



Management Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Understanding People's Preferences for Predictions: People Prioritize Being Right over Minimizing How Wrong They Are in Expectation

Berkeley J. Dietvorst

To cite this article:

Berkeley J. Dietvorst (2025) Understanding People's Preferences for Predictions: People Prioritize Being Right over Minimizing How Wrong They Are in Expectation. *Management Science*

Published online in Articles in Advance 19 Dec 2025

. <https://doi.org/10.1287/mnsc.2024.07257>

This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "*Management Science*. Copyright © 2025 The Author(s). <https://doi.org/10.1287/mnsc.2024.07257>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>."

Copyright © 2025 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Understanding People's Preferences for Predictions: People Prioritize Being Right over Minimizing How Wrong They Are in Expectation

Berkeley J. Dietvorst^a

^aThe University of Chicago Booth School of Business, Chicago, Illinois 60637

Contact: berkeley.dietvorst@chicagobooth.edu,  <https://orcid.org/0000-0002-8139-8983> (BJD)

Received: August 12, 2024

Revised: March 2, 2025; July 23, 2025

Accepted: July 29, 2025


Published Online in Articles in Advance:
December 19, 2025

<https://doi.org/10.1287/mnsc.2024.07257>

Copyright: © 2025 The Author(s)

Abstract. This work explores the preferences that laypeople exhibit when making and evaluating predictions in the form of point estimates (e.g., the high temperature will be 66°). I propose that people typically have diminishing sensitivity to prediction error: the absolute difference between a prediction and a realized outcome. As a result, people often prioritize “being right,” focusing on achieving near perfect predictions and placing less emphasis on the magnitude of errors when errors occur. Across 16 studies using varying methods and stimuli, participants exhibited multiple behaviors consistent with diminishing sensitivity to prediction error: (i) predicting the mode of distributions, (ii) restricting predictions to possible outcomes, (iii) reporting decreasing reactions to increasing marginal units of error, and (iv) preferring predictive models built with diminishing sensitivity to error. This behavior diverges from traditional methods of building predictive models and common interpretations of people’s predictions, which often prioritize avoiding large errors and assume that people are predicting the mean. Ultimately, this work not only highlights the discrepancies between our current practices and people’s preferences for predictions but also calls for a more thorough exploration of human objectives before we build models for them to use or make inferences about their beliefs in light of a decision they made.

History: Accepted by Jack Soll, behavioral economics and decision analysis.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “Management Science. Copyright © 2025 The Author(s). <https://doi.org/10.1287/mnsc.2024.07257>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>.”

Funding: I thank the University of Chicago Booth School of Business for financial support.

Supplemental Material: Preregistrations, materials, data, code, and supplements are available at <https://doi.org/10.1287/mnsc.2024.07257> and at ResearchBox at <https://researchbox.org/3130>.

Keywords: point estimate • judgment • decision making • algorithm • forecasting • modeling • algorithm aversion

Introduction

Prediction is a cornerstone of decision making, influencing both individual choices (e.g., investments, bets, and purchases) and organizational strategies (e.g., hiring, pricing, and strategic planning). Crucially for this work, we often give people predictions to inform them about the world. We also elicit predictions from people to learn about their beliefs. Often, these predictions are communicated as point estimates: specific numbers such as 66° for a high temperature instead of an entire distribution of possibilities. This simplification, although useful, relies on a critical process: selecting one value from a broader distribution of potential outcomes. But how do we choose that one point? This selection isn’t arbitrary; it is deeply influenced by how people perceive and respond to different prediction errors (i.e., the difference between a prediction and the realized outcome).

In this paper, I investigate the nature of laypeople’s preferences regarding error and discover that they diverge significantly from the preferences commonly assumed by statisticians, modelers, researchers, and other experts. The results of 16 studies suggest that, when making predictions, people typically choose what they believe to be a likely outcome, focusing on their odds of achieving near perfect predictions and placing less emphasis on the magnitude of errors when errors occur. This insight is crucial because it diverges from the traditional methods of both building models and interpreting people’s predictions, which often prioritize avoiding large errors and implicitly assume that people are predicting the mean, respectively. Ultimately, this work not only highlights the discrepancies between our current practices and people’s preferences for predictions but also calls for a more thorough exploration of human objectives before

we build models for them to use or make inferences about their beliefs in light of a decision they made.

Scope of This Work

This work investigates people's preferences regarding prediction error independent of any material outcomes (e.g., financial gains, purchases, awards, grades), using what I call "pure predictions:" predictions that have no consequences beyond the prediction error itself. Real-world examples of pure predictions include responses to surveys of consumer sentiment, market research studies, and certain psychological paradigms designed to study biased beliefs (see the general discussion for a concrete example). In addition to investigating preferences over different magnitudes of prediction error in pure predictions, this work also tests whether these preferences influence decisions involving material outcomes (e.g., monetary rewards).

I propose that people strive to make and receive what they believe to be good predictions even when material consequences are at stake. This proposal aligns with findings that people sometimes reject predictive algorithms after seeing them make errors even when adopting those algorithms yields greater financial rewards (Dietvorst et al. 2015, Dietvorst and Bharti 2020). This tendency reflects a broader pattern observed in decades of research, which suggests that people do not necessarily adopt the prediction strategy that maximizes the relevant material outcome of interest (see Estes 1950, Dawes 1979, Grove and Meehl 1996, Vulkan 2000, Highhouse 2008, Dietvorst 2025). Consistent with this proposal, Studies 3a–3d and S3e find that people's preferences regarding error influence their decisions even when material rewards are at stake. People's valuation of material rewards undoubtedly plays a major role in their decisions as described in many models of decision making under risk (see Starmer 2000); however, I suggest that their valuation of prediction error also plays a meaningful role even if it is not the sole determinant of their decisions.

Generating Predictions

Generating a prediction in the form of a point estimate represents a fundamental decision-making process under uncertainty (Goodwin 1996). In theory, the decision maker combines their beliefs regarding the likelihood of various outcomes occurring with preferences over different magnitudes of errors to derive the prediction. For example, a decision maker generating a prediction of the number of points scored in a sports match may consider how likely they believe different outcomes to be and how they feel about different magnitudes of accuracy (e.g., being off by two versus four) to determine which point maximizes the expected rewards (or, equivalently, minimizes the expected penalties). This process mirrors the underlying mechanics of many

predictive models built using cost-minimization calculations (e.g., minimizing the sum of squared residuals). In sum, predictions in the form of point estimates (\hat{y}) are a function of the predictor's preferences over different magnitudes of error (i.e., the penalty function, P) and the predictor's distribution of beliefs about the likelihood of various outcomes (B):

$$\hat{y} = f(P, B). \quad (1)$$

If this conception of the prediction process holds for laypeople, then understanding their preferences over different magnitudes of errors is crucial for anyone interpreting their predictions or supplying them with predictions. When it comes to interpreting predictions, consider a researcher trying to understand a person's beliefs given a prediction the person made. Without knowledge of the penalty function, the researcher faces one Equation (1) with two unknowns: the penalty function (P) and the distribution of beliefs (B). For example, when a respondent predicts that inflation will be 5% next year in a survey of consumer sentiment, what does this prediction represent? Does it represent the mean of the respondent's beliefs, the median, the outcome the respondent believes to be most likely (the mode), or something else? Without this knowledge, it is hard to make accurate inferences about the respondent's beliefs.

When it comes to providing predictions, consider trying to supply a recipient with the recipient's preferred point estimate given an expert's distribution of beliefs (or building an algorithm that does so). The predictor needs to understand the recipient's preferences over different magnitudes of error to select the optimal point estimate. Otherwise, the predictor once again faces one Equation (1) with two unknowns: the penalty function (P) and the prediction (\hat{y}). For example, does a sports fan seeking a football prediction (e.g., the Bears will lose by seven) from an expert source want the mean of the expert's beliefs, the mode, or something else? Without this knowledge, it is hard for the expert to provide the fan's preferred prediction. *In sum, accurately interpreting and optimally providing predictions requires understanding the relevant individual's preferences over different magnitudes of error.*

Minimizing Expected Penalties

Penalty functions describe how predictors trade off between different magnitudes of error (e), which I define as the absolute difference between a prediction (\hat{y}) and a realized outcome (y). For example, a predictor's penalty function (P) specifies the penalty the predictor faces for realizing an error of two (e.g., 2^2 , 2 , $2^{1/2}$, $\log(2+1)$...), four (e.g., 4^2 , 4 , $4^{1/2}$, $\log(4+1)$...), and all other potential outcomes. As a result, penalty functions define the optimal prediction(s) given a distribution of beliefs (B). The optimal prediction is the point estimate that results in

the lowest expected penalty (see Equation (2)), calculated by summing the believed likelihood of realizing each possible error ($p(e)$) multiplied by the corresponding penalty for each error ($P(e)$):

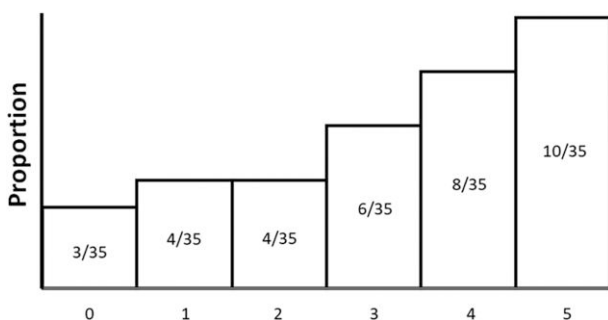
$$\min_{\hat{y}} \sum_e P(e) \cdot p(e). \quad (2)$$

Importantly, given a specific distribution of beliefs, different penalty functions can lead to very different predictions. For example, consider the skewed distribution pictured in Figure 1, representing the whole number outputs of a random number generator (0, 1, 2, 3, 4, or 5). A predictor tasked with generating a point estimate of the next outcome must choose among potential points, each offering a different trade-off between potential errors and aligning with different objectives. When it comes to standard penalty functions, the mean of the distribution (3.2) minimizes mean squared error (MSE), and the median (4) minimizes mean absolute error (MAE) (Lehmann and Casella 2006). However, there are countless potential objectives from which to choose. For instance, the mode (5) maximizes the chances of realizing zero error, the midpoint of the outcome space (2.5) minimizes the maximum possible error, and so on.

Note that, for symmetrical, unimodal distributions—in which the mean, median, and mode coincide—standard penalty functions (e.g., MSE, MAE) all select the midpoint of the distribution. Similarly, decision makers whose penalty functions satisfy the standard assumptions that penalties monotonically increase with the magnitude of error and that all errors of equal magnitude receive equal penalties would also predict the midpoint of any symmetrical, unimodal distribution. However, after relaxing these standard assumptions, every point in a distribution becomes the optimal estimate given some particular penalty function (see the general discussion for more on relaxing these assumptions). For instance, in Figure 1, predicting the minimum (0) minimizes the chances of realizing a positive error.

The penalty functions that statistical models use are well-defined. However, it is not as clear what

Figure 1. Histogram of a Random Number Generator's Whole Number Outputs



objectives laypeople adopt when making predictions or how they penalize different magnitudes of error.

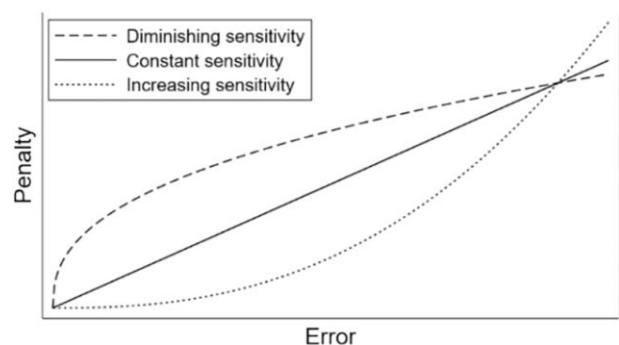
Diminishing Sensitivity to Error

I propose that people become less sensitive to marginal increases in error the larger the initial error is. That is, moving from an error of zero to an error of one results in a larger decrease in satisfaction with a prediction's performance than moving from an error of one to two, moving from one to two likely results in a larger decrease in satisfaction than moving from two to three, and so on. This decreasing reaction to each marginal unit of error creates a penalty function with diminishing sensitivity to error (see the dashed line in Figure 2) with each subsequent unit of error carrying a smaller (or equal) marginal penalty compared with the previous one. Dietvorst and Bharti (2020) provide initial evidence for diminishing sensitivity to error and find that it reduces algorithm adoption in more uncertain prediction domains. However, they do not examine how diminishing sensitivity influences people's predictions or their preferences across different predictive models within a domain: the subjects of the present work.

