Cognition 265 (2025) 106277

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

When metrics matter: How reasoning in different metrics impacts judgments of uncertainty

David Zimmerman [a,1,*], Stephen A. Spiller [a], Nicholas Reinholtz [b], Sam J. Maglio [c,d]

[a] Anderson School of Management, University of California, Los Angeles, 110 Westwood Plaza, Los Angeles, CA 90095, USA
[b] Leeds School of Business, University of Colorado, Boulder, Boulder, Colorado 80309, USA
[c] University of Toronto Scarborough, Toronto, Ontario M1C 1A4, Canada
[d] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada

ARTICLE INFO

Keywords:
Uncertainty
Numerical cognition
Metric
Judgment

ABSTRACT

The uncertainty of a point estimate, often conceptualized and quantified via a prediction interval, can vary in both magnitude (e.g., width) and symmetry. Further, when people make many types of estimates, they can use different but equitable metrics (e.g., feet vs. meters). In a series of experiments, we investigate whether using different metrics impacts people's estimates of uncertainty. Three empirical regularities guide our focus: First, people believe risk scales with magnitude, reporting greater uncertainty for bigger point estimates, leading to inconsistent prediction intervals across metrics differing by a fixed additive constant. Second, people are insufficiently sensitive to unit changes, leading to inconsistent prediction interval widths across metrics differing by a multiplicative constant. Third, people tend to assume that distributions are symmetric, leading to inconsistent symmetry across metrics differing by an inverse transformation. Together, these three regularities exemplify how uncertainty estimations are sensitive to metric in substantive ways.

1. Introduction

Alice receives an offer of $300,000 for her house. Should she accept it, or reject it in hopes for a better one? Bobby is throwing a party and has five six-packs of beer. Should he buy more to hedge against running out? Claire is shopping online for a new, fuel-efficient car. She finds one that gets 32 miles-per-gallon. Is that efficient enough, or should she keep looking for other, potentially more efficient options?

Each of these scenarios involves uncertainty: Is there someone who will offer more than $300,000 for Alice's house? How many beers will Bobby's guests drink? Can Claire find a car with substantively better fuel economy? The subjective distribution of outcomes for these estimates will influence what actions people take. For example, if Bobby thinks there is less than a 10 % chance his guests will drink more than 30 beers, he might decide against getting more.

At the same time, many quantities can be represented in different yet equivalent metrics.[2] Alice's $300,000 home with a $240,000 mortgage can be considered in terms of its value ($300,000) or her equity ($60,000). We examine whether the metric used in reasoning about uncertainty (e.g., whether Alice focuses on her home's value or her home equity, whether Bobby ponders his supply as "how many beers" or "how many six packs of beer", and whether Claire considers how many miles she gets from a gallon of gasoline or how many gallons it takes to go 1000 miles) influences people's perceptions of uncertainty. We focus on situations with multiple metrics that have a one-to-one mapping to test whether equivalent metrics can lead to inconsistent judgments of uncertainty.

Specifically, we focus on three types of transformations, both prevalent and potentially susceptible to inconsistencies: (i) addition of a constant (e.g., Alice's home value vs. home equity), (ii) multiplication by a constant (e.g., Bobby's count of beers vs. six packs), and (iii) inversion of a ratio (e.g., Claire's consideration of miles-per-gallon vs. gallons-per-thousand-miles). In each case, prior research suggests discrepancies may occur in judgments of uncertainty: (i) People's judgments of risk and uncertainty scale with numeric magnitude (Hogarth, 1975; Weber, Shafir, & Blais, 2004), suggesting potential discrepancies between metrics that differ by an additive constant. (ii) People tend to be insensitive to units (Burson, Larrick, & Lynch, 2009; Raghubir &

* Corresponding author.
E-mail addresses: ZimmermanDa@sec.gov (D. Zimmerman), stephen.spiller@anderson.ucla.edu (S.A. Spiller).
[1] Present address: Securities and Exchange Commission, Washington DC, 100 F St NE, Washington DC, 20549.

[2] Throughout, we use "metric" and "unit of measurement," or simply "unit" interchangeably.

https://doi.org/10.1016/j.cognition.2025.106277
Received 25 July 2024; Received in revised form 30 June 2025; Accepted 22 July 2025
Available online 5 August 2025
0010-0277/Published by Elsevier B.V.

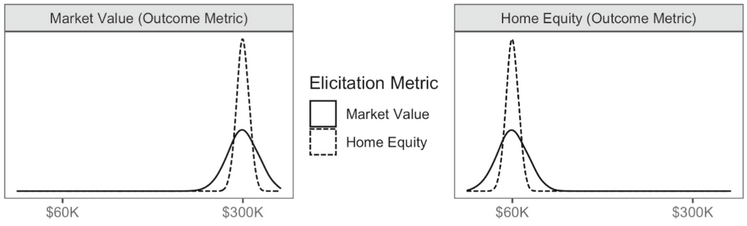
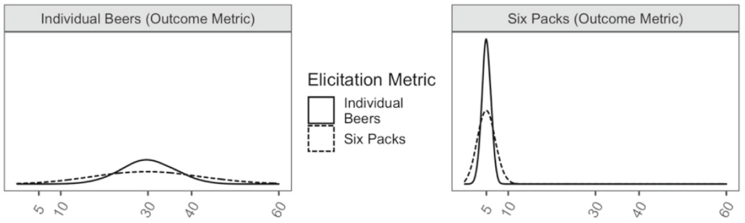
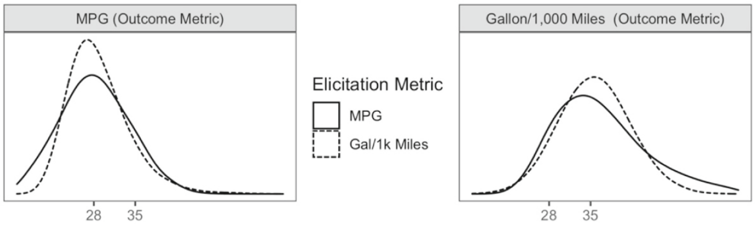
Metric Relationship	Outcome Metric A	Outcome Metric B	Prediction
Risk Scales with Magnitude (Units Differ by Constant)		Market Value = Home Equity + $240,000 (Mortgage Balance)	H1: Greater uncertainty for more numerous unit Tested in Experiments 1a-b
Unit Insensitivity (Units Differ by Multiple)		$Individual\ Beers = Six\ Packs * 6$	H2: Greater uncertainty for the larger unit Tested in Experiment 2
Symmetry Assumption (Units Inversely Related)		$\frac{Miles}{Gallon} = \frac{1,000}{\frac{Gallon}{1,000\ Miles}}$	H3: More symmetric uncertainty for elicited metric Tested in Experiments 3a–c

Fig. 1. Predicted Results of Metrics on Uncertainty Distributions. The line types indicate the elicitation metric for each estimated distribution. The outcome metric columns show estimated uncertainty in each outcome metric. When the outcome metric is different than the elicitation metric, it is the transformation, shown below each pair of plots, of the distribution where the outcome and elicitation metric match in the other plot (e.g., the home equity line in the market value outcome metric is just the elicitation of home equity plus $240,000). The expected differences in the uncertainty between elicitation metrics is stated in the prediction column.

