Judgment Errors in Naturalistic Numerical Estimation

Wanling Zou and Sudeep Bhatia

University of Pennsylvania

February 18, 2021

Word count: 14,917

Send correspondence to Wanling Zou, Department of Psychology, University of Pennsylvania, Philadelphia, PA. Email: wanlingz@sas.upenn.edu. Funding was received from the National Science Foundation grant SES-1626825 and the Alfred P. Sloan Foundation.

Abstract

People estimate numerical quantities (such as the calories of foods) on a day-to-day basis. Although these estimates influence behavior and determine wellbeing, they are prone to two important types of errors. Scaling errors occur when people make mistakes reporting their beliefs about a particular numerical quantity (e.g. by inflating small numbers). Belief errors occur when people make mistakes using their knowledge of the judgment target to form their beliefs about the numerical quantity (e.g. by overweighting certain cues). In this paper, we quantitatively model numerical estimates, and in turn, scaling and belief errors, in everyday judgment tasks. Our approach is unique in using insights from semantic memory research to specify knowledge for naturalistic judgment targets, allowing our models to formally describe nuanced errors in belief not considered in prior research. In Studies 1 and 2, we find that belief error models predict participant estimates and errors with very high out-of-sample accuracy rates, significantly outperforming the predictions of scaling error models. In fact, the best-fitting belief error models can closely mimic the inverse-S shaped patterns captured by scaling error models, suggesting that the types of responses previously attributed to scaling errors can be seen as errors of belief. In Studies 3 to 8, we find that belief error models are also able to predict people's responses in semantic judgment, free association, and verbal protocol tasks related to numerical judgment, and thus provide a good account of the cognitive underpinnings of judgment.

Keywords: judgment errors; numerical estimation; word vectors; knowledge representation; cognitive model; semantic cognition

Introduction

Decades of research on human cognition and judgment has established that people make systematic errors when estimating numerical quantities, such as the frequencies of lethal events, proportions of demographic groups, or the calories of food items (Chernev & Chandon, 2011; Landy, Guay, & Marghetis, 2018; Lichtenstein et al., 1978). Researchers studying these judgment errors have identified a number of factors responsible for numerical misestimation, including the use of non-linear weighting functions (e.g. Gonzalez & Wu, 1999; Hollands & Dyre, 2000; Landy et al., 2018; Tversky & Kahneman, 1992) or the use of heuristic cue-aggregation rules (Brown & Siegler, 1993; von Helversen & Rieskamp, 2008).

These factors can be understood as involving two types of errors – scaling errors and belief errors. Scaling errors occur when people make mistakes reporting their beliefs about a particular numerical quantity. In other words, people may have the correct belief about the numerical quantity (e.g. they may know that the calorie amount of 100g of walnuts is 654kcal[1]) but incorrectly scale this belief to a response (e.g. by reporting that the calorie amount of 100g of walnuts is 160kcal due to a deflation of large values). Prior literature has found evidence for such distortions of numerical representations, indicating that psychological magnitude does not always correspond to the same physical magnitude (Stevens, 1957, 1975; Stevens & Galanter, 1957). These distortions are the cause of the overestimation of small values and underestimation of large values, which leads to the common inverse-S-shape pattern when plotting participant estimates against objective statistics (e.g. Erlick, 1964; Hollands & Dyre, 2000; Landy et al., 2018; Varey, Mellers, & Birnbaum,1990). Inverse S-shaped patterns in numerical estimation can

[1] According to the United States Department of Agriculture (USDA), 100g of walnuts have 654kcal.

be modeled using various non-linear functions, e.g. polynomial functions and their variants (Curtis, Attmeave, & Harrington, 1968; Hollands & Dyre, 2000), log-odds transformations (Shepard, 1981; Zhang & Maloney, 2012), and probability weighting functions (Fennell & Baddeley, 2012; Tversky & Kahneman, 1992). Models that use non-linear functions to describe scaling errors typically assume that a systematic distortion takes place when transforming otherwise correct internal beliefs on to an explicit numerical response.

In contrast, belief errors occur when people make mistakes using their knowledge of the judgment target to form their beliefs. These can lead to the formation of incorrect beliefs through, for instance, the biased use of memory cues (e.g. thinking 100g of walnuts have only 160kcal because walnuts are associated with health and nutrition), though people may still be able to accurately report these beliefs (e.g. by answering 160kcal). For example, Chernev and Chandon (2011) have studied biases in food calorie estimation and have found that health-related cues are given an incorrectly high weight, which can then lead to the underestimation of food calories for healthy food sources. Media coverage and word frequency have also been shown to be used as cues in probability estimation (Dougherty, Franco-Watkins, & Thomas, 2008; Tversky & Kahneman, 1974) and frequency estimation (Hertwig, Pachur, & Kurzenhäuser, 2005; Lichtenstein et al., 1978), which can lead to the overestimation of the size of minority groups (Gallagher, 2003; Herda, 2013, 2015). More generally, many researchers in psychology have proposed that people use heuristics to weight and aggregate judgment cues, which can, at times, lead to belief errors in numerical estimation. These heuristics simplify the decision process by ignoring certain cues and thus assigning them incorrectly low weights, or by using equal weights for all cues and thus over-weighting irrelevant cues and under-weighting relevant cues (see

Hertwig, Hoffrage, & Martignon, 1999; Juslin, Olsson, & Olsson, 2003; von Helversen & Rieskamp, 2008).