Reasons for Diminishing Sensitivity to Error

There are many reasons why people may exhibit diminishing sensitivity to error, some of which are perceptual in nature. Research on sensory perception in domains such as vision and touch finds that people often exhibit diminishing marginal reactions to the same incremental change in a stimulus as its initial magnitude increases (Weber 1834, Fechner 1860/1966, Stevens 1957). In other words, the larger the starting value of a stimulus, the greater the change required for a person to notice a difference. Further, some work suggests that this pattern also characterizes people's numerical representations (see Siegler and Opfer 2003). Applied to prediction error, this suggests that people may easily distinguish between small errors (e.g., off

Figure 2. Examples of Penalty Functions Exhibiting Diminishing (Dashed Line), Constant (Solid Line), and Increasing (Dotted Line) Sensitivity to Error



by zero versus off by two), but the same difference may feel smaller—or even imperceptible—at larger error magnitudes (e.g., off by 20 versus off by 22). In short, diminishing sensitivity to error may stem from perceptual differences in how people process errors of varying magnitudes.

People's memory for prediction performance may also drive their diminishing sensitivity to error. For instance, people may make intuitive judgments about an entity's past prediction performance using gist memories of the performances they have witnessed (Brainerd and Gordon 1994, Reyna 2012). These gist representations may be encoded, stored, and recalled in such a way that better distinguishes between smaller errors. For example, if a decision maker encodes errors into categories such as right, almost right, and wrong, this could effectively create diminishing sensitivity to larger magnitudes of prediction error (also see work on efficient coding, e.g., Frydman and Jin 2022). At the extreme, people could even binarize prediction performance, encoding performance simply as either right or wrong (Fisher and Keil 2018). This is all to say that, when people judge prediction performance from memory, their tendency to represent continuous data coarsely could either cause or compound their diminishing sensitivity to error.

Beyond perception and memory, there are many reasons why people may feel that the differences among small magnitudes of error are more consequential than the differences between larger errors. Social comparison may play a key role (see Festinger 1954) as individuals often prioritize outperforming others. Specifically, individuals may care about their rank relative to others in addition to the magnitude of their error (Chun and Larrick 2022). As a result, they may strongly differentiate between small errors, which are likely to determine who performs best, but downplay distinctions among large errors, which are all unlikely to achieve a high rank. Additionally, some individuals may be motivated by demonstrating mastery in a domain (White 1959) and feel that only near perfect predictions achieve this goal. Both of these accounts could help explain why the differences between large errors may feel relatively inconsequential and lead to diminishing sensitivity to larger errors.

Beyond internal motivations, people may develop diminishing sensitivity to error by internalizing external performance standards (see Deci and Ryan 2000). For example, people may frequently encounter winner-takes-all incentives, by which only top performers are rewarded, or see performance standards that disproportionately reward near perfect answers (see Hogarth et al. 1991). Over time, people may internalize these standards and apply them to predictions, defining good prediction performance as achieving perfect predictions (see Einhorn 1986, Gao and Corter 2015), and treating

different magnitudes of large errors as effectively the same. Conversational norms may reinforce this tendency (see Grice 1975); when asked to make a prediction, people may feel that they are expected to be right rather than avoid large errors.

Rather than being driven by a single cause, I propose that people's diminishing sensitivity to error is likely influenced by multiple forces, which may operate independently or together in any given context. Importantly, regardless of the specific cause, this general pattern is provocative because it contrasts sharply with the assumptions typically embedded in predictive models and used to interpret people's predictions. As described next, the empirical evidence in this paper tests various implications of diminishing sensitivity using different paradigms that examine distinct participant behaviors such as making predictions, choosing between predictive models, directly evaluating errors of different magnitudes, and rating potential predictions. Given the diversity of these paradigms, it is unlikely that a single mechanism explains all participant behavior across all paradigms. Because the goal of this paper is to conceptualize diminishing sensitivity to error broadly and test its generality across contexts, I do not conduct deep examinations of the underlying mechanisms within any single paradigm. Instead, in the general discussion, I outline which studies align most closely with each proposed driver of diminishing sensitivity and suggest directions for future research.

Implications of Diminishing Sensitivity to Error

Diminishing sensitivity to increasing outcomes is a characteristic of many value functions, such as diminishing marginal utility of wealth (Arrow 1971) and diminishing sensitivity to increasing changes in wealth (Kahneman and Tversky 1979), but it has unique implications in the domain of error because of this domain's unique properties. Specifically, the domain of error has a strict upper bound on performance at zero error; the most accurate prediction possible is one that results in zero error with larger errors corresponding to poorer performance. Diminishing sensitivity from this strict upper bound of zero error leads predictors to be risk-seeking over error; they pursue the best outcomes (e.g., a prediction that is likely to be right) even at the cost of higher expected error (i.e., higher MAE). *Put plainly, people with diminishing sensitivity to error prefer increasing their chances of being right over reducing how wrong they are in expectation.*

It is worth highlighting one special case of diminishing sensitivity as it may represent a qualitatively distinct prediction process. Under what I term "maximally diminishing sensitivity," people only care about the frequency of perfect predictions. For example, using the penalty function $P(e) = e^\alpha$,¹ a decision

maker exhibits diminishing sensitivity to absolute error (e) for values of α between zero and one ($0 < \alpha < 1$). As α approaches zero, the penalty function converges to a binary form: a perfect prediction incurs a penalty of zero, and any nonzero error incurs a penalty of one. This represents maximally diminishing sensitivity, the most extreme degree of diminishing sensitivity to error, as a decision maker with an α approaching zero effectively distinguishes only between exactly right and not exactly right. Consequently, the goal becomes maximizing the frequency of perfect predictions without regard for the magnitude of any errors, allowing the decision maker to predict by simply selecting the single most likely outcome.

The notion that people make and evaluate predictions with diminishing sensitivity to error leads to four hypotheses that I describe below and test empirically.

Location of Point Estimates. People's preferences over error determine which point they pick as a prediction from a distribution of beliefs. Using the skewed distribution in Figure 1 as an example, a decision maker with constant sensitivity to error would select the median (4). Such a decision maker faces an equal penalty for each marginal unit of error (see the solid line in Figure 2), and the median minimizes the sum of absolute deviations from all outcomes (Lehmann and Casella 2006). A decision maker with increasing sensitivity to error would select a point to the left of the median (4) toward the midpoint of the outcome space (2.5). Decision makers with increasing sensitivity to error face larger marginal penalties for larger errors (see the dotted line in Figure 2) and, thus, prioritize avoiding estimates that may lead to large errors. For example, a decision maker minimizing squared error would select the mean (3.2) as the mean minimizes the sum of squared deviations from all outcomes (Lehmann and Casella 2006).

Unlike constant and increasing sensitivity, diminishing sensitivity to error leads people to generate point estimates representing what they believe to be a likely outcome. Given the distribution of beliefs in Figure 1, a decision maker with diminishing sensitivity to error would select a point equal to or greater than the median (4). Decision makers with diminishing sensitivity to error face relatively smaller marginal penalties for larger errors (see the dashed line in Figure 2) and, thus, seek estimates that are more likely to generate near perfect predictions even if those estimates also have the potential to generate large errors. For example, using the penalty function $P(e) = e^\alpha$, a decision maker predicts the mode (5) if α is between $\sim 2/9$ and zero.

In sum, decision makers with diminishing sensitivity to error prioritize their chances of being right over how wrong they are in expectation, and their

predictions represent outcomes they believe to be likely as a result. I test Hypothesis 1 in Studies 1a–1d and S1f–S1i.

Hypothesis 1. *People tend to place point estimate predictions near the modes of distributions.*

Predicting Possible Outcomes. An additional implication of diminishing sensitivity is that it leads decision makers to predict possible outcomes, such as producing whole number estimates when possible outcomes are exclusively whole numbers. In fact, when there is a discrete set of possible outcomes (e.g., scores in a football game) and a predictor has a strictly concave penalty function (i.e., diminishing sensitivity to error; see Figure 2), the estimate that minimizes the sum of their expected penalties has to be one of the possible outcomes (see Supplement 1, alongside the article in the supplemental materials tab on the INFORMS site and ResearchBox, for a proof).² In lay terms, this is because only predictions of possible outcomes can realize exactly zero error, thus evading the initial increment of error that carries the steepest penalty. For example, returning to Figure 1, a decision maker with diminishing sensitivity to error would select either the median (4) or the mode (5), both whole numbers; no degree of diminishing sensitivity would lead a decision maker to predict a decimal. The implication is that decision makers with diminishing sensitivity to error restrict their predictions to possible outcomes. I test Hypothesis 2 in Study 1e and supplemental analyses of Studies 1a, 1c, and 1d.

Hypothesis 2. *People tend to predict possible outcomes.*

This feature of diminishing sensitivity leads decision makers (and models) with diminishing sensitivity to behave markedly differently when selecting point estimates compared with those with increasing sensitivity. Returning to the example above, as increasing sensitivity intensifies (i.e., α increases from one toward ∞), the optimal point estimate gradually shifts from the median (4) toward the midpoint of the outcome space (2.5). For instance, the optimal estimates for $\alpha = 1.5$, $\alpha = 2$, and $\alpha = 3$ are 3.32, 3.20, and 3.01, respectively. In contrast, as diminishing sensitivity intensifies (α decreases from one toward zero), the optimal point estimate remains stable at 4 until α reaches approximately $2/9$, at which point it abruptly jumps to 5 and then remains stable. In summary, whereas intensifying increasing sensitivity causes a continuous drift in the location of a decision maker's optimal estimate, intensifying diminishing sensitivity leaves the optimal estimate static until a threshold is reached, resulting in a sudden jump from one possible outcome to another.

This behavior has implications for how we should interpret decision makers' point estimates. For

decision makers with increasing sensitivity ($\alpha > 1$), extreme predictions—such as the midpoint of the outcome space (2.5)—indicate maximally increasing sensitivity (α approaching ∞). In contrast, for decision makers with diminishing sensitivity ($1 > \alpha > 0$), extreme predictions, such as the mode (5), do not necessarily indicate maximally diminishing sensitivity (α approaching zero). In other words, even individuals with only moderate diminishing sensitivity may frequently predict the mode.

Evaluations of Errors. People's preferences over error should affect their reported satisfaction with prediction errors of different magnitudes. Specifically, people with diminishing sensitivity to error feel that each unit of error carries a marginal penalty smaller than (or equal to) the previous unit of error. For example, such preferences mean that moving from an error of zero to an error of one decreases satisfaction with one's performance more than moving from an error of one to two and so on. I test Hypothesis 3 in Studies 2a and 2b.

Hypothesis 3. *People report decreasing reactions to later marginal units of error.*

Preferences Between Models. People's preferences over error also affect how they prefer models to trade off between residuals when fitting coefficients. Specifically, people with diminishing sensitivity to error prefer models to prioritize their chances of being right over minimizing how wrong they are in expectation; they are relatively insensitive to the magnitude of models' errors when errors occur. Conversely, models are more typically built with increasing sensitivity to error (e.g., MSE) and, as a result, prioritize avoiding large errors. This may represent a fundamental disconnect between the way that we typically build models to perform (avoiding large errors) and how the users of those models want them to perform (seeking right answers). I test Hypothesis 4 in Studies 3a–3d and S3e.