Srivastava, 2002), suggesting potential discrepancies between metrics that differ by a multiplicative constant. (iii) People tend to believe distributions are symmetric (Flannagan, Fried, & Holyoak, 1986), suggesting potential discrepancies between metrics that differ via inversion of a ratio.

1.1. Risk scales with magnitude

Estimates of subjective risk are positively related to both the standard deviation of the observed data and the mean of the distribution (Hogarth, 1975; Weber et al., 2004). Thus, if people are told that that the mean of a distribution is higher, they expect the distribution to have a higher standard deviation (Hofstätter, 1939; Reinholtz, 2015). For example, if people see distributions with equal variability, they give estimates with more variability for the distribution with the higher mean (Beach & Scopp, 1968; Lathrop, 1967; Weber et al., 2004).

As a result, considering an unknown quantity in one of two alternative metrics that differ by an added constant might lead to inconsistent perceptions of the variability. We predict that metrics which are more numerous will lead people to report greater uncertainty than the uncertainty of estimates in their less-numerous counterpart metrics. For example, Alice could form expectations for the value for her home in terms of sale price or her equity. Because these two metrics differ by a known amount (her $240,000 mortgage), they should have the same variability. Nonetheless, we expect reasoning about sale price will have greater uncertainty than reasoning about her equity simply from the higher numerosity of the sale price.

1.2. Unit insensitivity

People are insufficiently sensitive to unit when making evaluations or predictions. This holds in contexts ranging from cell phone plans to calorie information for food (Burson et al., 2009; Pandelaere, Briers, & Lembregts, 2011; Shen & Urminsky, 2012; Wertenbroch, Soman, & Chattopadhyay, 2007). For instance, Americans spend lavishly in Europe but more stingily in Mexico, in part because they evaluate the number on the price tag without fully appreciating that the units are in euros (leading to smaller numbers on price tags) or pesos (larger numbers on price tags). More generally, people working in units that are a multiple of a familiar unit do not sufficiently adjust their evaluations in the less familiar unit (Raghubir & Srivastava, 2002, see also Maglio & Trope, 2011; Pelham, Sumarta, & Myaskovsky, 1994).

This suggests people who reason about uncertainty in different units that are multiples of each other may also be insufficiently sensitive to the unit. For example, Bobby might consider the quantity of beer he expects to be consumed at his party in terms of individual beers or six packs.

Uncertainty expressed in six-packs ought to be one-sixth that expressed in individual beers. We predict that when one unit is larger than the other, and thus each one-unit increase reflects a bigger change in the larger unit, people will insufficiently adjust their estimates of uncertainty. Thus, estimates of uncertainty will be higher for the larger unit.

1.3. *Default distribution is symmetric*

For estimates, uncertainty can be conceptualized as a distribution of possible values. People tend to assume the distribution of possible values is symmetric (Flannagan et al., 1986). Moreover, across several elicitation formats (e.g., CDF, PDF, hypothetical future samples, etc.) people express beliefs that are remarkably close to a normal distribution without any requirement to do so (Winkler, 1967).

Assuming uncertainty is symmetric might lead to an *asymmetric* distribution after a nonlinear transformation to convert an uncertainty estimate to an alternative metric. Likewise, if symmetric-expecting people initially reason in this alternative metric, converting would lead to asymmetric uncertainty in the original, focal metric. Previous work has shown that inversely-related metrics (e.g., miles-per-gallon and gallons-per-1000 miles) lead many people to make incorrect evaluations of the benefits from changes in the two metrics (de Langhe & Puntoni, 2016; Larrick & Soll, 2008; Peer, 2010). For example, Claire's car hunt might lead her to reason about the distribution of fuel efficiency for gas-powered cars in the US in miles-per-gallon. Given the findings above, she will likely assume a roughly symmetric distribution. This implies a positively skewed distribution of the scaled reciprocal: gallons-per-1000 miles. Yet if Claire were to reason about the distribution of fuel efficiency for gas powered cars in gallons-per-1000 miles, she would again likely assume a roughly symmetric distribution, necessitating a difference in the symmetry of the uncertainty between those equivalent metrics. Accordingly, we predict that when metrics are inversely related, people should give more symmetric uncertainty in the focal metric, creating asymmetry upon converting to other metrics.

Our key hypotheses are summarized in Fig. 1.

1.4. *Overview of experiments*

We test our three hypotheses in six experiments. Experiments 1a–b investigate the impact of an additive constant, eliciting estimates either of revenue or profit for a business with a fixed cost structure (i.e., revenue = profit + constant). If people are influenced by numerical magnitude when considering uncertainty, we should expect greater uncertainty, or equivalently, wider prediction intervals for revenue than for profit. Experiment 2 considers the impact of a multiplicative constant by eliciting estimates of seltzer water sales in 24-packs, 8-packs or individual cans (i.e., individual cans = 8-packs × 8) in a grocery store. If people insufficiently adjust to the unit when estimating values, then those estimating in 24- or 8-packs (e.g., the larger unit) will end up with greater uncertainty – or equivalently, wider, real prediction intervals – because people do not sufficiently adjust their nominal intervals for the units in which they report. Experiments 3a-c consider the impact of ratio transformations (e.g., estimates of Mechanical Turk HIT income in minutes/dollar or cents/min) on the symmetry of uncertainty. If people are more likely to assume a symmetric uncertainty distribution, then ratio transformations will make uncertainty more positively skewed once transformed.