Our division of numerical judgment errors into scaling and belief errors has precedent. For example, Lichtenstein et al. (1978) suggested that there are two types of biases in frequency estimation – a primary bias (specifically, an overestimation of small numbers and underestimation of large numbers) and a secondary bias (involving the formation of incorrect beliefs due to the overuse of certain memory cues, such as familiarity and imaginability). Likewise, Brown and Siegler (1993) argued that there are two types of knowledge that come into play in quantitative estimation – metric knowledge (which involves knowledge of the distributional properties in the judgment domain) and mapping knowledge (which involves knowledge of the relative ranks of different entities within the distribution). Von Helversen and Rieskamp (2008) similarly argued that people are likely to sample objects that are similar to the judgment target (where belief errors may occur) and make estimates based on some transformation of the sampled objects' values (where scaling errors may occur). Finally, Landy et al. (2018) suggested that two different features could lead to errors in domains such as demographic estimation – domain-general process (i.e. a log-odds mental representation of proportion) and topic-specific bias (e.g. media bias and xenophobia).

Although this work has greatly expanded our understanding of numerical estimation, almost all of it uses artificial experimental stimuli, such as the toxicity of fictional bugs (Juslin et al., 2003), the percentage of white dots in a mixture of black and white dots (Varey et al., 1990) and the proportion of certain letters in a random letter string (Erlick, 1964). There is work that studies more naturalistic judgment targets, though these targets are typically modeled using a small set of experimenter-generated cues that provide only an abstract representation of the rich

knowledge representations for the judgment targets in people's minds. For example, Lichtenstein et al. (1978) studied belief errors by predicting participant estimates of frequencies of lethal events from cues such as participant ratings of their experience with lethal events and lengths of news articles reporting lethal events. These experimenter-generated cues do predict participant estimates. However, they certainly do not represent all of what people know about and associate with lethal events, and may not be enough for understanding how people use knowledge to form beliefs and the judgment errors that could arise from these beliefs.

Research on semantic memory offers one solution to this problem. Specifically, high-dimensional semantic spaces, that have previously used to model knowledge representations for words and concepts (Günther, Rinaldi, & Marelli, 2019; Jones, Willits & Dennis, 2015; Landauer & Dumais, 1997; Mandera, Keuleers & Brysbaert, 2017; Mikolov et al., 2013), can also be used to specify people's knowledge for judgment targets (Bhatia, 2017, 2019a; Bhatia, Richie & Zou, 2019), allowing models of numerical estimation to permit more complex forms of knowledge, and thus more complex forms of belief errors, not considered in prior research. Doing this has numerous practical and theoretical benefits. Firstly, by allowing for richer and more nuanced belief errors we can greatly improve our ability to predict estimates (and misestimates) for judgment targets. This has direct policy implications, and our tests, which involve judgments of food calories and judgments of infant mortality rates, can be used to inform behavioral interventions for healthy eating and charitable giving (e.g. Butera & Houser, 2018; Glanz & Mullis, 1988; Goswami & Urminsky, 2016; Michie et al., 2009; Oppenheimer & Olivola, 2011).

Additionally, by understanding accuracy rates using richer knowledge representations such tests can also provide a better estimate of the explanatory power of belief error models (relative to scaling error models). This can help researchers better understand how and why

people make mistakes in numerical estimation, and the set of theoretical assumptions necessary to describe these mistakes. Some prior work has found that after allowing for a domain-general bias (which is due to a nonlinear psychophysical rescaling and is analogous to a scaling error), there is little evidence for a topic-specific bias (which is due to media bias or xenophobia and is analogous to a belief error) in participant estimates of demographic proportions, suggesting that errors in demographic proportion estimates are largely due to scaling errors (Landy et al., 2018). However richer knowledge representations for judgment targets, as well as formal models for scaling these representations to responses, could result in different conclusions if such representations accurately predict the structure of people's judgment errors.

Finally, formally modeling the knowledge at play in numerical judgment will generate a direct theoretical link between judgment research and research in semantic cognition (see e.g. Jones et al., 2015 or Mandera et al., 2017 for a review), which pertains directly to the structure of people's knowledge representations, and how these representations guide memory, language, and various high-order cognition tasks. These theoretical linkages can also be used to make novel predictions regarding the use of memory in numerical estimation, e.g. by specifying the words and concepts that are most strongly associated with or related to the estimated quantity. This in turn can provide a more detailed understanding of the cognitive processes at play in numerical estimation tasks.

Overview of Approach

The key insight underpinning our tests is that knowledge for naturalistic judgment targets can be captured using high-dimensional vector representations obtained from semantic space models. Semantic space models are powerful theoretical tools in semantic memory research, that use the distributional structure of words and concepts in large natural language datasets, such as

text corpora, to derive knowledge representations for the words and concepts. The corpora can be a collection of books, Wikipedia, product reviews, and news articles, all of which capture the knowledge shared among humans in a rich linguistic environment. As the linguistic environment informs learning, and human behavior is shaped by how knowledge is communicated and exchanged (Estes, 1955; Simon 1969), text corpora can serve as an approximation of language experience (Johns, Jones, & Mewhort, 2016, 2019) in which the distributional patterns of words correspond to their meanings or knowledge representations (Günther et al., 2019; Mandera et al., 2017; Lenci, 2018; Hollis, 2017). For this reason, the proximity between words in the semantic space derived from word distributions in text corpora can be used to predict judgments of semantic similarity and relatedness (Bhatia et al., 2019; Jones et al., 2015; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mandera et al., 2017; Mikolov et al., 2013; Pennington, Socher & Manning, 2014). Vector space semantics have also been shown to mimic representations at play in other settings, such as those involving semantic priming (Jones, Kintsch & Mewhort, 2006; Günther, Dudschig, & Kaup, 2016), semantic memory search (Hills, Jones & Todd, 2012), free recall (Manning & Kahana, 2012; Steyvers, Shiffrin, & Nelson, 2004), probability judgment and factual judgment (Bhatia, 2017; Bhatia & Walasek, 2019), risk perception (Bhatia, 2019a), multiattribute choice (Bhatia, 2019b; Bhatia & Stewart, 2018), and stereotype judgment (Bhatia, 2017; Caliskan, Bryson, & Narayanan, 2017; Garg et al., 2018). In all of these domains, semantic vectors for items in memory, judgment, and decision making tasks, specify what people know and associate with the items and thus predict their responses in the tasks.