Hypothesis 4. *People prefer models built with diminishing sensitivity to error.*

Methods

I organized the studies into three sets under the umbrellas of Study 1, Study 2, and Study 3. The studies under the Study 1 umbrella (Studies 1a–1e and S1f–S1i) investigate the point estimates that people generate given a distribution of beliefs. The studies under the Study 2 umbrella (Studies 2a and 2b) investigate people's reported reactions to errors of different magnitudes. The studies under the Study 3 umbrella (Studies 3a–3d and S3e) are incentivized experiments investigating people's choice of models built with different penalty functions. This third set of studies is

intended not only to test the implications of people's diminishing sensitivity to error for their adoption of predictive models but also test, in an incentive-compatible and causal way, whether people have diminishing sensitivity to error.

Open Science

For each study, I preregistered the sample size before any data were obtained. I report all exclusions (if any), all manipulations, and all measures. I found no meaningful evidence of differential attrition across conditions in applicable studies (see Supplement 10 alongside the article in the supplemental materials tab on the INFORMS site and ResearchBox). I have posted preregistrations, materials, data, and code for all studies along with supplements alongside the article in the supplemental materials tab on the INFORMS site and on this project's ResearchBox page: <https://researchbox.org/3130>.

Studies 1a–1d

In Studies 1a–1d, participants learned about a distribution of outcomes before predicting the next observation. Whereas each study applied the same general design, they differed in scenarios and distributions. The aim was to identify where participants positioned a point estimate within the distribution, shedding light on their approach to balancing different magnitudes of error. To this end, I designed the distributions such that people with diminishing, increasing, and constant sensitivity to error would make different estimates. However, it is important to note that in certain cases—such as symmetrical, unimodal distributions—diminishing, increasing, and constant sensitivity to absolute error lead to the same prediction as the mean, median, mode, and midpoint of the outcome space all coincide.

Participants. I preregistered that I would recruit 250 participants from MTurk using CloudResearch approved participants for each of Studies 1a through 1d. In Studies 1a through 1d, respectively, 260, 265, 259, and 260 MTurk workers responded to the surveys, 250, 250, 251, and 250 of whom passed the attention check and completed the last forced response questions in the study (the comprehension checks). The average ages in the final samples were 38, 40, 45, and 40, and 39%, 39%, 43%, and 41% were females, respectively.

Design. After giving consent and passing an attention check, participants were asked to pay close attention and read: “You can earn a \$0.25 bonus for correctly answering two questions (clearly labeled at the end of the survey) about the information that you learn.” This incentive was designed to encourage attentiveness without influencing participants' predictions. For example,

scoring participants' estimates with a linear incentive might bias them toward the median, rewarding only perfect estimates might bias them toward the mode, etc. Thus, I intentionally did not incentivize participants' predictions as the goal of these studies was to learn where people place their estimates based on their own preferences without influence of external incentives.

Subsequently, participants were introduced to the scenario, which differed between the studies (see descriptions in the next section), followed by a textual and graphical summary of the distribution. They were then asked to predict the next outcome's value by generating a point estimate, having access to all previously presented information. In designing the studies, I opted for participants to predict the single next outcome rather than predicting multiple outcomes or setting a policy. This decision was informed by the observation that laypeople typically engage in predictions with a focus on immediate, singular events, such as predicting how long a specific drive will take, the winner of an upcoming sports match, or how much they will like a particular product. I believe that it is less common for laypeople to formulate policies for future decisions (e.g., establishing a general approach for all future driving time estimations) or to make a series of estimates simultaneously (e.g., estimating the duration of each of their next 10 commutes). Thus, this single-estimate approach is designed to closely mirror how laypeople typically make predictions with the goal of enhancing external validity.

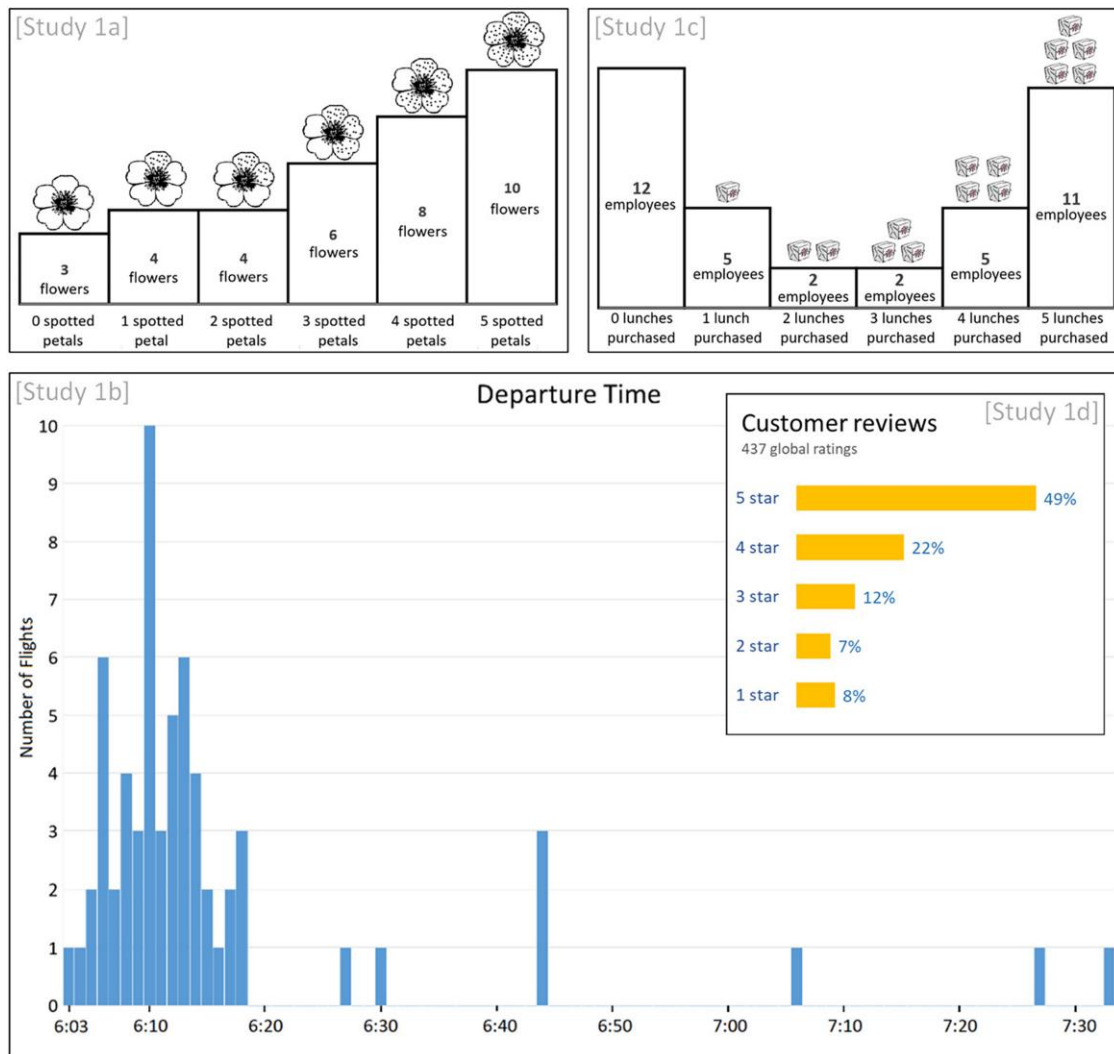
After making their estimate, participants estimated the distribution's mean, again with full information access. The study concluded with a two-question comprehension check related to the scenario (with a \$0.25 bonus for two correct responses) and demographic data collection (age, sex, and education).

Scenarios. In Study 1a, participants read a scenario involving blooming flowers described as follows: "A small garden just started growing a unique type of flower. This type of flower has 5 petals, and can have spots on all, none, or some of its petals. So far, 35 of these flowers have bloomed in the garden." This scenario was selected because of its constrained set of discrete outcomes (0–5) and the assumption that participants would not bring much outside knowledge to the prediction task. Participants were presented with a skewed distribution (see Figure 3), in which the mean (3.2) median (4), and mode (5) were spread out, and thus, participants' treatment of error (increasing, constant, or diminishing) was identifiable. For the estimate, participants read: "The next flower is about to bloom, and you are tasked with predicting the number of spotted petals that it will have. How many spotted petals do you predict the next flower will have? (please enter a number between 0.00 and 5.00)."

In Study 1b, participants read a scenario involving a flight's departure time: "You will see the departure times for the last 63 instances of a specific flight (#792) from O'Hare International Airport to LaGuardia Airport. This flight occurs once per day. It is always scheduled for the same time, but it can depart early or late given each day's particular circumstances. After learning the departure times for the previous 63 days, you will predict the departure time of the next instance of this flight (on day number 64)." This scenario was selected to test whether the results of Study 1a replicated in a domain with a wide range of outcomes, a few dramatic outliers, and a mode to the left of the distribution. Participants were presented with a skewed distribution of departure times (see Figure 3) with a mean equal to 6:18.3, a median equal to 6:13, and a mode equal to 6:10. The data were based on real data for flight UA792 that was altered slightly to produce a round mode and remove a few extreme outcomes. For the estimate, participants read: "You are tasked with predicting the departure time of the next instance of this flight (on day number 64). When do you predict the flight will depart on day number 64? (please enter a number 6–7 in the 'hours' box and 0–59 in the 'minutes' box)."

In Study 1c, participants read a scenario involving lunch purchases: "An employee at a small office building is interested in how often their colleagues purchase lunch (instead of bringing it from home) during the work week. The employee surveys their colleagues asking how many times they purchase lunch in a typical 5-day work week. So far, 37 of their colleagues have responded." The purpose of this study was to test whether the results of previous studies replicate with a bimodal distribution. Participants were presented with a bimodal distribution of lunch purchases (see Figure 3) with a mean equal to 2.43, a median equal to 2, and modes equal to 0 and 5. For the prediction, participants read: "Another employee (#38) is about to submit their response to the survey, and you are tasked with predicting the number of times they report purchasing lunch in a typical week. How many times do you think the employee (#38) will report purchasing lunch in a typical week? (please enter a number between 0.00 and 5.00)." After completing their estimates, participants completed a three-item numeracy scale from Lipkus et al. (2001) that was not included in Studies 1a and 1b. This scale was included to measure participants' capability regarding working with numbers and test whether participants who showed more capability to work with numbers were more or less likely to predict near the mode.

In Study 1d, participants read a scenario involving product reviews: "Imagine that you are looking at product ratings on an online shopping platform. Specifically, ratings for the USB charger pictured below. You scroll down to the ratings for the charger and see that 437 people have submitted a rating." Participants

Figure 3. (Color online) Stimuli Used in Studies 1a Through 1d

were presented with a skewed distribution of ratings based on review distributions of products on Amazon (see Figure 3) with a mean equal to ~ 3.91 , a median equal to 4, and a mode equal to 5. For the prediction, participants read: “Your task is to predict the next rating that this charger will get. That is, what will rating #438 be? How many stars do you think the next person (#438) to review this product will give it? (please enter a number between 1.00 and 5.00).” This study also included the three-item numeracy scale from Study 1c. The purpose of this study was to provide another replication and test whether the results regarding the numeracy scale replicate with a skewed distribution.