In all studies, participants gave an estimate for the upper and lower bounds of a prediction interval, often with a centrality estimate (like the interval elicitation procedure of Soll & Klayman, 2004). The order of the elicitation (i.e., 90th percentile, central estimate, then 10th percentile) was counterbalanced, between-subjects. Participants were randomly assigned to give estimates in one of two metrics, except in Experiment 2 where participants gave one estimate in individual cans and one in packs. The focal dependent measure was either the width of the prediction interval (e.g., the upper bound - the lower bound) or the symmetry of the prediction interval. Results for the dependent measure that were not predicted to differ are reported in the supplemental materials, with simplified results in Table 1. All data collection, except for experiment 1b and 2, was approved by UCLA's IRB. Experiment 1b and 2 were approved by CU Boulder's IRB. Pre-registrations, data, code, and codebooks for all experiments can be found here: https://researchbox.org/490. The stargazer package (Hlavac, 2018) was used to generate regression tables and the printy package (Mahr, 2024) was used to generate in-text statistics.

2. Experiment 1a: Fortune teller

2.1. *Participants*

200 participants (78 women, median age = 36) recruited from Amazon's Mechanical Turk[3] (AMT) completed this study. Participants were paid $1.40 and the median response time was about 8 minutes, for an hourly rate of $11.13/h. The sample size was based on having 80 % power to detect a moderate effect size, Cohen's d of about 0.4.

2.2. *Procedure*

Participants completed a training in which they were given detailed information about a hypothetical person's subjective expectations for a specific quantity (e.g., the temperature in Denver the next day) and how this person should respond to the type of prediction interval questions the participant would later have to answer. Before advancing to the focal task, participants went through two more hypothetical examples, answering questions designed to assess their understanding of prediction intervals and receiving feedback on the first question when incorrect. Full details of this training are available in the supplemental materials.

Next participants were randomly assigned to make predictions about either *revenue* or *profit*. They then learned about the specific estimation task. First, participants read basic information about a fortune telling booth at a carnival. Specifically, that this fortune teller could complete three fortunes per hour for eight hours each day, for a total of 24 fortunes per day, for five days per week. Further, they read that the fortune teller earned $30 per completed fortune and that renting the booth cost the fortune teller $1500 for the week. Participants had to correctly answer three multiple choice questions to advance to the next phase: how much the booth cost to rent, how much money would be brought in if the fortune teller had clients for a full day, and the payment from telling fortunes for 70 people over the course of the week in their assigned metric.

Next participants made predictions about either the *revenue* or *profit*, that this fortune teller would make in a five-day period. They provided a point prediction and their 80 % prediction interval in their assigned metric. Due to the problem parameters described above, revenue was bounded between $0 and $3600. Participants estimated the 90th percentile (i.e., "I am 90% sure that the [revenue (total sales)/profit (total sales minus the booth rental)] earned will be **less** than ___"), the average (i.e., "My **best guess** for the [revenue (total sales)/profit (total sales minus the booth rental)] earned is___"), and the 10th percentile (i.e., "I am 90% sure that the [revenue (total sales)/profit (total sales minus the booth rental)] earned will be **more** than ___") of revenue or profit.

After answering demographic questions, participants were given an instructional attention check question (bogus text with the instructions to select "other" and type "pen"; see supplemental materials for the full question).

[3] Eligibility for 1a and all additional AMT samples was limited to participants with U.S. IP addresses who had at least 100 approved HITs, with a minimum 95 % approval rating, and had not taken part in any other HITs from this project.

Table 1
Summary of Main Findings from Six Experiments.

	Width		Symmetry	
Experiment	Predicted Effect	Result (95 % CI Cohen's *d*)	Predicted Effect	Result (95 % CI Cohen's *d*)
1a - 80 % PI (Profit vs Revenue)	Higher for revenue metric	(0.00, 0.62)	No impact	(−0.22, 0.39)
1b - 90% PI (Profit vs Revenue)	Higher for revenue metric	(0.31, 0.76)	No impact	(−0.30, 0.15)
2–80 % PI (Cans vs Packs)	Higher for larger metric	(0.19, 0.52)	No impact	(−0.20, 0.14)
3a - 80 % PI (Cents/Min vs Min Dollar)	No impact	(0.09, 0.90)	Reduced symmetry after inverse transformation	(−1.13, −0.32)
3b - 80 % PI (US Cents/Lira vs Lira/USD)	No impact	(−0.35, 0.26)	Reduced symmetry after inverse transformation	(−0.89, −0.27)
3c - 80 % PI (Gallons/$40 vs $/Gallon)	No impact	(−0.48, −0.02)	Reduced symmetry after inverse transformation	(−0.70, −0.24)

2.3. Exclusions

As pre-registered, we excluded data from participants for whom any of the three elicited or implied revenue estimates fell outside of the range [$0, $3600] (23 participants) and anyone who failed the instructional attention check at the end of the study (10 additional participants). This left a final sample size of 167.

2.4. Impacts on interval width

Consistent with our proposal, prediction intervals were significantly wider on average when participants gave estimates of revenue (larger magnitudes) than of profit (smaller magnitudes; $M_{Revenue} = 1123.3$, $M_{Profit} = 900.8$; $t(165) = 2.00$, $p = .047$, 95 % CI = [1.65, 220.91], Cohen's $d = 0.31$; Fig. 2; see Table 1 for summary of all results).

3. Experiment 1b: Fortune teller replication

3.1. Participants

403 (147 women, median age = 39) recruited from CloudResearch's Connect panel completed this study (Hartman et al., 2023). Participants were paid $1.40 and the median response time was about 10 minutes, for an hourly rate of $8.24/h. The sample size was based on having 80 % power to detect an effect size with Cohen's d of about 0.2.

3.2. Procedure

The procedure for this study was identical to Experiment 1a except all 80 % prediction intervals were replaced with 90 % prediction intervals, both in the training and in the main elicitation. Participants were asked for upper and lower bound estimates where they were 95 % sure the outcome would be more than and 95 % sure the outcome would be less than, respectively.

3.3. Exclusions

As pre-registered, we excluded data from participants for whom any of the three elicited or implied revenue estimates fell outside of the range [$0, $3600] (53 participants) and anyone who failed the instructional attention check at the end of the study (45 additional participants). This left a final sample size of 305.