Due to their ability to describe knowledge representations at play in people's judgments, we propose that semantic space models can also be used to study the types of belief errors that accompany these judgments. Specifically, the vector representations for judgment targets

specified by semantic space models are analogous to cue vectors used in existing research on numerical estimation. These semantic vectors can be multiplied by cue weights to predict participant responses, and the weighting vectors used by participants can be inferred by fitting linear judgment models to participant numerical estimation data (as in e.g. Brunswik, 1952; Gigerenzer & Kurz, 2001; Karelaia & Hogarth, 2008). Belief errors emerge if certain vector dimensions are given disproportionally high or low weights (relative to the optimal weight vector for predicting the criterion variable). Note that such a linear model of judgment implicitly assumes that there is no further non-linear transformation of beliefs onto the response scale. In this sense, this model permits only belief errors. The ability of such a model to predict participant estimates is a measure of the explanatory scope of belief error models in describing numerical estimation. Figure 1 provides an illustration of this approach. Additional technical details are provided in subsequent sections.

Semantic space models are not only capable of predicting judgment errors. They can also provide insights regarding the conceptual underpinnings of these errors. Particularly, the weight vectors used by participants to map vector representations of judgment targets onto the response variable, are themselves vectors in the semantic space (see Figure 1). These vectors occupy regions of the semantic space that are most associated with or related to high responses on the dimension being judged. Words and concepts that are close to the weight vectors in this space are, in this sense, highly associated with or related to the judgment dimension. These words and concepts may not themselves be judgment targets, but if they were, they would be given the highest estimates by participants. Contemporary semantic space models have rich vocabularies, spanning hundreds of thousands of words and concepts, and can thus be used to identify which of a very large set of words are most associated with the judgment dimension in consideration.

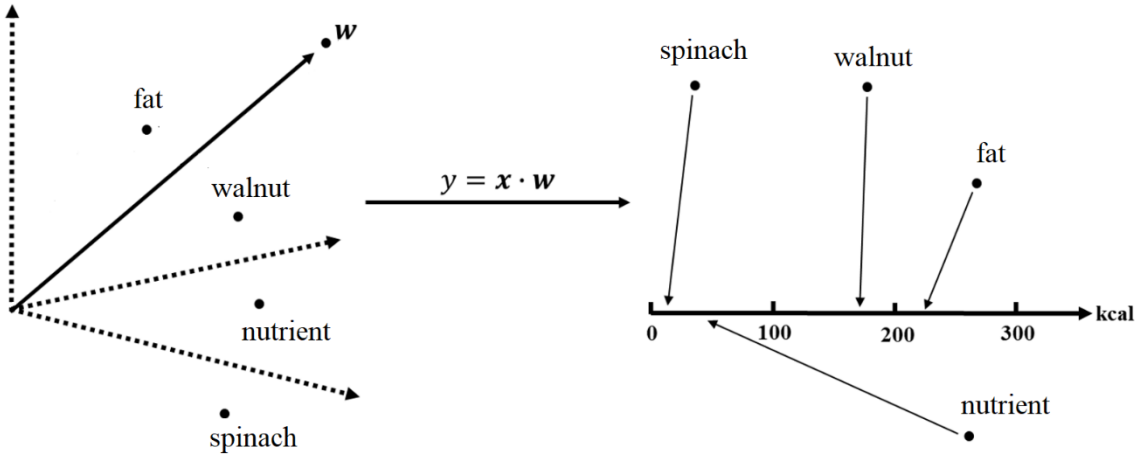


Figure 1. Illustration of a vector semantic space specification of a belief error model. Each judgment target *i* (e.g. walnuts or spinach) has a normalized vector representation, $\boldsymbol{x_i}$, in the semantic space. A vector of weights $\boldsymbol{w}$ multiplies against $\boldsymbol{x_i}$ to generate participant estimates of target *i*, y_i, on the judgment dimension (e.g. calories). This approach can be applied to non-target concepts (e.g. nutrient and fat) to predict participant estimates of that concept's calories as if it were a judgment target. This serves as a measure of the strength of association between that concept and the judgment dimension. As vectors are normalized, multiplying $\boldsymbol{w}$ against $\boldsymbol{x_i}$ is proportional to measuring the cosine similarity between $\boldsymbol{w}$ and $\boldsymbol{x_i}$.

We will use this approach to test for the conceptual underpinnings of numerical estimation predicted by our belief error models.

The predictions of the belief error models can be contrasted with those of scaling error models. These are quite simply non-linear functions of the true value of the criterion variable being judged, and thus implicitly assume that people have correct beliefs (i.e. know the true value of this variable), but map these beliefs onto the response scale in some biased manner. As scaling models are monotonic (e.g. see Hollands & Dyre, 2000; Litchtenstein et al., 1978), they also predict that the rank-ordering of participant estimates should be fully accurate. Thus, judgment targets with higher values on the criterion should always be given higher estimates, even if the estimates themselves are incorrect. Belief error models, in contrast, can generate participant estimates whose rank-ordering deviates from the correct ordering. For example,

judgment targets with high values on overweighted cue dimensions can be given incorrectly high ranks.