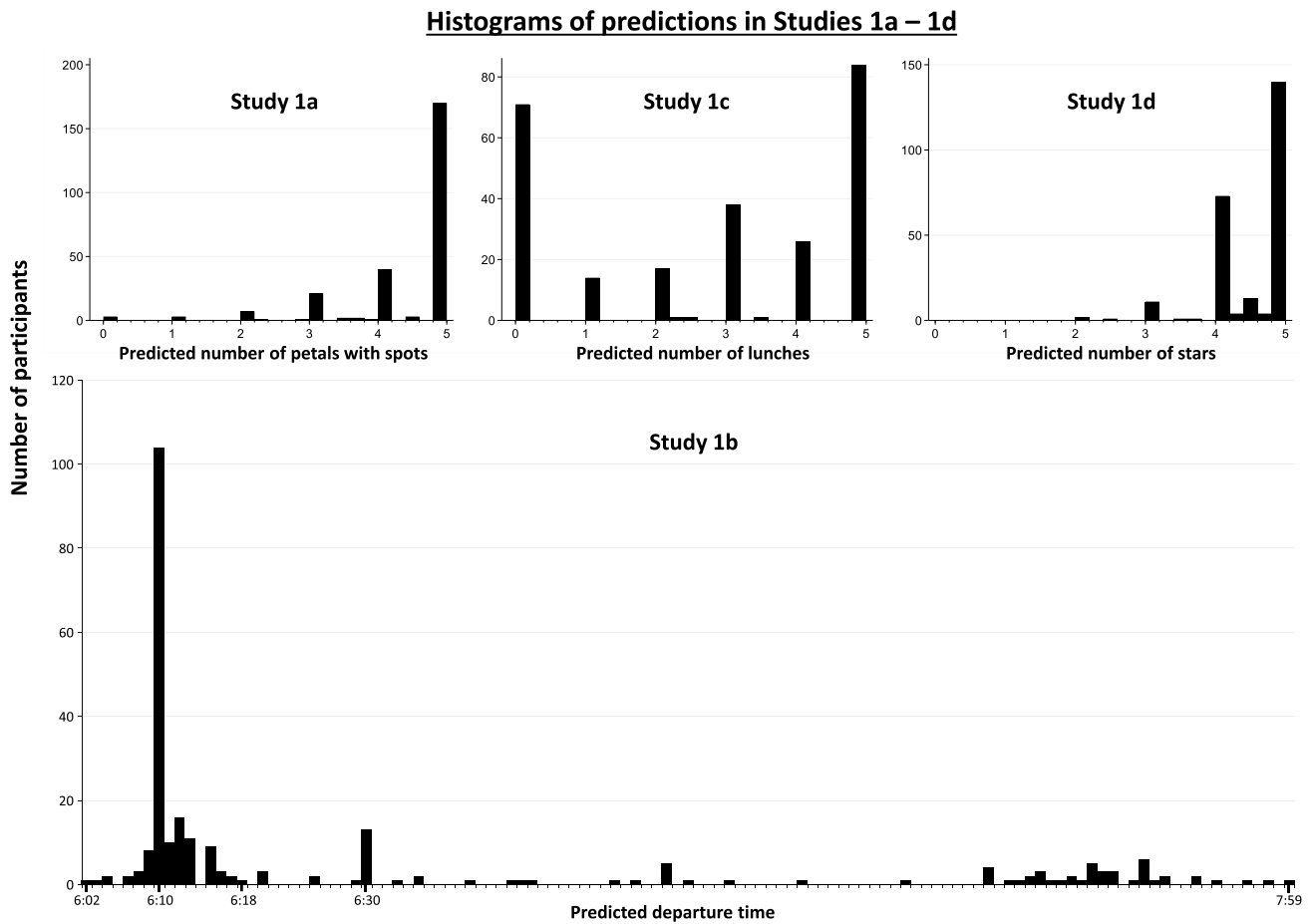
Results

Location of Participants’ Estimates. Across Studies 1a–1d, participants consistently placed their point estimates near the mode(s) of the distribution (see Figure 4), supporting the hypothesis that people generate predictions with diminishing sensitivity to error

(Hypothesis 1). In each study, a plurality—and, in most cases, a majority—of participants made predictions consistent with the mode:³ 68.11% predicted within 0.5 of the mode (5) in Study 1a, 41.11% predicted within 0.5 of the mode (6:10) in Study 1b, 28.06% predicted within 0.5 of the first mode (0) and 33.20% predicted within 0.5 of the second mode (5) in Study 1c, and 62.80% predicted within 0.5 of the mode (5) in Study 1d. The vast majority of these participants predicted exactly the mode (see Supplement 2, alongside the article in the supplemental material tab on the INFORMS site and ResearchBox, for detailed results of each study). In contrast, far fewer participants made predictions consistent with other points in the distribution including the mean and median, and Wilcoxon signed-rank tests confirmed that significantly more participants predicted values consistent with the mode than either the mean or median: $z(N \geq 250) \geq 4.35, p < 0.001$.

These findings were robust across different distribution structures, including those with a wider range of

Figure 4. Histograms of Participants' Predictions in Studies 1a–1d



possible outcomes (Study 1b) and bimodal distributions (Study 1c). Importantly, restricting analyses to participants who passed comprehension checks or provided accurate estimates of the mean did not meaningfully alter results (see Supplement 2 alongside the article in the supplemental materials tab on the INFORMS site and ResearchBox), suggesting that participants' preference for the mode was not driven by confusion or inattention. These findings suggest that people make point estimates with diminishing sensitivity to error, prioritizing their odds of being right over minimizing their expected deviation from the right answer.

Numeracy Scores from Studies 1c and 1d. Analyses of participants' numeracy scores in Studies 1c and 1d find that those participants who showed greater ability to work with numbers were *more* likely to predict the mode. Specifically, in Study 1c, 33.33% (5/15) of participants who answered no numeracy questions correctly predicted within 0.5 of a mode, 55.88% (19/34) of participants who answered one numeracy question correctly predicted within 0.5 of a mode, 56.25% (45/80) of participants who answered two numeracy questions correctly predicted within 0.5 of a mode,

and 69.67% (85/122) of participants who answered all three numeracy questions correctly predicted within 0.5 of a mode: odds ratio (OR) = 1.53 (95% confidence interval (CI): 1.15, 2.03), $z = 2.89$, $p = 0.004$.⁴ Similarly, in Study 1d, 28.57% (2/7) of participants who answered no numeracy questions correctly predicted within 0.5 of the mode, 54.05% (20/37) of participants who answered one numeracy question correctly predicted within 0.5 of the mode, 62.77% (59/94) of participants who answered two numeracy questions correctly predicted within 0.5 of the mode, and 67.86% (76/112) of participants who answered all three numeracy questions correctly predicted within 0.5 of the mode: OR = 1.43 (95% CI: 1.04, 1.97), $z = 2.20$, $p = 0.028$. I interpret these results as evidence that participants did not predict the mode because of a lack of understanding.

Supplemental Studies. I find that these results regarding people's predictions are robust to different paradigms in supplemental Studies S1f through S1i. In Studies S1f and S1g, participants once again made a prediction; however, they provided a distribution of beliefs themselves instead of learning about the distribution in the study. In counterbalanced order,

participants both predicted a future outcome on a seven-point scale and used a distribution builder tool (Andre 2016, Hu and Simmons 2025) to indicate their distribution of beliefs regarding that outcome on the same seven-point scale. In both studies, I found that participants were most likely to predict the mode of their distribution of beliefs, consistent with the results of Studies 1a–1d. See Supplements 3 and 4, alongside the article in the supplemental materials tab on the INFORMS site and ResearchBox, for full write-ups.

In Study S1h, participants chose among models' predictions instead of making predictions themselves. The introduction of the survey was the same as Study 1a using the same flower prediction task. However, when it came time to make the prediction, participants chose among three models (model X, model Y, and model Z) with each model randomly assigned to predict one of the mean (3.2), median (4), and mode (5). Participants learned each model's prediction before making their choice. In line with the results of Studies 1a–S1g, 60.80% (152/250) of participants chose the model that predicted the mode, 26.00% (65/250) of participants chose the model that predicted the median, and 13.20% (33/250) of participants chose the model that predicted the mean. These results suggest that participants in Studies 1a through S1g did not simply predict the mode because the mean and median were not salient to them or because calculating the mean or finding the median was effortful. Further, because predictions were randomly assigned to models, this study addresses the descriptive nature of Studies 1a through S1g and finds causal evidence that people prefer to predict the mode. A logistic regression found that participants were significantly more likely to choose model X when model X was randomly assigned to predict the mode: OR = 7.69 (95% CI: 4.21, 14.03), $z = 6.64$, $p < 0.001$. See Supplement 5, alongside the article in the supplemental materials on the INFORMS site and ResearchBox, for a full write-up.

Study S1i examined preferences over prediction error by having participants rate three potential predictions on seven-point scales after viewing a distribution rather than simply selecting their most preferred prediction. The distribution was structured so that only specific rating patterns aligned with increasing or diminishing sensitivity to error. Notably, giving the mode the highest rating was consistent with both increasing and diminishing sensitivity, meaning that, unlike previous studies, the distinction between these sensitivities depended on preferences beyond simply favoring the mode. The results found that 50.40% (127/252) of participants provided rating patterns consistent only with diminishing sensitivity, 30.56% (77/252) provided patterns consistent with both increasing and diminishing sensitivity, 10.32% (26/252) provided patterns consistent only with increasing sensitivity, and 8.73% (22/252) provided patterns

consistent with neither. Overall, significantly more participants exhibited patterns consistent with diminishing (80.95%) than increasing (40.87%) sensitivity to error ($z = -8.17$, $p < 0.001$). These findings suggest that evidence for diminishing sensitivity to error extends beyond a simple preference for the mode. See Supplement 6, alongside in the supplemental materials on the INFORMS site and ResearchBox, for a full write-up.

Discussion. Studies 1a–1d used a diverse set of stimuli and measures and found evidence suggesting that people make and evaluate predictions with diminishing sensitivity to error with participants placing predictions around the modes of distributions (Hypothesis 1). These results generalized across cover stories and distributions with varying characteristics (skewed versus bimodal, having outliers versus not, wide versus narrow range of outcomes, etc.). The supplements report studies (Studies S1f–S1i) finding that these results generalize across different experimental designs. The results of these eight studies suggest that people's predictions often represent outcomes they consider to be likely rather than, for instance, the mean of their distribution of beliefs.

The very high percentage of participants predicting the mode in these studies might give the impression that they are exclusively focused on maximizing the frequency of perfect predictions, that is, exhibiting maximally diminishing sensitivity. However, this interpretation is misleading as predicting the mode is also consistent with less extreme diminishing sensitivity in all of these studies. For example, a decision maker with the penalty function $P(e) = e^\alpha$ would predict the mode if α is below 0.224, 0.800, 0.585, and ~ 0.943 in Studies 1a–1d, respectively. Thus, a participant predicting the mode in these studies could have either maximally or moderately diminishing sensitivity. Studies 2a, 2b, and 3c are better designed to evaluate whether people often exhibit maximally diminishing sensitivity.

Study 1e

Study 1e tests whether participants restrict their predictions to possible outcomes (Hypothesis 2), an additional implication of diminishing sensitivity to error. Preliminary support for this hypothesis can already be seen in Studies 1a, 1c, and 1d because participants were significantly more likely to enter whole numbers when making predictions (for which realized outcomes were always whole numbers) than when estimating distribution means (which could result in decimals; see analyses in Supplement 7 alongside in the supplemental materials on the INFORMS site and ResearchBox). However, alternative explanations exist; participants may have favored whole numbers specifically rather than simply restricting their predictions to possible outcomes, or they may have chosen whole numbers because they believed it was expected of them. Study 1e was

designed to test Hypothesis 2 in a way that is robust to these alternative explanations.

Participants. I preregistered that I would recruit 500 participants from MTurk using CloudResearch approved participants. Five hundred twenty-nine MTurk workers responded to the survey, 497 of whom passed the attention check and completed the last forced response questions in the study (the comprehension checks). The sample averaged 45 years of age and was 45% female.

Scenario and Design. Participants read a restaurant review scenario: “A restaurant reviewer recently started a blog in which they review restaurants selected by readers and rate those restaurants using a 5-star system. So far, the reviewer has completed 27 reviews.” The design of Study 1e was similar to Study 1a but with three key differences. First, instead of entering their prediction in a text box, participants selected from 11 possible predictions, ranging from zero to five stars in 0.5-star increments (e.g., 5 stars, 4.5 stars, 4 stars, 3.5 stars ...). This response format was intended to make all possible predictions equally salient and to clearly indicate that both whole-number and decimal predictions were valid responses. Second, unlike in Study 1a, participants did not estimate the average of the distribution.