3.4. Impacts on interval width

Consistent with our proposal and replicating Experiment 1a's results, prediction intervals were significantly wider on average when participants gave estimates of revenue (larger magnitudes) than of profit (smaller magnitudes; $M_{Revenue} = 1420.5$, $M_{Profit} = 1017.6$; $t(303) = 4.70$, $p < .001$, 95 % CI = [117.06, 285.84], Cohen's $d = 0.54$; Fig. 3).

3.5. Discussion

In Experiments 1a-b, we found that participants reported wider prediction intervals, and thus greater uncertainty, for metrics with more numerous estimates, revenue, than less numerous estimates, profit. This supports the proposal that metrics can impact relative uncertainty. Next, we test whether people are insensitive to unit when the units are multiplicatively related instead of differing by a constant.

4. Experiment 2: Seltzer water sales

Our second hypothesis is that unit insensitivity affects uncertainty. If people are insufficiently sensitive to unit, they are likely to report prediction intervals that imply greater uncertainty (after transforming all intervals into the same units) when each individual unit represents a greater quantity (e.g., 8-packs of seltzer water) than when each

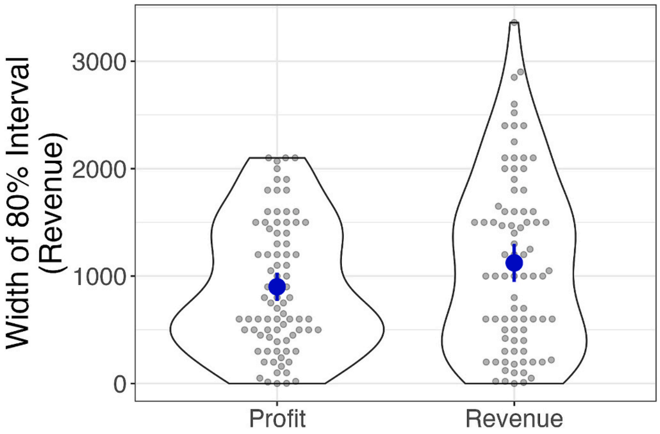


Fig. 2. Prediction Interval Width in Revenue by Elicitation Metric. Error bars represent 95 % confidence intervals of the condition means.

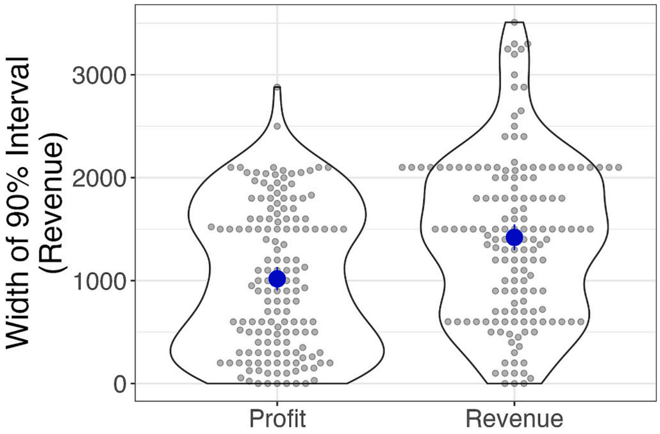


Fig. 3. Prediction Interval Width in Revenue by Elicitation Metric. Error bars represent 95 % confidence intervals of the condition means.

individual unit represents a smaller quantity (e.g., individual cans of seltzer water). As in Experiment 1a, Experiment 2 (and the remaining studies) elicit an 80 % prediction interval.

4.1. Participants

1003 participants (477 women, median age = 39) recruited from Connect completed this study. Participants were paid \$0.80 and the median response time was about 3 and a half minutes, for an hourly rate of \$13.65/h.

4.2. Procedure

Participants were introduced to the task and saw a picture of each of two beverage aisles meant to represent the two specific stores to help them estimate sales. They were told they would be estimating all seltzer water sales for these two stores, labeled Store A and Store B, on a specific day about a week away.

Next, participants made predictions about sales of seltzer water. They were randomly assigned to estimate *packs* at either Store A or Store B and *individual cans* at the other store. They learned Store A sold 8-packs and Store B sold 24-packs. Next, they estimated the 90th percentile (i.e., "I am 90% sure that the [number of **(8 or 24)-packs** of cans/total number of **individual cans**] sold will be **less** than ___"), the average (i.e., "My **best guess** for the [number of **(8 or 24)-packs** of cans/total number of **individual cans**] that will be sold is ___"), and 10th percentile (i.e., "I am 90% sure that the [number of **(8 or 24)-packs** of cans/total number of **individual cans**] sold will be **greater** than ___") at one store followed by the other. Store order was counterbalanced between participants. Finally, they responded to demographic questions and the same attention check item from Experiment 1a.

4.3. Exclusions

As pre-registered, we excluded data from participants for whom any of the three elicited or implied can estimates were greater than three standard deviations away from the estimated statistic (e.g., a 90th percentile estimate for cans sold that was greater than three standard deviations away from the average of all 90th percentile estimates; 8 participants), anyone who failed the attention check at the end of the study (253 additional participants), anyone with a duplicate geolocation (155 additional participants), and anyone with a duplicate IP address (0 additional participants). This left a final sample size of 587.

4.4. Impacts on interval width

The distribution of interval widths was highly skewed, thus we log transformed one plus the widths of the prediction intervals for analysis and report the exponentiated averages below. Consistent with our proposal, the transformed prediction interval widths were significantly wider when participants gave estimates of sales in packs of seltzer water than in individual cans (geometric means: $M_{Packs} = 398.1$, $M_{Individual} = 190$; $b = 0.34$, $t(583) = 4.26$, $p < .001$, 95 % CI of difference in transformed means = [0.18, 0.50], Cohen's $d = 0.37$; Fig. 4). Further, the impact of metric was significantly larger for the 24-packs than the 8-packs, where the difference in prediction interval widths between 24-packs and individual cans is larger for Store B (Store B: $M_{Individual} = 258.3$, $M_{24-packs} = 718.2$) than the difference in prediction interval width between 8-packs and individuals cans for Store A (Store A: $M_{Individual} = 143.6$, $M_{8-packs} = 203.4$; interaction: $b = 0.17$, $t(583) = 2.10$, $p = .036$, 95 % CI of interaction of difference in transformed means = [0.01, 0.33]). Within-subjects analysis showed qualitatively similar results, with the effect of 24-packs being larger (24-packs vs individuals cans for Store B: $b = 1.11$, $t(1170) = 12.84$, $p < .001$, 95 % CI of difference in transformed means = [0.94, 1.28]) than the effect of 8-packs (8-packs vs individual cans for Store A: $b = 0.72$, $t(1170) = 8.96$, $p < .001$, 95 % CI of difference in transformed means = [0.57, 0.88]; interaction: $b = 0.38$, $t(1170) = 2.44$, $p = .015$, 95 % CI of interaction of difference in transformed means = [0.08, 0.69]; see Online Supplement for analysis details).