In this paper, we will be examining the predictions of both belief and scaling error models. Such tests will allow us to directly compare the explanatory scope of these two causes of judgment errors. Additionally, the use of a flexible non-linear function to scale correct criterion values to participant estimates provides a particularly stringent test of our proposed semantic space specification of belief errors. We can be confident in the power of our belief error models if they are able to provide a more accurate account of participant data than the scaling error models (despite the belief error models not having access to the correct value of the criterion variable).

We will consider two domains to test our models: estimates of calories in food items and estimates of infant mortality rates in countries. Judgments in these domains influence health decision making and charitable decision making, making these domains suitable for applied (as well as theoretical) research on judgment. Using the insights developed in this paper, it may be possible to craft behavioral interventions to increase healthy and charitable choice (Butera & Houser, 2018; Glanz & Mullis, 1988; Goswami & Urminsky, 2016; Michie et al., 2009; Oppenheimer & Olivola, 2011). Similar judgment domains have been previously used to study numerical estimation. For example, Chandon and Wansink (2007) found that (in a between-subject design) calorie estimates were significantly lower when participants were given a healthy menu than when they were given an unhealthy one, and Hertwig et al. (2005) and Litchtenstein et al. (1978) found that participants on average overestimated small (and underestimated large) incidence rates and mortality rates of lethal events (e.g. accidents and diseases).

In Studies 1 and 2, we will examine participant estimates for 400 distinct naturalistic entities in these two domains and fit scaling and belief error models separately to predict participant estimates. We will compare the accuracy of the scaling and belief error models using their out-of-sample predictions. In Studies 3 and 4, we will examine the accuracy of the word and concept associations predicted by the belief error models by asking participants to rate the relatedness of these words and concepts to the corresponding judgment dimensions. In Studies 5 and 6, we will use free association data to examine the words that are most associated with the high and low ends of the two judgment dimensions. Finally, in Studies 7 and 8, we will use verbal protocols to further explore the mental process underlying estimation. Through the semantic judgments in Studies 3 and 4, free associations in Studies 5 and 6, and verbal protocols in Studies 7 and 8, we can obtain a detailed understanding of how well our belief error models predict the conceptual underpinnings of numerical estimation.

Studies 1 and 2: Modeling and Predicting Judgment Errors

Experimental Methods

Participants. We recruited a total of 50 participants (mean age = 30 years, 52% female) in Study 1 and 51 participants (mean age = 31 years, 60% female) in Study 2 from Prolific Academic, an online experiment platform. All participants were residents of the United States and had an approval rate of 80% or above.

Stimuli. For Study 1, we obtained 200 food items and their calorie amounts from the United States Department of Agriculture database (https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/). Sample items included walnuts, spinach, lamb, mint, etc. For Study 2, we obtained the infant mortality rates of 200 countries from the Central Intelligence Agency World Factbook

(https://www.cia.gov/library/publications/the-world-factbook/rankorder/2091rank.html). Food items and countries were chosen to ensure that they were present in the vocabulary of the semantic space used for our belief error models (see computational methods below).

Procedure. In Study 1, participants were asked to estimate how many calories (in kcal) there are in 100 grams of a particular food item; in Study 2, they were asked to estimate the infant mortality rate (i.e. number of deaths of children younger than one year, per 1,000 live births) in a particular country. The sample prompt for Study 1 read "How many calories are there in 100g of [*food*]?" and that for Study 2 read "In the [*country*]**,** what is the infant (child under 1-year-old) mortality rate, in the number of deaths per 1,000 live births?". Participants entered answers in a textbox in both studies. Each participant estimated all 200 judgment targets and saw only one target on each screen. The order of the 200 targets was randomized and there was a 30-second break after every 50 targets. Additionally, for Study 1, participants were also told that 100g is approximately the weight of a stick of butter, allowing them to calibrate their responses for the response scale in the study.

Computational Methods

Scaling Error Model. For each target i (e.g. walnuts), we obtained the (aggregate or individual-level) participant estimate y_i (e.g. estimated calories in walnuts) and the correct answer z_i (e.g. actual calories in walnuts). Our first model for describing participant estimates was a linear model that assumes that y_i is simply z_i multiplied by some constant and added to an intercept (Eq. 1). Although linear patterns are rarely found in previous literature, we included it as a baseline. Note that the baseline model serves as a useful contrast against both scaling models (that use z_i to predict y_i in a non-linear manner) and belief error models (which use a linear transformation of a knowledge or cue vector to predict y_i).

$$y_i = \alpha + \beta z_i \quad (1)$$

To quantitatively study scaling explanations for estimation errors, we fit two different scaling error models that transformed correct answers into participant responses. Formally, our scaling error models predicted y_i as some non-linear function of z_i. The first function we used was a third-degree polynomial (Eq.2), and the second function was a power function with a constant term (Eq.3). A third-degree polynomial can capture a potential S-shape or inverse-S-shape pattern in participant responses (as in e.g. Erlick, 1964; Landy et al., 2018; Varey et al.,1990). We incorporated a power function due to its prevalence in prior work (e.g. see Hollands & Dyre, 2000).

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 \quad (2)$$

$$y_i = \gamma + \theta z_i^{\delta} \quad (3)$$

Parameters for the baseline model and the two scaling models were estimated by minimizing the residual sum of squares to participant responses. We used established Python functions numpy.polyfit for Eq. 1 and Eq. 2 (which gives least squares polynomial fit) and scipy.optimize.fmin for Eq. 3 (which minimizes sum squared errors using the downhill simplex algorithm) with 100 iterations and random starting points for each iteration.