Third, participants were randomly assigned to one of two between-subjects conditions, in which all of the reviewer’s past ratings were either whole numbers or decimals (see Figure 5). The distributions were identical except for a 0.5 shift. This manipulation was designed to test whether participants specifically favored predictions of possible outcomes (as predicted by Hypothesis 2) rather than simply preferring whole numbers. I hypothesized that participants would be significantly more likely to predict a whole number when the reviewer had exclusively used whole numbers and significantly more likely to predict a decimal when the reviewer had exclusively used decimals. Notably, both distributions were structured such that minimizing squared error and predicting the average would result in a contrary outcome: 3.5 in the whole number condition and 3 in the decimal condition.

Results. In line with Hypothesis 2, the majority of participants predicted a possible outcome. Specifically, 94.88% (241/254) of participants in the whole number condition predicted a whole number, whereas 86.06% (216/251) in the decimal condition predicted a decimal. These results remained stable when restricting the analysis to participants who answered both comprehension check questions correctly (95.00% (209/220) in the whole number condition, 88.29% (196/222) in the decimal condition). A two-sample Wilcoxon rank-sum test confirmed that participants were significantly more likely to predict a decimal when all of the reviewer’s past ratings were decimals: $z = -18.25$, $p < 0.001$. Finally, consistent with Studies 1a–S1i, the majority of participants predicted the mode: 79.53% (202/254) selected 4 in the whole number condition, whereas 74.10% (186/251) selected 3.5 in the decimal condition.

Discussion. The results of Study 1e, along with the supplemental analyses of Studies 1a, 1c, and 1d (see Supplement 7 alongside the article in the supplemental materials tab on the INFORMS site and Research-Box), find strong support for the notion that people tend to restrict their predictions to possible outcomes (Hypothesis 2). This pattern held across different distributions and a design specifically intended to rule out alternative explanations, namely, that participants in Studies 1a, 1c, and 1d simply favored whole numbers (rather than possible outcomes) or believed that whole number predictions were expected of them. These findings highlight a key implication of diminishing sensitivity to error: when predicting discrete outcomes, people may prefer estimates that align with a possible outcome, often the mode.

I propose that this tendency may frequently apply to laypeople as they often encounter effectively discrete outcomes. Many services routinely round continuous data to whole numbers when reporting information, such as temperatures (e.g., 70°) or travel durations (e.g., 25 minutes), a practice especially prevalent in consumer-facing contexts. Similarly, people often categorize continuous variables into a finite set

Figure 5. (Color online) Distributions Used in the Whole Number (Left) and Decimal (Right) Conditions of Study 1e



of outcomes; for instance, a baby's weight is typically recorded to the nearest pound and ounce (e.g., 7 lbs., 8 oz.). As a result, even when true outcomes are technically continuous, people may report and evaluate predictions within a discrete framework, particularly in settings involving laypeople.

Studies 2a and 2b

The plurality of participants in Studies 1a through 5I made predictions near the mode of a distribution, and participants were relatively likely to predict possible outcomes, consistent with diminishing sensitivity to error. In Studies 2a and 2b, I test whether participants report similar diminishing sensitivity to prediction error with a completely different elicitation method. In these studies, participants reported happiness with different prediction outcomes (corresponding to different magnitudes of error) in a prediction task. I hypothesized that participants would report the largest difference between a perfect prediction and an error of one and smaller differences between consecutive errors of larger magnitudes.

Participants. I preregistered that I would recruit 250 participants from MTurk using CloudResearch approved participants for each of Studies 2a and 2b. In Studies 2a and 2b, respectively, 267 and 280 MTurk workers responded to the surveys, 245 and 252 of whom passed the attention check and completed the last forced response questions in the study (the comprehension checks). The average ages in the final samples were 38 and 43, and 34% and 43% were females, respectively.

Design. After giving consent and passing an attention check, participants were asked to pay close attention and read: "You can earn a \$0.25 bonus for correctly answering two questions (clearly labeled at the end of the survey) about the information that you read." Subsequently, participants were introduced to the prediction scenario, which differed between the studies (see the next section). Participants read that they would see six different scenarios with different prediction performances and rate each one. Next, participants saw and rated each of the six scenarios with each representing one possible level of performance (off by zero to off by five) in random order. After making their ratings, the study concluded with a two-question comprehension check related to the prediction scenario (with a \$0.25 bonus for two correct responses) and demographic data collection (age, sex, and education).

Scenarios. In Study 2a, participants read a scenario based on the flower prediction task in Study 1a: "The garden just started growing a unique type of flower. This type of flower has 5 petals, and can have spots on

all, none, or some of its petals. So far over 100 flowers have bloomed, and an equal number have had 0, 1, 2, 3, 4, and 5 spotted petals. In other words, all of the outcomes (0, 1, 2, 3, 4, & 5 spotted petals) are equally likely." Study 2a used a uniform distribution because it was easier to explain to participants, and a skewed distribution (as in Study 1a) was not needed to identify participants' reactions to different magnitudes of error. Next, participants read: "Imagine that the next flower is about to bloom, and you are tasked with predicting the number of spotted petals that it will have. In each of the 6 scenarios below, you will rate how you would feel given different prediction performances." During these scenarios, participants read "Imagine that you predicted the flower will have 5 spotted petals. Imagine that the flower actually has [0, 1, 2, 3, 4, or 5] spotted petal[s]." The scenarios were presented in random order. Participants responded to the question "How happy would you be with your performance?" on a nine-point bipolar scale from "Very Unhappy" to "Very Happy" with a midpoint of "Neutral."

In Study 2b, participants read a scenario about predicting a die roll: "You will read about Jim's predictions of the outcome of a die roll. The die has 6 sides numbered 1, 2, 3, 4, 5, 6. In each of the 6 scenarios, you will rate how you think Jim would feel given different outcomes." During the six scenarios, participants read "Jim bets that the die will land on 1. Imagine that the die lands on [1, 2, 3, 4, 5, or 6]." The scenarios were presented in random order. Participants responded to the question "How happy do you think Jim will be with his performance?" on a nine-point bipolar scale from "Very Unhappy" to "Very Happy" with a midpoint of "Neutral."

Results

Study 2a. In Study 2a, participants' aggregate ratings were consistent with diminishing sensitivity to error (see Figure 6). The rated difference between a perfect estimate and an error of one was significantly larger (2.30) than the difference between any other neighboring levels of error (≤ 0.95): $t(253) \geq 9.00$, $p < 0.001$. I conducted an ordinary least squares (OLS) regression to analyze the difference in ratings between adjacent errors (using a variable coded 1, 2, 3, 4, and 5 to represent the rated differences between errors off by zero versus off by one, off by one versus off by two, and so on) with standard errors clustered by a participant ID. This analysis revealed that, on average, the rating gap between neighboring errors decreased by 0.39 for each one-unit increase in error magnitude: $t(254) = -12.49$, $p < 0.001$.

Next, I analyze participants' individual response patterns by examining whether they reported a step down in happiness for each incremental increase in error. On average, participants reported 2.61 steps down (out of a maximum of five) with only 27.56%

Figure 6. Participants' Ratings in Study 2a



Notes. Error bars represent 95% CIs; labels represent means. The pattern is similar when excluding participants who failed the comprehension check ($n = 193$): 3.21, 0.69, -0.36 , -1.03 , -1.60 , -2.11 .

(70/254) indicating a single step down. The majority (68.5%, 174/254) exhibited multiple steps down,⁵ suggesting more continuous diminishing sensitivity rather than a simple step function. However, among those who indicated only one step down, the vast majority (91.43%, 64/70) placed it at the first unit of prediction error: precisely where maximally diminishing sensitivity predicts the step should occur. Overall, these results suggest that most participants demonstrated continuous diminishing sensitivity to error, and a notable minority exhibited what appeared to be binary thinking about error consistent with maximally diminishing sensitivity.

Study 2b. In Study 2b, participants' aggregate ratings were again consistent with diminishing sensitivity to prediction error (see Figure 7). The rated difference between a perfect prediction and an error of one was significantly larger (5.25) than the difference between any other neighboring levels of error (≤ 0.20): $t(252) \geq 28.11$, $p < 0.001$. I conducted an OLS regression to analyze the difference in ratings between adjacent errors (using a variable coded 1, 2, 3, 4, and 5 to represent the rated differences between errors off by zero versus off by one, off by one versus off by two, and so on) with standard errors clustered by a participant ID. This analysis revealed that, on average, the rating gap between neighboring errors decreased by 1.03 for each one-unit increase in error magnitude: $t(252) = -33.12$, $p < 0.001$.

As opposed to the results of Study 2a (the flower scenario), the diminishing sensitivity in Study 2b (the die roll scenario) appears to be largely driven by the

rated difference between errors of zero and one, consistent with preferences approaching maximally diminishing sensitivity to error.⁶ However, a regression of participants' raw ratings on an error magnitude variable excluding zero (coded as one unit of error, two units, ... five units) finds that participants' ratings did significantly decrease as the magnitude of positive errors increased, $b = -0.16$, $t(252) = -5.59$, $p < 0.001$, consistent with the notion that they did discriminate between different magnitudes of nonzero errors.⁷

Next, I analyze participants' individual response patterns by examining whether they reported a step down in happiness for each incremental increase in error. On average, participants reported 1.82 steps down (out of a maximum of five). Roughly half indicated only one step down (50.59%, 128/253), whereas the other half reported multiple steps down (47.43%, 120/253).⁸ Among those who indicated only one step down, the vast majority (96.88%, 124/128) placed it at the first unit of prediction error: precisely where maximally diminishing sensitivity predicts the step should occur. These findings suggest notable heterogeneity in participants' responses: some exhibited what appeared to be binary thinking about error, whereas others demonstrated more continuous diminishing sensitivity. However, both groups expressed preferences that aligned with diminishing sensitivity, differing in the degree.

Discussion. Studies 2a and 2b found that participants report decreasing reactions to marginal units of error on average (Hypothesis 3) with the first unit of

Figure 7. Participants' Ratings in Study 2b

Notes. Error bars represent 95% CIs; labels represent means. The pattern is similar when excluding participants who failed the comprehension check ($n = 248$): 3.41, -1.96, -2.15, -2.29, -2.46, -2.60.

error eliciting the largest reaction and subsequent units of error eliciting smaller reactions. This evidence of diminishing sensitivity replicated across two studies with different stimuli. However, the degree of diminishing sensitivity differed between scenarios: participants exhibited stronger diminishing sensitivity when predicting the outcome of a die roll. One potential explanation is that the die roll scenario evoked binary performance standards (e.g., a prediction is either exactly right or wrong), which are common in dice games and gambling. In contrast, the flower scenario may not have evoked such binary standards. Overall, these results suggest substantial heterogeneity in people's preferences over error across different prediction contexts and individuals.

Studies 3a–3d

Studies 1a–2b found convergent evidence that people exhibit diminishing sensitivity to prediction error across different methods. Studies 3a–3d and S3e test whether people are more likely to adopt models built with diminishing sensitivity to error using incentivized experiments. In these studies, participants first experienced the performance of two prediction options through a set of practice predictions. Then, they selected on which to rely for an incentivized prediction that determined their bonus. I manipulated the penalty functions of the models across two between-subjects conditions and hypothesized that participants would be more likely to choose models designed with diminishing sensitivity to prediction error. As in Studies

1a–1d, I designed the stimuli in these studies to differentiate between varying degrees of sensitivity to error. However, it is important to note that, in many cases, models with different penalty functions do not produce such divergent predictions.

Studies 3a–3c had participants view two models' predictions alongside realized outcomes across a series of training trials before making an incentivized choice between the models. These studies aimed to present model performance as clearly and simply as possible. In contrast, Studies 3d and S3e provided participants with the full prediction process; participants first observed the three available inputs for a prediction before seeing the models' predictions or the realized outcomes. This approach allowed participants to generate their own predictions rather than solely evaluating model-generated predictions and tested whether the findings from Studies 3a–3c generalized to situations in which participants had richer information about the prediction process.