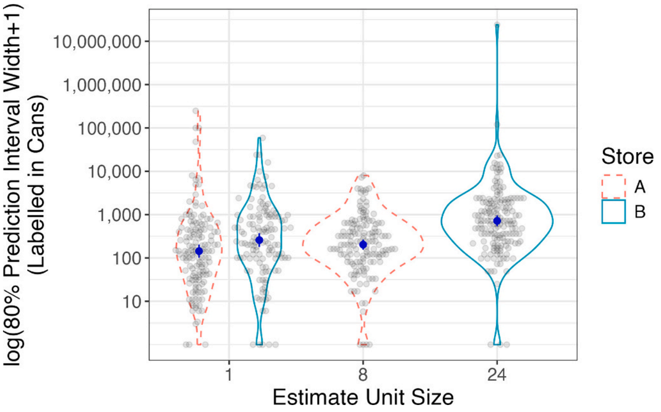


Fig. 4. 80 % Prediction Interval Width in Individual Cans by Elicitation Metric and Store. The outcome is the natural log of prediction interval width plus one, with the labels exponentiated back into individual cans. Error bars represent 95 % confidence intervals for condition means. Results only include first observation for each participant (i.e., the between-subject analysis).

4.5. Discussion[4]

In Experiment 2, we found that people give wider (real) prediction intervals when they estimate values with larger units. This is consistent with people insufficiently adjusting for the elicitation unit, where estimates that should be 8 or 24 times larger, on average, are not scaled sufficiently. This leads to (real) intervals in the packs condition that are wider when all their estimated values (10th percentile and 90th percentile) are scaled up. This effect appears sensitive to the magnitude of the unit difference: larger differences in metrics create larger differences in prediction intervals.

5. Experiment 3a: Wage rates

In Experiments 3a–3c, we explore where metrics that differ by a ratio transformation may lead to inconsistent beliefs about uncertainty. If people are more likely to assume a symmetric uncertainty distribution, then ratio transformations will impact the bounds of the prediction intervals. We expect that participants will believe that uncertainty is relatively symmetric in whatever focal metric they consider. Thus, we propose this will create skewness upon converting to an alternative metric.

5.1. Participants

153 participants (61 women, median age = 35) recruited from AMT completed this study. Participants were paid \$1.05 and the median response time was about 5 minutes, for an hourly rate of \$12.56/h. The sample size was chosen to have 80 % power to detect a moderate effect

[4] We also tested the effect of metric differing by a multiplicative constant in the context of egg sales. Participants either estimated sales for a specific day in individual eggs or dozens of eggs. We found the same general result: Estimates in the larger unit (i.e., dozens) led to wider prediction intervals than estimates in the smaller unit (i.e., individual eggs). This experiment had an incidental error in the instrument programming for a measure other than the main dependent variable, thus it has been excluded. See the Online Supplement for details.

size, Cohen's d of around 0.5.

5.2. Procedure

Participants first completed a set of training exercises describing 80 % prediction intervals and how to construct them. The training had the same types of information and structure as Experiment 1a, but the specific examples we used were different (see supplemental materials). Then they learned about the specific estimation task.

Next, participants gave estimates about their earnings rate in the last 100 paid human intelligence tasks (HITs) they completed. They were randomly assigned to make estimates in *cents/*min or in *minutes/dollar*. As guidance, they were asked to mentally order, from the lowest to highest payment rate or minutes spent on the task (based on assigned metric), their last 100 paid HITs, inclusive of any bonuses. Participants estimated the 90th (i.e., "What were the [minutes spent per dollar earned/payment rate in cents per minute] for the 90th HIT in your list of 100?"), 50th and 10th percentiles of earnings rate in minutes spent per dollar earned or cents earned per minute. The metrics are not exact inversions so that magnitude of expected estimates remains similar across metrics, thereby avoiding the possible confounding effects of uncertainty scaling with magnitude. Finally, they responded to demographic questions and the same attention check item from Experiment 1a.

5.3. Exclusions

As pre-registered, we excluded data from participants for whom any of the three elicited or implied estimates were outside of the range [1, 100] in either metric, implying hourly wages of less than $0.60 or greater than $60 (33 participants), anyone who gave non-monotonic estimates for the percentiles of heights of US adults in the training exercise (16 additional participants), and anyone who failed a simple attention check (5 additional participants). This left a final sample size of 99. Note that all estimates are converted to the same metric and then the range exclusion criterion is applied.

To equate across the two metric conditions, we convert our outcome variables for all participants into minutes/dollar. All the results are qualitatively the same, and the effect sizes are very similar, when estimates are converted to cents/min (see supplemental materials). Both the width and symmetry outcomes were highly skewed, thus we applied a log transformation. For ease of interpretation, we exponentiated the condition average for the outcome statistics.

5.4. Impacts on interval symmetry

Symmetry was operationalized as a ratio between distances of the central estimate to the edges: log((90th percentile – 50th percentile) / (50th percentile – 10th percentile)). We refer to this measure as the skew score. Consistent with our predictions, intervals left in their elicited minutes/dollar metric had lower skew scores than estimates transformed from cents/min (geometric means: $M_{cents/minute} = 2.5$, $M_{minutes/dollar} = 1.4$, $t(95) = -3.57$, $p < .001$, 95 % CI of difference in skew score = [−0.42, −0.12], Cohen's $d = 0.72$). Equivalently, intervals were more positively skewed when transformed from their elicited metric relative to intervals where the elicitation and outcome metrics are the same (Fig. 5).

6. Experiment 3b: Exchange rates

6.1. Participants

202 participants (95 women, median age = 36) recruited from AMT completed this study. Participants were paid $1.30, with a median response time of about 7 minutes, for an hourly rate of $11.77/h. The sample size was chosen to have 80 % power to detect a moderate effect size, Cohen's d of around 0.5.