Belief Error Model. To examine belief errors, we used the pre-trained Word2Vec semantic space model (Mikolov et al., 2013) to obtain semantic representations for food items and countries. This model was trained on a large dataset of Google News articles (roughly 100 billion words in size with a vocabulary of three million unique tokens) using the continuous bag-of-words (CBOW) method (which predicts words from their neighbors) and the skip-gram method (which predicts neighboring words of a given word). These two methods allow words that appear in similar contexts and share related meanings to be located in close proximity in the

high dimensional semantic space, and the Word2Vec semantic space has previously been shown to generate accurate predictions of human similarity judgment and associative judgment (Baroni, Dinu, & Kruszewski, 2014; Bhatia, 2017; Bhatia & Walasek, 2019; Caliskan et al., 2017; Pereira et al., 2016; Hollis, Westbury, & Lefsrud, 2017).

Each of the three million words and phrases in the Word2Vec vocabulary is described using a 300-dimensional vector. This vocabulary includes a large number of food items and countries, which we use in our studies. Specifically, for each food target, we used the Word2Vec vector representation corresponding to the lower case of the food word (e.g. walnuts). We took the plural form only for foods that are normally consumed in bulk (e.g. walnuts, blueberries, and bran flakes). Everything else was in the singular form. It is also worth noting that in the Word2Vec model, singular and plural forms have very similar vectors. For each country, we used the Word2Vec representation corresponding to the first letter uppercased and remaining letters lowercased for the country word (e.g. France), except for a few countries whose vector representations could only be found with all letters uppercased. Foods and countries with multiple words in their names had each word separated by an underscore (e.g. peanut_butter or South_Africa). Although the vectors in the original Word2Vec vocabulary have different magnitudes, we normalized all vectors to unit norm prior to analysis.

Ultimately, the Word2Vec space gave us a 300-dimensional vector $\boldsymbol{x_i}$ for each target i, where x_{ij} is the value of target on the j^{th} dimension, and $||\boldsymbol{x_i}|| = 1$. These vectors quantify the knowledge representations people use to make numerical estimates. Our belief error model, inspired by prior work on numerical judgment (Brunswik, 1952; Gigerenzer & Kurz, 2001; Karelaia & Hogarth, 2008), used a linear transformation of $\boldsymbol{x_i}$, with weight vector $\boldsymbol{w}$, to predict y_i (Eq. 4). As discussed above, each dimension of $\boldsymbol{x_i}$ can be seen as a cue that participants might

rely on to facilitate estimation, and the weight vector, $\boldsymbol{w}$, can be seen as capturing cue weights on knowledge and transforming knowledge represented in a 300-dimensional space to a one-dimensional numerical estimation line (Figure 1). Mathematically, multiplying two n-dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ results in a scalar c, with $c = \sum_{i=1}^{n} a_i * b_i$. For example, (1,2,3) * (3,0,5) = 1*3 + 2*0 + 3*5 = 3 + 0 + 15 = 18.

The weight vector and intercept of this model were estimated by minimizing the residual sum of squares. Due to the high dimensionality of the weight vectors $\boldsymbol{w}$, we estimated the weights with a ridge regression (also known as L2 regularization). This involves a loss function with an additional penalty based on the sum of squares of the weight vector multiplied by hyperparameter λ (Eq. 5). Using this penalty mitigates the problem of multicollinearity in such high-dimensional modeling problems and leads to better fits.

$$y_i = w_0 + \boldsymbol{w}\boldsymbol{x_i} \tag{4}$$

$$\sum_i \left(y_i - w_0 - \sum_j x_{ij} w_j\right)^2 + \lambda \sum_j {w_j}^2 \tag{5}$$

We implemented the ridge regression in the Python scikit-learn machine learning library (Pedregosa et al., 2011). There were a set of hyperparameters in this library. To avoid manipulating the hyperparameters to improve model performance, we took the default values of all these hyperparameters (including, for example, $\lambda = 1$). We focused on ridge regression because previous results (e.g. Bhatia, 2019a; Richie, Zou, & Bhatia, 2019) suggested that compared to other regression techniques such as lasso regression and support vector regression, ridge regression often works best in predicting human judgments from vector semantic representations.

To avoid overfitting, we compared scaling and belief error models through leave-one-out cross-validation (LOOCV), on both the aggregate and the individual level. Specifically, for each

domain, we trained our models on 199 judgment targets and then used the trained model to predict participant estimates of the left-out target. This procedure was repeated for each judgment target to get LOOCV predictions.

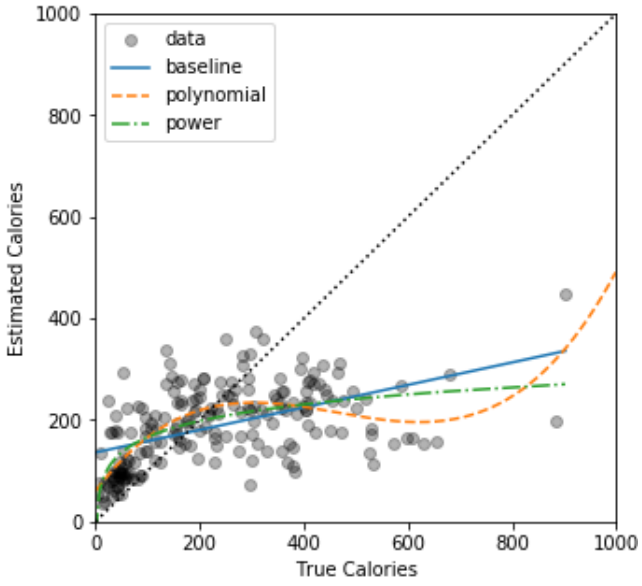
Results

Summary of data. Study 1 and 2 elicited a total of 40,400 participant estimates for 400 distinct naturalistic entities in two domains – calories of 200 common food items and infant mortality rates of 200 countries. Consistent with prior work, we found that participants made substantial errors. The average absolute differences between the participant estimates, y_i, and the correct answers, z_i, for food calories and infant mortality rates were -45.28kcal per 100g (*SE* = 10.80) and 53.44 deaths per 1,000 live births (*SE* = 1.35) respectively, indicating an overall underestimation of food calories and overestimation of infant mortality rates.