Participants. I preregistered that I would recruit 500 participants from MTurk using CloudResearch approved participants for each of Studies 3a through 3d. In Studies 3a, 3b, 3c, and 3d, respectively, 520, 542, 564, and 543 MTurk workers responded to the surveys, 505, 502, 504, and 501 of whom passed the attention check and completed the last forced response question in the study (the incentivized choice). The average ages in the final samples were 42, 43, 43, and 42, and 41%, 45%, 49, and 48% were females, respectively.

Studies 3a–3c Design. After giving consent and passing an attention check, participants read that they would learn about two models, model A and model B, that each make estimates for a prediction task. They learned that the study would have a trial stage and a bonus stage. During the trial stage, they would see 9 (Studies 3a and 3b) or 10 (Study 3c) prediction trials. For each trial, they would see three pieces of information: the estimates that model A and model B made along with the actual answer. During the bonus stage, they would choose between relying on either model A's estimate or model B's estimate for one more trial that determined their bonus payment. After the introduction, the survey software randomized which model was labeled model A and model B.

Next, participants completed the trials in random order (see Figure 8 for an example of a trial). After completing the trials, participants proceeded to the bonus stage. Participants were first presented with the study's bonus scheme. On the next page, participants made a binary choice between using model A's estimate and model B's estimate to determine their bonus. After making their choice, one of the trials was randomly reselected, and participants learned their chosen model's performance and their earnings. Participants reported their age, sex, and highest level of education to complete the study.

Study 3a Task. The prediction task in Study 3a is represented in Figure 9. The outcomes follow the linear equation $y = x$ for eight values of x (1, 2, 3, 4, 6, 7, 8, 9), and for the ninth value of x (5), y equals 14. The increasing model, built to minimize squared error, follows the equation $y = 1 + x$. For this model, the reduction in penalties from making an error of eight ($8^2 = 64$ penalty) versus nine ($9^2 = 81$ penalty) outweighs the cost of introducing an error of one for the remaining eight estimates ($8 \times 1^2 = 8$ penalty). In contrast, the diminishing model follows the equation $y = x$. Any model built with diminishing (or constant) sensitivity to error would avoid introducing errors into the eight perfect estimates to proportionally reduce the error of the ninth. I hypothesized that participants would be significantly more likely to choose the diminishing model.

Figure 8. Example of Prediction Task Trial

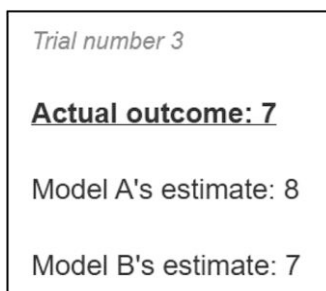
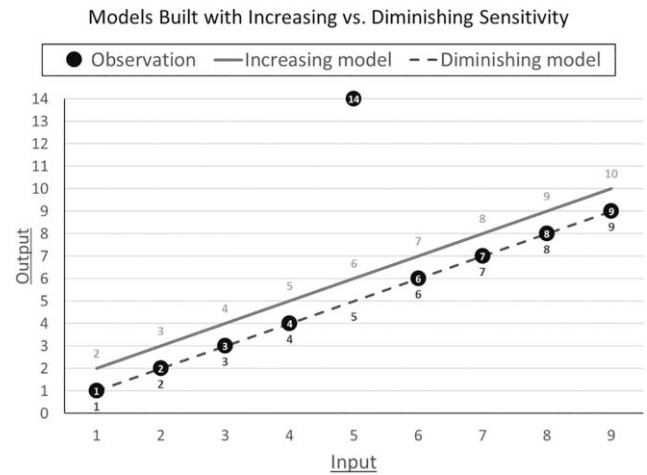


Figure 9. Observations (Black Dots) and Models with Increasing (MSE: Solid Gray Line) and Diminishing (Dotted Black Line) Penalty Functions for the Prediction Task in Study 3a



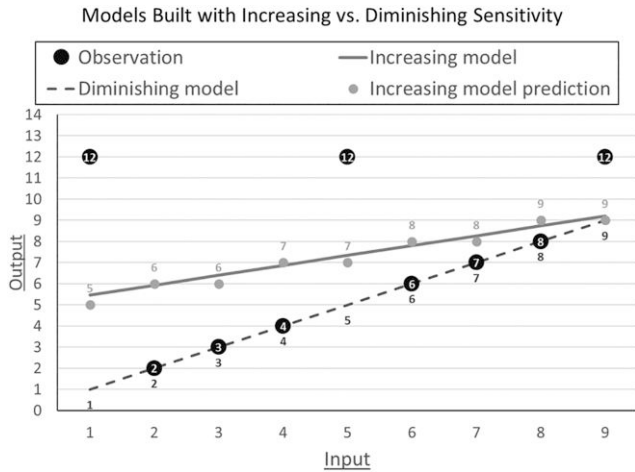
Note. Participants saw this information trial by trial as shown in Figure 8.

The bonus scheme in Study 3a equated the expected value of choosing the increasing and diminishing models so that participants were not biased toward either option. Participants earned \$0.50 for a perfect prediction, \$0.49 for an error of one, \$0.48 for an error of two, \$0.46 for an error of three, \$0.43 for an error of four, \$0.39 for an error of five, \$0.34 for an error of six, \$0.28 for an error of seven, \$0.21 for an error of eight, \$0.13 for an error of nine, \$0.03 for an error of 10, and \$0 for an error of 11 or larger. Given this incentive, the expected value of choosing both the increasing and diminishing model was \$0.458. Notably, this bonus scheme exhibited increasing penalties for larger magnitudes of error.

Study 3a Results. As hypothesized, participants were unlikely to choose the model built with increasing sensitivity to error. Overall, 86.53% (437/505) of participants chose the model built with diminishing sensitivity to error: $z(n = 505) = 16.42, p < 0.001$, using a Wilcoxon signed-rank test versus 50%. A logistic regression of whether participants chose model A (coded one if yes, zero if not) on an indicator of whether model A was the diminishing model (coded one if yes, zero if not) found that participants were substantially more likely to choose model A when it was built with diminishing sensitivity to error: $OR = 41.44$ (95% CI: 24.84, 69.13), $z = 14.26, p < 0.001$.

Study 3b Task. The prediction task in Study 3b is illustrated in Figure 10. This task was designed to test whether the results of Study 3a are robust to the diminishing model making more than one error, the models exhibiting different slopes, and an incentive

Figure 10. Observations (Black Dots) and Models with an Increasing (MSE: Solid Gray Line) and Diminishing (Dotted Black Line) Penalty Functions for the Prediction Task in Study 3b



scheme that favors the increasing model. The data follows the linear equation $y = x$ for six values of x (2, 3, 4, 6, 7, 8), and for the remaining three values of x (1, 5, 9), y equals 12. The model built with increasing sensitivity to error, which minimizes squared error, followed the equation $y = 5 + 0.46x$. This model's predictions were the output of this equation rounded to the nearest whole number. In contrast, the model built with diminishing sensitivity to error resulted in the equation $y = x$. I hypothesized that participants would be more likely to choose the diminishing model.

Participants were incentivized to minimize squared error. Specifically, they earned \$1.21 for a perfect prediction and lost one cent for each unit of squared error that their chosen model realized. For example, this resulted in a \$1.20 bonus for a prediction off by one ($\$1.21 - 1¢ \times 1^2$), a \$1.17 bonus for a prediction off by two ($\$1.21 - 1¢ \times 2^2$), down to a \$0 bonus for a prediction off by 11 ($\$1.21 - 1¢ \times 11^2$), which was the largest error that either model made. This incentive structure was optimized for the increasing model, meaning it yielded higher expected earnings than the diminishing model, and thus, choosing the diminishing model was costly.

Study 3b Results. As hypothesized, participants were again unlikely to choose the model built with increasing sensitivity to error even though it generated larger bonuses in expectation. Overall, 87.25% (438/502) of participants chose the model built with diminishing sensitivity to error: $z(n = 502) = 16.69$, $p < 0.001$ using a Wilcoxon signed-rank test versus 50%. A logistic regression of whether participants chose model A (coded one if yes, zero if not) on an indicator of whether model A was the diminishing model (coded

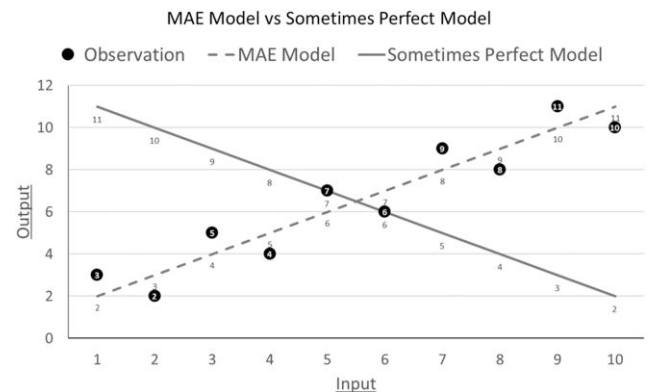
one if yes, zero if not) found that participants were substantially more likely to choose model A when it was built with diminishing sensitivity to error: OR = 50.77 (95% CI: 29.47, 87.47), $z = 14.15$, $p < 0.001$.

Study 3c Design. Studies 3a and 3b found additional evidence that people exhibit diminishing sensitivity to error, but they do not speak to the degree of this diminishing sensitivity. Because any degree of diminishing sensitivity would lead participants to prefer the diminishing model in these studies, it remains unclear whether participants simply preferred models that made more perfect predictions (consistent with maximally diminishing sensitivity) or if they exhibited a more continuous decline in sensitivity to models' increasing errors. Study 3c was designed to address this question.

The data follows the linear equation $y = x$ for five values of x (2, 4, 6, 8, 10), and $y = 2 + x$ for five values of x (1, 3, 5, 7, 9). The "sometimes perfect" model produces two perfect predictions but makes eight relatively large errors (four errors of four and four errors of eight; see Figure 11). In contrast, the "MAE model" is always off by one. This trade-off means that a decision maker would need near maximal diminishing sensitivity to error—placing overwhelming importance on being exactly right—to prefer the sometimes perfect model. For instance, using the penalty function $P(e) = e^\alpha$, this would require $\alpha < 0.128$. On the other hand, a decision maker with more moderate diminishing sensitivity to error would favor the MAE model. I hypothesized that participants would prefer the MAE model as I believe true maximally diminishing sensitivity—by which people focus solely on the frequency of perfect predictions and completely disregard the magnitude of errors—may be rare.

The bonus scheme in Study 3c equated the expected value of choosing the sometimes perfect and MAE models so that participants were not biased toward

Figure 11. Observations (Black Dots), the Sometimes Perfect Model (Solid Gray Line), and MAE Model (Dotted Gray Line) for the Prediction Task in Study 3c



either option. Participants earned \$1 for a perfect prediction, \$0.80 for errors of one to three, \$0.75 for errors of four to eight (the largest error either model made), and smaller bonuses were listed for larger magnitudes of error. Given this incentive, the expected value of choosing both the sometimes perfect and MAE model was \$0.80.

Study 3c Results. As hypothesized, participants were more likely to choose the MAE model than the sometimes perfect model even though the MAE model was never perfect. Overall, 92.06% (464/504) of participants chose the MAE model: $z(n=504) = -18.89$, $p < 0.001$, using a Wilcoxon signed-rank test versus 50%. A logistic regression of whether participants chose model A (coded one if yes, zero if not) on an indicator of whether model A was the sometimes perfect model (coded one if yes, zero if not) found that participants were substantially less likely to choose model A when it was the sometimes perfect model: OR = 0.007 (95% CI: 0.004, 0.014), $z = -14.87$, $p < 0.001$.