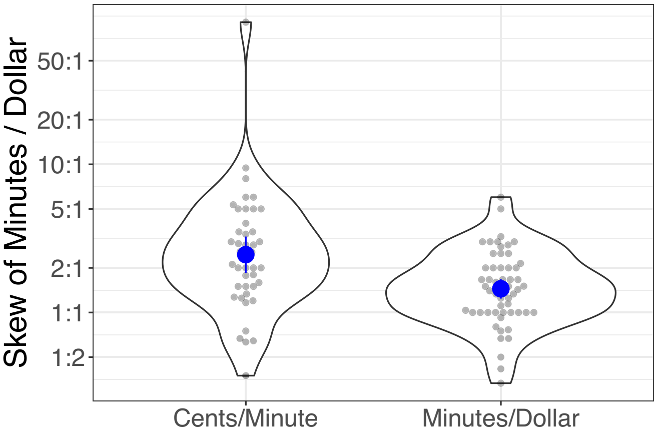


Fig. 5. Prediction Interval Skew Scores in Minutes / Dollar by Elicitation Metric. The outcome is the skew score with the labels exponentiated and transformed into ratios for interpretability. Error bars represent 95 % confidence intervals for condition means.

6.2. Procedure

Participants first completed a set of training exercises describing 80 % prediction intervals and how to construct them, the same as Experiment 1a. Next, they learned about exchange rates and had to answer two multiple choice questions translating between US dollars and Australian dollars for a given exchange rate. For questions answered incorrectly, participants were shown the correct answer and the logic behind it. Then they learned about the specific estimation task. They reviewed historical exchange rate information between Turkish and US currencies. The information covered the last 11 months in the metric they would use for estimation.

Next, participants made predictions about the exchange rate between Turkish Lira and US dollar or cents in two weeks to evoke a sense of uncertainty. They were randomly assigned to make predictions in *cents/Lira* or *Lira/Dollar*. Based on historical rates, the magnitudes of these metrics would be similar, thereby reducing the confounding effects of uncertainty scaling with magnitude. Participants estimated the 90th percentile (i.e., "I am 90% sure that the [US cents per Turkish Lira/ Turkish Lira per US dollar] exchange rate will be **less** than ___"), their best guess, and 10th percentile of the exchange rate in US cents per Turkish Lira or Turkish Lira per US dollar. Finally, they responded to demographic questions and the same attention check item from Experiment 1a.

6.3. Exclusions

As pre-registered, we excluded data from participants for whom any of the three estimates were outside of the range [13.915, 2.4562] when converted to Lira/Dollar or, equivalently, [40.713, 7.186] in US cents/ Lira (25 participants) and anyone who failed a simple attention check (8 additional participants); these values were chosen as they represented a plausible range of the expectations someone could hold for the exchange rate based on historical rates.[5] This left a final sample size of 169.

To equate across the two metric conditions, we converted our outcome variables for all participants into Lira/US dollar. When we use US cents/Lira all the results are qualitatively the same, and the effect sizes are very similar (see supplemental materials). Both the width and

[5] See the pre-registration, Experiment 3b – AsPredicted #64058, for the precise calculations and justification.

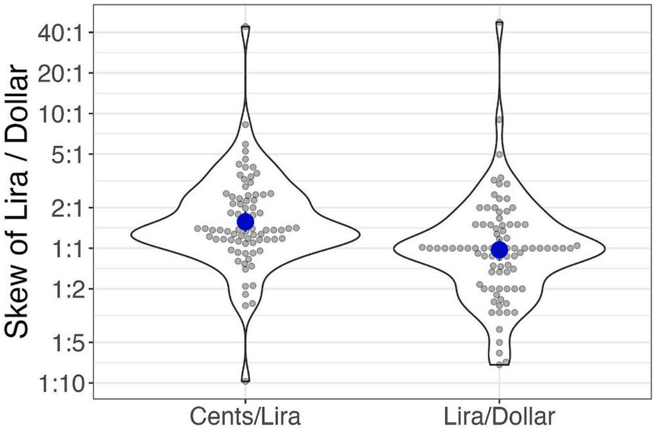


Fig. 6. Prediction Interval Skew Scores in Lira / Dollar by Elicitation Metric. The outcome is the skew score with the labels exponentiated and transformed into ratios for interpretability. Error bars represent 95 % confidence intervals for condition means.

symmetry outcomes were highly skewed, thus we applied a log transformation. For ease of interpretation, we exponentiated the condition average for the outcome statistics.

6.4. *Impacts on interval symmetry*

Symmetry was operationalized as a ratio between distances of the central estimate to the edges: log((90th percentile – average) / (average - 10th percentile)). We refer to this measure as the skew score. Consistent with our predictions, intervals left in their elicited Lira/US dollar metric had lower skew scores than estimates transformed from US cents/ Lira (geometric means: $M_{cents/lira} = 1.6$, $M_{lira/dollar} = 1.0$, $t(160) = -3.67$, $p < .001$, 95 % CI of difference in skew score = [−0.37, −0.11], Cohen's d = 0.58; Fig. 6). Equivalently, intervals were more positively skewed when transformed from their elicited metric relative to intervals where the elicitation and outcome metrics are the same. Note that seven additional participants were excluded because they gave estimates where the lower bound was equal to the average or the average was equal to the upper bound, thus we cannot calculate the outcome statistic for them in both metrics.

7. Experiment 3c (gas prices)

7.1. *Participants*

398 participants (160 women, median age = 37) recruited from AMT completed this study. Participants were paid $1.50 with a median response time of about 7 minutes, and an hourly rate of $13.42/h. This sample size was chosen to have 80 % power to detect a moderate effect size, Cohen's d of around 0.5.

7.2. *Procedure*

Participants were given a summary of the main task and a multiple-choice question confirming they understood the main study task. Then they completed a filler task in which they calculated miles traveled based on odometer readings.

Next, participants made predictions about gas prices across the US. They were randomly assigned to make predictions in *dollars/gal* or *gallons/$40* for regular gasoline in the US. Participants estimated the 90th percentile and the 10th percentile of the gas prices in the US (i.e., "[At **90%** of gas stations the price of gas is **lower** than $__ per gallon and at **10%** of gas stations the price of gas is **higher** than $_ per gallon./ If someone had $40, at **10%** of gas stations they would be able to buy **fewer** than __ gallons and at **90%** of gas stations they would be able to buy **more** than __ gallons.]"). Unlike all the prior experiments, participants were given the average price in their elicitation metric (dollars/gal or gallons/$40). Participants were asked to imagine a road trip and rate how likely they would be to stop for gas based on prices in either dollars/ gal or gallons/$40. The prices were selected to be symmetric in dollars/ gal, but asymmetric when converted to gallons/$40. Finally, they responded to demographic questions.