Figure 2 plots the average participant estimates for the calories and mortality rates against the true calories and mortality rates in the two studies. Here each point is a judgment target. Figure 2 shows some overestimation of low calories, significant underestimation of high calories, and overall overestimation of infant mortality rates (with slightly more overestimation of small infant mortality rates and less overestimation of large infant mortality rates). This is consistent with an inverse-S shaped pattern previously documented in judgment and decision making research.

Participants largely underestimated the calories of small-size foods (e.g. nuts, dressing, and cookies) and overestimated the calories of creamy and savory foods (e.g. milkshake, cappuccino, and chicken gumbo). Overestimation of infant mortality rates was mainly driven by African countries (e.g. Zimbabwe, Rwanda, and Ethiopia) and Middle Eastern countries (e.g. Syria, Kuwait, and Iran). It is not immediately clear what leads to these judgment errors. For

(a)



(b)

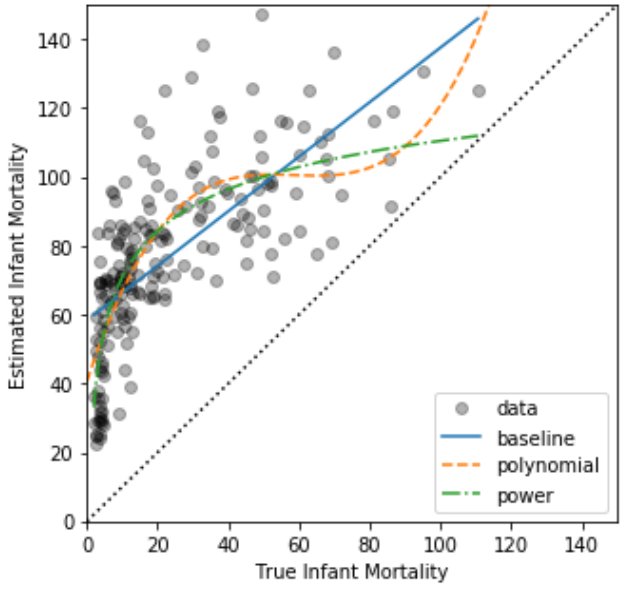


Figure 2. Scatterplots of average participant estimates vs. correct answers in (a) Study 1 (food calories) and (b) Study 2 (infant mortality). The dotted lines represent perfect calibration where participant estimates are equal to the correct answers. Curves represent the baseline model (baseline – Eq. 1) and scaling error models with different scaling functions (polynomial – Eq. 2, power – Eq. 3), which were trained on the aggregate-level data.

example, it could be that small-size foods have high calories per 100g and people systematically underestimate high calories (scaling error). Or it could be that people overweight portion size as a cue and thus think that small-size foods tend to have low calories (belief error). Note that

scaling error models predict that the rank ordering of participant estimates should be correct, which means that the rank correlation between average participant estimates and correct answers should be close to one. However, the rank correlation was 0.56 for food calories ($p < 10^{-16}$) and 0.78 for infant mortality rates ($p < 10^{-41}$), which indicates that scaling error models cannot provide a complete account of the data. To formally test whether belief errors are present in our data, and if belief error models can outperform the predictions of scaling error models, we need to quantitatively fit and evaluate the models.

Aggregate-level fits. Figure 2 shows the predictions of the best-fitting baseline model, as well as the best-fitting scaling error models with two different scaling functions (all trained on the entire dataset of aggregate participant responses). Table 1 summarizes the parameter values of these models. Although it appears based on Figure 2 and Table 1 that nonlinear scaling functions like the polynomial and power functions capture a lot of variance in participant estimates, these tests do not control for model flexibility, and there may be problems with overfitting. To avoid this, we evaluated model accuracy using leave-one-out cross-validation (LOOCV). The results of the LOOCV analysis are presented in Table 2. This table also contrasts the predictions of the baseline and scaling error models with belief error models trained on the aggregate participant data. Model performance is evaluated using the Pearson correlation, r, and root mean square error (RMSE), between the observed y_i and the out-of-sample prediction of y_i, for each i. Figures 3 and 4 show scatterplots of predicted estimates for each judgment target on the aggregate data using LOOCV and average participant estimates, along with Pearson correlations. Here each point is a different judgment target, and the prediction for each judgment target is obtained by fitting the models on all other targets excluding that target.

	Study 1	Study 2
Baseline		
α	136.13	58.35
β	0.22	0.79
Scaling Error -- Polynomial		
β_0	60.19	40.48
β_1	1.37	3.15
β_2	$-3.35*10^{-3}$	-0.05
β_3	$2.42*10^{-6}$	$3.09*10^{-4}$
Scaling Error -- Power		
γ	-714.25	-1010.51
θ	696.81	1035.95
δ	0.05	0.02

Table 1. Parameter estimates of the baseline model (Eq. 1) and scaling error models trained on entire data with different scaling functions – polynomial (Eq. 2), power (Eq. 3) for Study 1 (food calories) and Study 2 (infant mortality).