Discussion of Studies 3a–3c. Studies 3a–3c found evidence that people prefer models built with diminishing sensitivity to error (Hypothesis 4), supporting the idea that people evaluate prediction performance with diminishing sensitivity. These preferences held across multiple sets of stimuli and remained robust even in Study 3b, in which the incentive structure favored the model with increasing sensitivity to error. Notably, the results of Study 3c suggest that most participants do not simply prefer models that make more perfect predictions but instead exhibit more moderate diminishing sensitivity that distinguishes between smaller and larger errors.

Study 3d Design. Study 3d was designed to test whether the results so far generalize to choices between a participant's own predictions and a model's and situations in which participants have richer information about the prediction process. After giving consent and passing an attention check, participants read that they would learn about a prediction task by making a series of predictions and then have an opportunity to earn a bonus. They read that the task involved predictions of the output of a number machine. They learned that “the specific number machine that you will make judgments about takes in 3 input numbers (between 0 and 10) to generate 1 output number” and that the number machine always uses the same exact process to generate outputs. Next, participants learned that, along with their own estimates, they would also see estimates generated by a statistical model and that the model uses the same information that they would receive.

Participants learned that the survey would have a trial stage and a bonus stage. During the trial stage,

participants saw a set of inputs and then made a prediction for 20 sequential prediction trials. They saw the statistical model's prediction along with the right answer immediately after making each prediction. Then, during the bonus stage, they chose between using their own prediction or the model's prediction in one future trial to determine their bonus. Participants read that the trial stage does not count toward their bonus but is meant to help them learn how to make a choice in the bonus stage.

Next, participants proceeded through the trial stage. Behind the scenes, a set of formulas determined the predictions and performances that participants saw. The three inputs were random integers from 0 to 10. The correct answer was a linear combination of the three input numbers ($0.35a + 0.15b + 0.5c$) plus a random integer (equal to 1 or -1 with equal probability) all rounded to the nearest whole number (see Supplement 8, alongside the article in the supplemental materials tab on the INFORMS site and ResearchBox, for full details).

The predictions and performance of the model differed across the two between-subjects conditions. In the “increasing” condition, the model predicted the correct linear combination of the three inputs, rounded to the nearest whole number, disregarding the random integer. This meant that the model was always off by one (either one too low or one too high given the outcome of the random integer). In the “diminishing” condition, the model predicted the correct linear combination of the three inputs rounded to the nearest whole number and guessed the random integer. This meant that the model was off by zero half of the time (when correctly guessing the random integer) and off by two half of the time (when incorrectly guessing the random integer).

For each of the 20 trials in the trial stage, participants saw one of 24 sets of preconstructed stimuli randomly selected without replacement. Each trial consisted of two pages. On the first page, participants saw the three input numbers and made their own prediction from 0 to 10 (restricted to whole numbers) in a blank text box. Then, on the second page, they saw their prediction, the model's prediction, and the correct answer.

After completing the 20 trials, participants learned how their bonus would be calculated given the performance of their chosen prediction method. They would earn a maximum of \$1, losing \$0.10 for each unit of error in the prediction. That is, they would earn \$1 if the prediction was off by zero, \$0.90 if the prediction was off by one, \$0.80 if the prediction was off by two, and so on. Given this bonus scheme, both models generated the same amount of bonus money in expectation (\$0.90). Participants learned that they would make a prediction themselves no matter which option they chose. Next, participants made a binary choice between using their own prediction or the model's

prediction to determine their bonus. After making their choice, participants completed the final prediction and reported their age, sex, and highest level of education to complete the study.

Study 3d Hypothesis. I hypothesized that participants would be significantly more likely to choose to use the model in the diminishing condition as people with diminishing sensitivity to error would prefer an equal chance of realizing zero and two units of error to realizing one unit of error for certain. Note that choosing not to use the model in either condition was likely costly to participants because it was impossible to earn more than either model in expectation. Realistically, participants' own predictions would almost surely earn less than the model's in both conditions as participants were unlikely to perfectly learn the equation that generated the correct answers.

Study 3d Results. As expected, the model performed significantly better than participants in both conditions. Participants' predictions were off by 1.77 on average in the trial stages, and (as designed) the model's predictions were off by 1.00 on average: $t(500) = 31.19$, $p < 0.001$.⁹ Participants' performance when generating their own predictions was not significantly different between the increasing ($M = 1.74$) and diminishing ($M = 1.80$) conditions: $t(499) = -1.13$, $p = 0.257$. Only 2.20% of participants (11/501) outperformed the model (by chance) during the trial stage. Similarly, the model performed significantly better ($M = 0.96$) than participants ($M = 1.67$) in the single bonus stage ($t(500) = -9.80$, $p < 0.001$) and generated larger bonuses as a result.

Although the model was significantly better at generating bonuses than participants were in both conditions, participants were significantly more likely to use the model in the diminishing condition. Specifically, whereas 36.00% (90/250) of participants chose to use the model in the increasing condition, 51.39% (129/251) of participants chose to use the model in the diminishing condition: $\chi^2(1, n = 501) = 12.06$, $p = 0.001$. This result is consistent with the notion that participants preferred a model built with diminishing sensitivity to error and that more participants were willing to pay a cost in expected bonus to avoid using a model built with increasing sensitivity to error.

Discussion of Study 3d. As in Studies 3a–3c, Study 3d found that people prefer models built with diminishing sensitivity. Study S3e replicates these results with participants making a choice between two models (see Supplement 9 alongside the article in the supplemental materials tab on the INFORMS site and ResearchBox). These studies highlight the potential benefits of designing models to reflect users' preferences and, conversely, the potential costs of implementing standard penalty

functions (e.g., minimizing squared error) without considering users' preferences. For example, most participants in Study 3d chose not to use a model when their only option minimized squared error, leaving money on the table as a result. On the other hand, presenting participants with a model that reflected their apparent diminishing sensitivity to error significantly boosted their willingness to use a model, and increased their earnings as a result. This finding suggests that designing user-centered algorithms around human preferences may enhance algorithm adoption and help individuals to better capitalize on the benefits of algorithmic decision making (see Dawes et al. 1989).

General Discussion

Studies 1–3 found convergent evidence that people generate and evaluate predictions in the form of point estimates with diminishing sensitivity to error. Study 1 found that people place their predictions approaching the modes of distributions and restrict predictions to possible outcomes. Study 2 found that people report decreasing marginal reactions to larger units of error. Study 3 found that people prefer models built with diminishing sensitivity to error over those with more typical penalty functions. The consistency of participants' responses across very different experimental designs suggests that people may approach prediction as a decision under uncertainty; however, they may do so with different preferences over error than statisticians, modelers, researchers, and other experts typically assume.

People's diminishing sensitivity to error is likely to be multiply determined, and the different behaviors exhibited across Studies 1–3 likely reflect the influence of multiple underlying processes. Perceptual differences in how people process errors of varying magnitudes may have played a role in all studies. Particularly in Studies 2 and 3, in which participants directly evaluated errors, magnitudes of larger errors may have been perceived as less distinct, leading participants to prioritize the realization of small errors. However, for this perceptual account regarding errors to explain the results of Study 1's participant-generated predictions, participants would have needed to anticipate the potential errors associated with different possible predictions and then perceive those potential errors with diminishing sensitivity. Whereas this mirrors how many statistical models operate, it may be less likely to reflect participants' decision processes when formulating a prediction given a distribution.

Alternatively, in Study 1, participants may have formulated predictions based on internal motivations to be right or internalized performance standards that disproportionately reward perfect answers. This could have led participants to exhibit diminishing sensitivity

by seeking out predictions that were more likely to be perfect, disregarding the potential for large errors as relatively inconsequential. Additionally, because Study 1 did not provide a concrete performance goal, participants may have inferred that they were expected to provide the most likely outcome rather than one that minimized expected error, driving their predictions toward the mode. However, this mechanism is less likely in Study 3, in which participants chose among models with a clear financial incentive to minimize expected error, and especially in Study 3b, in which they were incentivized to minimize squared error.

Memory may have played a significant role in participants' diminishing sensitivity to error in Study 3, in which they observed multiple prediction trials before selecting a model, requiring them to evaluate performance from memory. Specifically, participants may have encoded, stored, and recalled model performances in a way that distinguished smaller errors, grouping larger errors into broader categories (e.g., right, close, and way off). However, memory should have played a smaller role in Study 1, in which all relevant information was visible on-screen when participants made their predictions. Further, it is unlikely that memory had a strong influence in Study 2 as participants directly evaluated errors of different magnitudes in real time with no clear need for recall.

Overall, diminishing sensitivity to error may be driven by multiple factors with their influence potentially varying across contexts and tasks. Future research could refine these mechanisms, explore how they interact, investigate when each is most influential, and test how they generalize to different tasks and prediction domains. Beyond the clear theoretical contribution, understanding the role of each mechanism may be useful for both building models that represent users' preferences across different domains and accurately interpreting people's predictions across different domains.

Implications for Supplying Laypeople with Predictions

This work has significant implications for prediction generation. Consumers of predictions may evaluate them with diminishing sensitivity to error, finding the most satisfaction when estimates reflect likely outcomes. To maximize consumer satisfaction, prediction suppliers (and their models) should consider prioritizing feasible outcomes (e.g., avoiding decimals when outcomes are inherently whole numbers) and being right rather than avoiding large errors. For example, consumers may prefer a prediction that a football team will win by 3 points instead of 2.48 points because (i) football scores are whole numbers and (ii) three-point margins of victory are much more common than two-point margins of victory (Greer 2020). These

preferences contrast with many standard algorithms—such as those that minimize squared error—which often produce decimal predictions for inherently whole-number outcomes and prioritize reducing large errors even if it means introducing smaller ones.

Importantly, aligning algorithmic predictions with people's preferences for predictions would likely increase the usefulness and adoption of those predictions. People's apparent reluctance to use algorithms that make errors (e.g., Dietvorst et al. 2015) could stem not just from a general distrust of imperfect algorithms but also from opposition to the traditional methods used to balance errors within these algorithms. This mismatch suggests that the dissatisfaction with algorithmic predictions could be reduced by developing models that align more closely with human preferences over error, emphasizing algorithms that better reflect how people naturally assess errors. This shift could not only improve user satisfaction but also enhance the practical adoption of algorithmic decisions in everyday applications (see Dietvorst 2025).

Implications for Interpreting Laypeople's Predictions

These findings also have important implications for those of us who interpret people's predictions for research or other purposes. People's predictions are often consequential. For example, market researchers may ask consumers to make predictions in order to learn about their beliefs, and surveys of consumer sentiment, which often ask participants to make predictions, are used as economic indicators (e.g., Carroll et al. 1994). However, as highlighted in the introduction, accurately deducing people's beliefs from their point estimates requires a clear grasp of their prediction goals. This work indicates that people's point estimates often reflect the outcomes they deem likely rather than outcomes that minimize error in expectation (see also Dimitriadis et al. 2019). Recognizing this tendency can enhance the interpretation of consumers' predictions for research and practical applications.

This work also suggests that some common research practices misinterpret participants' predictions. For example, researchers often elicit predictions from participants to learn about their beliefs and test whether those beliefs are biased. These tests often compare the mean of participants' beliefs to the mean of true outcomes (e.g., using a *t*-test) and interpret a significant difference as evidence of biased beliefs. As just one of many potential examples, Snyder (1978) interpreted students predicting that they will live 10 years longer than average (conditional on their age) as a self-serving biased belief. However, such interpretations assume that participants' predictions represent the mean of their distribution of beliefs and, therefore, implicitly assume that participants generate point

estimates with the goal of minimizing the expectation of squared error. Otherwise, we should not necessarily expect predictions to match the mean of actual outcomes when beliefs are unbiased. For example, considering the generally left-skewed distribution of age at death with a mean of 78.9 years and a mode of 87 years in the United States in 2015 (Fuchs and Eggleston 2018), Snyder's (1978) students could have been accurately estimating the most common age at death (i.e., the mode) instead of inflating the mean age at death.