7.3. *Exclusions*

As pre-registered, we excluded data from participants for whom any of the three estimates were outside of the range [1.00, 6.66] when converted to dollars/gal (104 participants); these values were selected as they represent a plausible range for gas prices someone could hold given historical rates and geographic variation.[6] This left a final sample size of 294.

To equate across the two metric conditions, we converted our outcome variables for all participants into dollars/gal. Treatment effects on interval symmetry are qualitatively the same and have very similar effect sizes when estimates are converted to gallons/$40 (see supplemental materials). Both the width and symmetry outcomes were highly skewed, thus we applied a log transformation. For ease of interpretation, we exponentiated the condition average for the outcome statistics.

7.4. *Impacts on interval symmetry*

Symmetry was operationalized as a ratio between distances of the central estimate to the edges: log((90th percentile – average) / (average - 10th percentile)). We refer to this measure as the skew score. Consistent with our predictions, intervals left in their elicited $/gallon metric had lower skew scores than for intervals transformed from gallons/$40 (geometric means: $M_{gallons/\$40} = 2.2$, $M_{\$/gallon} = 1.36$, $t(290) = -4.04$, $p < .001$, 95 % CI difference in skew scores = [−0.35, −0.12], Cohen's d = 0.47; Fig. 7). Equivalently, intervals were more positively skewed when in the transformed metric from their elicited metric relative to intervals where the elicitation and outcome metrics are the same.

7.5. *Discussion*

In Experiments 3a–c, we test whether people tend to give prediction intervals that are more symmetric when the elicitation and outcome metrics are the same than when they are different. We see this result for all three experiments, suggesting that people act as though they are assuming relatively symmetric distributions regardless of which metric they are using for estimation. When their estimates are inverted, the intervals become less symmetric because of the symmetry prior combined with this mathematical transformation.

8. General discussion

We examined three important means by which simple unit transformations impact uncertainty: addition of a constant, multiplication by a constant, and inversion of a ratio. Each metric transformation impacts estimated uncertainty: adding a constant increases estimated uncertainty, larger quantity metrics increase uncertainty, and estimates in elicited metrics are more symmetric than inversely transformed estimates. More complex metric relationships likely produce a combination of the findings that we observe and are ripe for examination as the literature matures.

Our work suggests that humans do not represent uncertainty in an absolute manner and then translate this absolute uncertainty onto any

[6] See the pre-registration, Experiment 3c – AsPredicted #73452.pdf, for details about the calculation

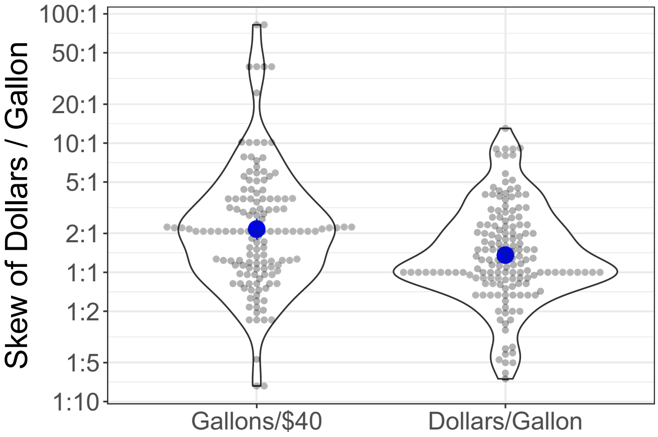


Fig. 7. Prediction Interval Skew Scores for Dollars/Gallon by Elicitation Metric. The outcome is the skew score with the labels exponentiated and transformed into ratios for interpretability. Error bars represent 95 % confidence intervals for condition means.

given metric. Rather, different elicitation methods can result in violations of invariance and lead to meaningful consequences (e.g., Filippin & Crosetto, 2016). Here, we show that that judgments of uncertainty are—at least to some degree—constructed through elicitation rather than simply retrieved and reported (cf. Slovic, 1995). Consistent with the principle of What-You-See-Is-All-There-Is (WYSIATI; Kahneman, 2011), participants seem to reason about uncertainty in the metric that is provided to them. This is reflected most clearly in our findings about symmetry. When asked about Lira/Dollar exchange rates, people seem to start with the assumption that the distribution should be symmetric (Winkler, 1967). When asked about the (inverse) Dollar/Lira exchange rate, people also seem to start with the assumption that the distribution should be symmetric. But mathematically, both cannot be true.

Our results also highlight the importance of prior beliefs in judgments about numerical distributions: using coefficient of variation (σ/μ) to judge variability and assuming symmetry. When one of two metrics has a higher mean, because there is a constant difference between them, someone using the coefficient of variation would expect greater uncertainty in the higher mean metric to maintain the same variability perceptions. If a person uses the coefficient of variation, rather than standard deviation, as their statistic for judging variability, then changes in the mean impact their variability perceptions, even if this change is unwarranted. When metrics are inversely related, at most, one can be symmetric, while the other must be asymmetric. If someone always assumes the uncertainty is approximately symmetric, then switching the elicitation metric by inverting it will impact the skew of their uncertainty, creating an inconsistency in beliefs.

For all three effects, the effect size should depend on the magnitude of the manipulation. Given that variance estimates scale with magnitude, we would expect that as the ratio of the metric magnitudes go from 1.5:1 to 2:1 to 3:1, the inconsistencies in uncertainty estimation would increase, with the more numerate metric having larger and larger uncertainty estimates relative to the less numerate metric. This is because people use the magnitude of the metric to scale their perceptions of uncertainty and larger differences in the magnitudes of the metrics translate to larger differences in perceived uncertainty. A very similar prediction is made for unit insensitivity. When the difference between the units is a factor of 10, we expect that the less numerate metric will result in larger estimates of uncertainty, because people do not fully adjust their estimates by a factor of 10. If the difference between units is a factor of 50, we expect the difference in uncertainty to be even larger, with the less numerate metric resulting in even greater uncertainty estimates than the more numerate metric because people will insufficiently adjust their estimates even more by the larger factor, 50 vs 10. We test this in experiment 2 and find suggestive evidence that larger differences in metric magnitudes increase the impacts on elicited uncertainty. For small differences in metrics (e.g., suppose 1.5 Australian dollars equals 1 US dollar), the effects should be attenuated and may no longer be practically significant.