	Study 1			Study 2		
	Correlation	R^2	RMSE	Correlation	R^2	RMSE
Baseline	0.48	0.23	67.89	0.68	0.46	18.67
Scaling Error – Polynomial	0.61	0.37	61.42	0.74	0.55	17.08
Scaling Error – Power	0.59	0.35	62.47	0.78	0.60	16.08
Belief Error	0.83	0.68	43.86	0.83	0.68	14.43

Table 2. Aggregate level out-of-sample predictive accuracy of the baseline model (Eq. 1), and scaling error model with different scaling functions – polynomial (Eq. 2), power (Eq. 3), and belief error model, using leave-one-out cross-validation (LOOCV) for Study 1 (food calories) and Study 2 (infant mortality).

As can be seen in Table 2 and Figures 3 and 4, the belief error model was able to predict participant estimates fairly accurately, with out-of-sample accuracy rates of $r = 0.83$ for both domains on the aggregate level. In contrast, the highest aggregate level out-of-sample correlation rates achieved by scaling error models were 0.61 for calorie estimates and 0.78 for infant mortality rate estimates. The baseline model had the lowest correlation of 0.48 for calories and 0.68 for mortality rates (all correlations are significantly different from 0, at $p < 10^{-12}$). These differences in model correlations were accompanied by differences in out-of-sample RMSE error rates. The belief error model generated an RMSE of 43.86 and 14.43 for Study 1 and 2. In contrast, the lowest RMSEs for the scaling error models were 61.42 and 16.08, and the RMSEs for the baseline models were 67.89 and 18.67. The differences in RMSEs between the belief error models and the best-performing scaling error models are statistically significant when evaluated with a paired t-test on the target-level ($df = 199$, $t = -5.13$, $p < 10^{-6}$ for Study 1; $df = 199$, $t = -2.14$, $p = 0.03$ for Study 2). Both belief and scaling error models outperformed the baseline model (all $p < 0.02$ in paired t-tests).

Note that there appear to be floors in the baseline predictions for both domains (Baseline in Figures 3a and 4a) and caps in the polynomial predictions for both domains (Scaling Error – Polynomial in Figures 3b and 4b) and in the power prediction for calorie estimates (Scaling Error – Power in Figure 3c). These were likely due to the intercepts that overrode the effect of other parameters (see Table 1), and suggest a substantial limitation in the predictive power of scaling error models. Belief error predictions, in contrast, do not appear to suffer from these limitations.

