.

Griffiths TL, Vul E, Sanborn AN (2012) Bridging levels of analysis for probabilistic models of cognition. *Current Directions Psych. Sci.* 21(4):263–268.

Hamilos CA, Pitz GF (1977) The encoding and recognition of probabilistic information in a decision task. *Organ. Behav. Human Performance* 20(2):184–202.

Hertwig R, Erev I (2009) The description–experience gap in risky choice. *Trends Cognitive Sci.* 13(12):517–523.

Hertwig R, Pleskac TJ (2010) Decisions from experience: Why small samples? *Cognition* 115(2):225–237.

Hertwig R, Barron G, Weber EU, Erev I (2004) Decisions from experience and the effect of rare events in risky choice. *Psych. Sci.* 15(8):534–539.

Hoffart JC, Jarecki JB, Dutilh G, Rieskamp J (2022) The influence of sample size on preferences from experience. *Quart. J. Experiment. Psych.* 75(1):1–17.

Hotaling JM, Donkin C, Jarvstad A, Newell BR (2022) MEM-EX: An exemplar memory model of decisions from experience. *Cognitive Psych.* 138:101517.

Jeong J, Chong SC (2021) Perceived variability reflects the reliability of individual items. *Vision Res.* 183:91–105.

Johnson RA (2017) *Miller & Freund's Probability and Statistics for Engineers*, 9th ed. (Pearson, London).

Jorion P (1995) Predicting volatility in the foreign exchange market. *J. Finance* 50(2):507–528.

Juslin P, Winman A, Hansson P (2007) The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psych. Rev.* 114(3):678–703.

Kacelnik A, Bateson M (1996) Risky theories—The effects of variance on foraging decisions. *Amer. Zoologist* 36(4):402–434.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–292.

Kareev Y, Arnon S, Horwitz-Zeliger R (2002) On the misperception of variability. *J. Experiment. Psych. General* 131(2):287–297.

Konovalova E, Le Mens G (2020) An information sampling explanation for the in-group heterogeneity effect. *Psych. Rev.* 127(1): 47–73.

Konovalova E, Pachur T (2021) The intuitive conceptualization and perception of variance. *Cognition* 217:104906.

Kruger J, Dunning D (1999) Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Personality Soc. Psych.* 77(6):1121–1134.

Lathrop RG (1967) Perceived variability. *J. Experiment. Psych.* 73(4):498–502.

Lejarraga T, Lejarraga J (2020) Confidence and the description–experience distinction. *Organ. Behav. Human Decision Processes* 161:201–212.

Levitt SD, List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? *J. Econom. Perspect.* 21(2):153–174.

Lieder F, Griffiths TL, Hsu M (2018) Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psych. Rev.* 125(1):1–32.

Lindskog M, Nyström P, Gredebäck G (2021) Can the brain build probability distributions? *Front. Psych.* 12:596231.

Downloaded from informs.org by [2409:8a6c:e2:3654:944e:32b0:6e7f:100e] on 19 November 2026, at 10:51 . For personal use only, all rights reserved.

Lovie P (1978) Teaching intuitive statistics: II. Aiding the estimation of standard deviations. *Internat. J. Math. Ed. Sci. Tech.* 9(2): 213–219.

Lovie P, Lovie AD (1976) Teaching intuitive statistics I: Estimating means and variances. *Internat. J. Math. Ed. Sci. Tech.* 7(1): 29–39.

Markowitz H (1952) Portfolio selection. *J. Finance* 7(1):77–91.

Mason A, Szollosi A, Newell BR (2024) Learning the lie of the land: How people construct mental representations of distributions. *J. Experiment. Psych. Learning. Memory Cognition* 51(4): 515–525.

Mason A, Madan CR, Simonsen N, Spetch ML, Ludvig EA (2022) Biased confabulation in risky choice. *Cognition* 229:105245.

McElreath R (2018) *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (CRC Press, Boca Raton, FL).

McKenzie CRM, Liersch MJ, Yaniv I (2008) Overconfidence in interval estimates: What does expertise buy you? *Organ. Behav. Human Decision Processes* 107(2):179–191.

Oaksford M, Chater N (2007) *Bayesian Rationality: The Probabilistic Approach to Human Reasoning* (Oxford University Press, Oxford, UK).

Olschewski S, Scheibehenne B (2024) What's in a sample? Epistemic uncertainty and metacognitive awareness in risk taking. *Cognitive Psych.* 149:101642.

Peterson CR, Beach LR (1967) Man as an intuitive statistician. *Psych. Bull.* 68(1):29–46.

Pollard P (1984) Intuitive judgments of proportions, means, and variances: A review. *Current Psych. Res. Rev.* 3(1):5–18.

Ramirez A, Cox C (2012) Improving on the range rule of thumb. *Rose-Hulman Undergraduate Math. J.* 13(2):1–13.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16(2):225–237.

Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science* 237(4820):1317–1323.

Simon HA (1990) Invariants of human behavior. *Annual Rev. Psych.* 41(1):1–20.

Soll JB, Klayman J (2004) Overconfidence in interval estimates. *J. Experiment. Psych. Learn. Memory Cognition* 30(2):299–314.

Spitzer B, Waschke L, Summerfield C (2017) Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behav.* 1(8):1–8.

Stan Development Team (2023) Stan reference manual. Accessed October 25, 2026, https://mc-stan.org/docs/reference-manual/.

Sun J, Li J, Zhang H (2019) Human representation of multimodal distributions as clusters of samples. *PLoS Comput. Biol.* 15(5):e1007047.

Sundh J (2024) Human behavior in the context of low-probability high-impact events. *Humanities Soc. Sci. Comm.* 11(1):1–10.

Tran R, Vul E, Pashler H (2017) How effective is incidental learning of the shape of probability distributions? *Roy. Soc. Open Sci.* 4(8):170270.

Triola MF (2004) *Elementary Statistics*, 9th ed. (Pearson/Addison-Wesley, Boston, MA).

Tversky A, Kahneman D (1992) Advances in prospect-theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4):297–323.

Ülkümen G, Fox CR, Malle BF (2016) Two dimensions of subjective uncertainty: Clues from natural language. *J. Experiment. Psych. Generation* 145(10):1280–1297.

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statist. Comput.* 27(5):1413–1432.

Vul E, Goodman N, Griffiths TL, Tenenbaum JB (2014) One and done? Optimal decisions from very few samples. *Cognitive Sci.* 38(4):599–637.

Walters DJ, Ülkümen G, Tannenbaum D, Erner C, Fox CR (2023) Investor behavior under epistemic vs. aleatory uncertainty. *Management Sci.* 69(5):2761–2777.

Weber EU, Shafir S, Blais AR (2004) Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psych. Rev.* 111(2):430–445.

Wulff DU, Mergenthaler-Canseco M, Hertwig R (2018) A meta-analytic review of two modes of learning and the description–experience gap. *Psych. Bull.* 144(2):140–176.