The present work suggests that people do not typically aim to minimize squared error or predict the mean of a distribution, potentially invalidating the types of comparisons described above. Moreover, laypeople's preferences over error do not necessarily conform to conventional assumptions; they may have asymmetric penalties for positive and negative errors, unexpected discontinuities in their penalties, or other nonstandard preferences. Further, they may treat probabilities nonlinearly (see Gonzalez and Wu 1999). As a result, drawing conclusions about biased beliefs based solely on a point estimate may require either unrealistic assumptions or more information than is typically available.

Limitations and Future Directions

This work is intended to be an initial investigation and has many limitations. Future work should refine and build on the present findings. This paper made several simplifying assumptions to facilitate the investigation of people's preferences over prediction error. Whereas it proposes that people exhibit diminishing sensitivity to prediction error, it does not take a definitive stance on the exact functional form of people's penalty functions, which could be power functions, logarithmic functions, hyperbolic functions, quasi-hyperbolic functions, or other alternatives. I assumed a power function, $P(e) = e^\alpha$, for concrete examples in the paper, but this choice is purely for illustrative simplicity and not intended to advocate for this particular functional form. Moreover, different functional forms might indicate distinct or partially distinct psychological processes, which future work should explicitly investigate. Additionally, this paper assumes that people evaluate absolute error (e) and apply equal penalties to all errors of the same magnitude (e.g., errors of 2 and -2 generate the same penalty). These are also assumptions of convenience rather than theoretically driven claims. Future research should investigate if and when these assumptions hold.

Separately, whereas the evidence in this paper found a tendency for participants to make and evaluate predictions with diminishing sensitivity to error, participants exhibited notable heterogeneity. For example, some participants did not make predictions approaching the mode, and some participants chose models built with increasing sensitivity. This variability was also evident across contexts; for example, participants

demonstrated more pronounced diminishing sensitivity when making predictions about dice rolls in Study 2b compared with predicting flower characteristics in Study 2a. Further, there may be domains in which most people exhibit constant or increasing sensitivity to prediction error although I have not found any yet. Thus, anyone providing people with predictions or interpreting people's predictions should ideally investigate the relevant people's preferences over error in the relevant domain, possibly employing methods from this research, such as eliciting reactions to different magnitudes of error or having individuals rate potential predictions given a distribution.

This work focuses on individual preferences regarding error when people make or evaluate single predictions. However, preferences might shift when individuals are involved in setting policies or making multiple predictions simultaneously. It is possible that people show increasing or constant sensitivity to error when considering many predictions. This suggests a potential difference in perspective: model builders, who naturally consider overall performance across multiple predictions, might prioritize aggregate measures of performance (e.g., MSE, MAE), whereas users may simply care whether the model gives them a good prediction this time. Exploring this potential disparity—whether people indeed have different preferences over error when supplying many estimates to others versus receiving one estimate for themselves—offers an intriguing avenue for future research.

The notion that diminishing sensitivity to error drives people to seek predictions that are “right” may also have implications for the types of estimates they prefer. For example, Hu et al. (2024) find that people prefer range estimates over point estimates for time prediction tasks likely because ranges increase the chance of being perceived as right. This raises open questions about how diminishing sensitivity influences people's evaluation of ranges, and whether alternative prediction formats could better satisfy people's desire to be right. Exploring these open questions may help extend the present framework to a broader class of prediction formats.

Finally, the present work investigated people's preferences over error isolated from other outcomes. However, people often use predictive algorithms and make predictions in domains in which they realize outcomes beyond error, such as gains or losses to wealth, products they like or dislike, and early or late arrivals at a destination. Whereas Study 3 found that preferences over error still influence decisions even when material outcomes are involved (see also Dietvorst and Fei 2021), further research is needed to understand how people balance prediction error against these other considerations. For example, this paper proposes that people are typically risk-seeking over prediction error, whereas prior research finds that people are

generally risk-averse concerning financial outcomes (Arrow 1971, Kahneman and Tversky 1979), and it remains unclear how people integrate these distinct preferences to make a prediction that informs a financial decision. It should be emphasized that predictions in contexts with material consequences may not always reflect diminishing sensitivity to error as people's preferences regarding the material outcome may strongly influence their predictions. Additionally, it remains unclear which penalty functions models should adopt when these objectives conflict, highlighting an important area for future inquiry.

Although this paper cannot specify which penalty functions people universally apply to predictions and other decisions, it emphasizes an important general lesson: The penalty function should never be merely an assumption or an afterthought. We should not assume that people are minimizing squared error or reporting the mean of their distribution of beliefs when interpreting their predictions. Similarly, we should not automatically build models to minimize squared error just because it is traditional or because it is the default setting in our statistical software, especially if we aim for our models to align with users' preferences. Instead, we should thoroughly investigate users' preferences concerning different outcomes and construct models that are maximally beneficial for the user.

Acknowledgments

The author is deeply grateful to Avner Strulov-Shlain for generously providing the mathematical proof presented in Supplement 1. The author is also grateful to Daniel Bartels, Andreas Kraft, Sanjog Misra, Uri Simonsohn, Bradley Shapiro, Avner Strulov-Shlain, Abigail Sussman, Alex Todorov, Oleg Urminsky, George Wu, and Yvette Yang for their helpful feedback. The author thanks the editor, the associate editor, and three anonymous reviewers for their exceptionally constructive and thoughtful feedback throughout the review process. Preregistrations, materials, data, code, and supplements are found in the supplemental material on the INFORMS site and at the following ResearchBox URL: <https://researchbox.org/3130>.

Endnotes

¹ I use $P(e) = e^x$ for concrete examples throughout the paper, but I do not advocate for this particular functional form. I chose this function because it is simple, and it can accommodate squared error (a commonly used metric).

² I give full credit for this proof to Avner Strulov-Shlain. I take full responsibility for any errors.

³ In all studies, the preregistered standard for making a prediction consistent with a summary statistic of the distribution (i.e., the mean, median, and mode) was making a prediction within 0.5 of that point.

⁴ The odds ratios in this section are from a logistic regression of whether participants predicted within 0.5 of a mode (coded 0/1) on participants' numeracy scores, which ranged from zero to three.

⁵ Here, 12.99% reported two steps, 24.02% reported three, 21.65% reported four, and 9.84% reported five steps.

⁶ In Study 2b, when excluding the first difference, the rating gap between neighboring errors no longer decreased significantly on average: $b = -0.02$, $t(252) = -0.80$, $p = 0.424$. In Study 2a, the decrease in ratings gap persisted after excluding the first difference: $b = -0.14$, $t(253) = -4.30$, $p < 0.001$.

⁷ This coefficient is much larger when "0 error" is included in the regression: $b = -0.89$, $t(252) = -34.01$, $p < 0.001$.

⁸ Here, 20.95% reported two steps, 17.79% reported three, 7.11% reported four, and 1.58% reported five steps.

⁹ I summarize performance in terms of mean absolute error because participants were incentivized based on absolute error.

References

- Andre Q (2016) distBuilder. <https://doi.org/10.5281/zenodo.166736>.
- Arrow KJ (1971) *Essays in the Theory of Risk-Bearing* (Markham, Chicago).
- Brainerd CJ, Gordon LL (1994) Development of verbatim and gist memory for numbers. *Developmental Psych.* 30(2):163–177.
- Carroll CD, Fuhrer JC, Wilcox DW (1994) Does consumer sentiment forecast household spending? If so, why? *Amer. Econom. Rev.* 84(5):1397–1408.
- Chun JS, Larrick RP (2022) The power of rank information. *J. Personality Soc. Psych.* 122(6):983–1003.
- Dawes RM (1979) The robust beauty of improper linear models in decision making. *Amer. Psych.* 34(7):571–582.
- Dawes RM, Faust D, Meehl PE (1989) Clinical versus actuarial judgment. *Science* 243(4899):1668–1674.
- Deci EL, Ryan RM (2000) The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psych. Inquiry* 11(4):227–268.
- Dietvorst BJ (2025) Understanding when laypeople adopt predictive algorithms. *Nature Human Behav.* 9(5):851–853.
- Dietvorst BJ, Bharti S (2020) People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psych. Sci.* 31(10):1302–1314.
- Dietvorst BJ, Fei L (2021) People take more risk when their prospects are tied to future states of the world. Preprint, submitted November 8, <https://doi.org/10.2139/ssrn.3955918>.
- Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych. General* 144(1):114–126.
- Dimitriadis T, Patton AJ, Schmidt PW (2019) Testing forecast rationality for measures of central tendency. Preprint, submitted October 28, <https://arxiv.org/abs/1910.12545>.
- Einhorn HJ (1986) Accepting error to make less error. *J. Personality Assessment* 50(3):387–395.
- Estes WK (1950) Toward a statistical theory of learning. *Psych. Rev.* 57(2):94–107.
- Fechner GT (1860/1966) *Elemente Der Psychophysik [Elements of Psychophysics]*, vol. 1 (Adler HE, trans.) (Holt, New York).
- Festinger L (1954) A theory of social comparison processes. *Human Relations* 7(2):117–140.
- Fisher M, Keil FC (2018) The binary bias: A systematic distortion in the integration of information. *Psych. Sci.* 29(11):1846–1858.
- Frydman C, Jin LJ (2022) Efficient coding and risky choice. *Quart. J. Econom.* 137(1):161–213.
- Fuchs VR, Eggleston K (2018) Life expectancy and inequality in life expectancy in the United States. Policy Brief, Stanford Institute for Economic Policy Research, Stanford, CA.
- Gao J, Corter JE (2015) Striving for perfection and falling short: The influence of goals on probability matching. *Memory Cognition* 43(5):748–759.

- Gonzalez R, Wu G (1999) On the shape of the probability weighting function. *Cognitive Psych.* 38(1):129–166.
- Goodwin P (1996) Statistical correction of judgmental point forecasts and decisions. *Omega* 24(5):551–559.
- Greer R (2020) Margin probabilities from NFL spreads. Accessed June 6, 2024, <https://www.nfeloapp.com/analysis/margin-probabilities-from-nfl-spreads/>.
- Grice HP (1975) Logic and conversation. *Syntax Semantics* 3:43–58.
- Grove WM, Meehl PE (1996) Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psych. Public Policy Law* 2(2):293–323.
- Highhouse S (2008) Stubborn reliance on intuition and subjectivity in employee selection. *Indust. Organ. Psych.* 1(3):333–342.
- Hogarth RM, Gibbs BJ, McKenzie CR, Marquis MA (1991) Learning from feedback: Exactingness and incentives. *J. Experiment. Psych. Learn. Memory Cognition* 17(4):734–752.
- Hu B, Simmons JP (2025) Different methods elicit different belief distributions. *J. Experiment. Psych. General* 154(2):476–496.
- Hu B, Gaertig C, Dietvorst BJ (2024) How should time estimates be structured to increase customer satisfaction? *Management Sci.* 71(9):7497–7515.
- Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.
- Lehmann EL, Casella G (2006) *Theory of Point Estimation* (Springer Science & Business Media, New York).
- Lipkus IM, Samsa G, Rimer BK (2001) General performance on a numeracy scale among highly educated samples. *Medical Decision Making* 21(1):37–44.
- Reyna VF (2012) A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment Decision Making* 7(3):332–359.
- Siegler RS, Opfer JE (2003) The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psych. Sci.* 14(3):237–250.
- Snyder CR (1978) The “illusion” of uniqueness. *J. Humanistic Psych.* 18(3):33–41.
- Starmer C (2000) Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *J. Econom. Literature* 38(2):332–382.
- Stevens SS (1957) On the psychophysical law. *Psych. Rev.* 64(3):153–181.
- Vulkan N (2000) An economist’s perspective on probability matching. *J. Econom. Surveys* 14(1):101–118.
- Weber EH (1834) *De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae* (CF Koehler, Leipzig).
- White RW (1959) Motivation reconsidered: The concept of competence. *Psych. Rev.* 66(5):297–333.