There is the potential that these biases may be mitigated by broadening decision frames (Larrick, 2004). In this case, people may simply need to consider the alternative metric to form a belief about the focal construct which is consistent for elicitations between the two metrics. For example, explicitly encouraging people to confront both metrics may reduce the impact of metric. Given the strength of numerical processing biases (Thomas & Morwitz, 2009), a decision aid may be needed to facilitate the realignment between estimates in the two metrics.

8.1. *Constraints on generality*

All participants were recruited from internet-based convenience samples. The tasks were not incentivized, were somewhat artificial judgments, and participants were not selected based on their expertise in the tested contexts. Experts who regularly deal with multiple metrics may not show similar inconsistencies in judging uncertainty. We present multiple paradigms for the default symmetry assumption, but only a single paradigm testing that uncertainty scales with magnitude and unit insensitivity. The evidence underlying the uncertainty scaling with magnitude prediction covers a range of contexts (Weber et al., 2004) with similar variation in the evidence for unit insensitivity (Pandelaere et al., 2011; Raghubir & Srivastava, 2002). We expect the two effects to generalize to other contexts even though our experiments do not provide direct evidence of this proposition. Furthermore, our study designs primarily assess only a single trial per participant. However, Experiment 2 has participants make two assessments, suggesting that the effects documented in our investigation should extend to naturalistic conditions under which people make multiple judgments.

CRediT authorship contribution statement

David Zimmerman: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Stephen A. Spiller:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Nicholas Reinholtz:** Writing – review & editing, Methodology, Conceptualization. **Sam J. Maglio:** Writing – review & editing, Methodology, Conceptualization.

Funding

This work was supported by the Morrison Center at UCLA Anderson School of Management.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106277.

Data availability

All data, code, and stimulus files are available at https://researchbox.org/490.

References

Beach, L. R., & Scopp, T. S. (1968). Intuitive statistical inferences about variances. *Organizational Behavior and Human Performance, 3*(2), 109–123. https://doi.org/10.1016/0030-5073(68)90001-9

Burson, K. A., Larrick, R. P., & Lynch, J. G. (2009). Six of one, half dozen of the other. *Psychological Science, 20*(9), 1074–1078.

Filippin, A., & Crosetto, P. (2016). A reconsideration of gender differences in risk attitudes. *Management Science, 62*(11), 3138–3160.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(2), 241–256. https://doi.org/10.1037/0278-7393.12.2.241

Hartman, R., Moss, A. J., Jaffe, S. N., Rosenzweig, C., Litman, L., & Robinson, J. (2023). *Introducing connect by CloudResearch: Advancing online participant recruitment in the digital age*. https://doi.org/10.31234/osf.io/ksgyr

Hlavac, M. (2018). *stargazer: Well-formatted regression and summary statistics tables*. Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Hofstätter, P. R. (1939). Über die Schätzung von Gruppeneigenschaften. *Zeitschrift für Psychologie, 145*, 1–44.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association, 70*(350), 271–289. https://doi.org/10.1080/01621459.1975.10479858

Kahneman, D. (2011). *Thinking, fast and slow*. Straus and Giroux: Farrar.

de Langhe, B., & Puntoni, S. (2016). Productivity metrics and consumers' misunderstanding of time savings. *Journal of Marketing Research, 53*(3), 396–406. https://doi.org/10.1509/jmr.13.0229

Larrick, R. P. (2004). Debiasing. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 316–337). Blackwell Publishing Ltd.

Larrick, R. P., & Soll, J. B. (2008). The MPG illusion. *Science, 320*(5883), 1593–1594. https://ccle.ucla.edu/pluginfile.php/2297402/modresource/content/0/LarrickSoll2007.pdf.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology, 73*(4), 498–502. https://doi.org/10.1037/h0024344

Maglio, S. J., & Trope, Y. (2011). Scale and construal: How larger measurement units shrink length estimates and expand mental horizons. *Psychonomic Bulletin & Review, 18*, 165–170. https://doi.org/10.3758/s13423-010-0025-1

Mahr, T. (2024). *Printy (Version 0.0.0.9003)*. R package. GitHub https://github.com/tjmahr/printy.

Pandelaere, M., Briers, B., & Lembregts, C. (2011). How to make a 29% increase look bigger: The unit effect in option comparisons. *Journal of Consumer Research, 38*(2), 308–322. https://doi.org/10.1086/659000

Peer, E. (2010). Exploring the time-saving bias: How drivers misestimate time saved when increasing speed. *Judgment and Decision making, 5*(7), 477–488.

Pelham, B. W., Sumarta, T. T., & Myaskovsky, L. (1994). The easy path from many to much: The numerosity heuristic. *Cognitive Psychology, 26*(2), 103–133. https://doi.org/10.1006/cogp.1994.1004

Raghubir, P., & Srivastava, J. (2002). Effect of face value on product valuation in foreign currencies. *Journal of Consumer Research, 29*(3), 335–347. https://doi.org/10.1086/344430

Reinholtz, N. (2015). *Persistence in consumer search*. Columbia University.

Shen, L., & Urminsky, O. (2012). Making sense of nonsense: The visual salience of units determines sensitivity to magnitude. *Psychological Science, 24*(3), 297–304. https://doi.org/10.1177/0956797612451470

Slovic, P. (1995). The construction of preference. *American Psychologist, 50*(5), 364–371. https://doi.org/10.1037/0003-066x.50.5.364

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 299–314. https://doi.org/10.1037/0278-7393.30.2.299

Thomas, M., & Morwitz, V. (2009). Heuristics in numerical cognition: Implications for pricing. In *Handbook of pricing research in marketing* (pp. 132–149).

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111*(2), 430–445. https://doi.org/10.1037/0033-295x.111.2.430

Wertenbroch, K., Soman, D., & Chattopadhyay, A. (2007). On the perceived value of money: The reference dependence of currency numerosity effects. *Journal of Consumer Research, 34*(1), 1–10. https://doi.org/10.1086/513041

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association, 62*(319), 776–800. https://doi.org/10.1080/01621459.1967.10500894