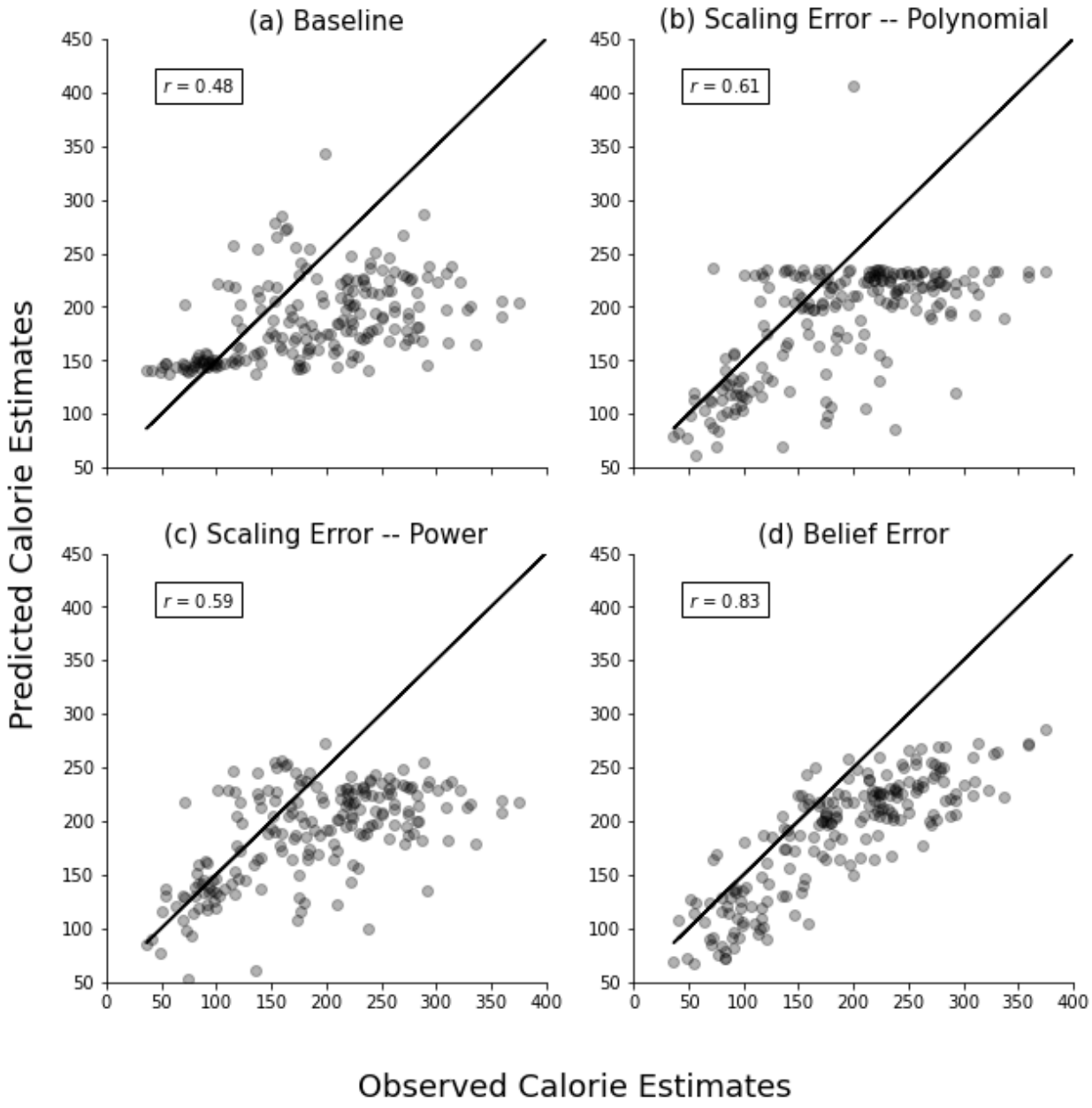


Figure 3. Scatterplots of predicted estimates by (a) baseline model (Eq. 1), (b) scaling error model with polynomial scaling function (Eq. 2), (c) scaling error model with power scaling function (Eq. 3), (d) belief error model (Eq. 4) using leave-one-out cross-validation (LOOCV) vs. actual participant estimates in Study 1 (food calories), along with Pearson correlations.

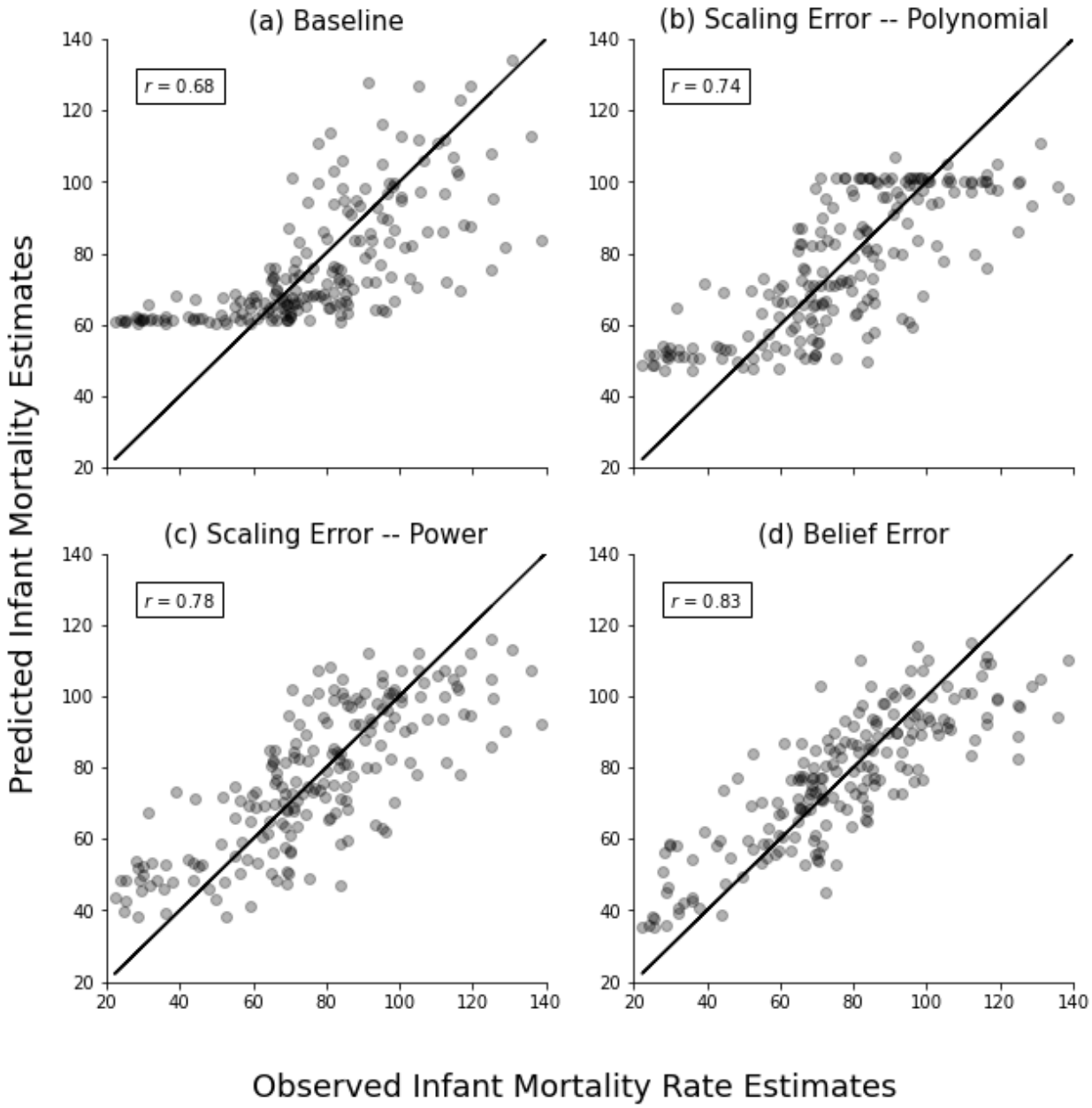


Figure 4. Scatterplots of predicted estimates by (a) baseline model (Eq. 1), (b) scaling error model with polynomial scaling function (Eq. 2), (c) scaling error model with power scaling function (Eq. 3), (d) belief error model (Eq. 4) using leave-one-out cross-validation (LOOCV) vs. actual participant estimates in Study 2 (infant mortality), along with Pearson correlations.

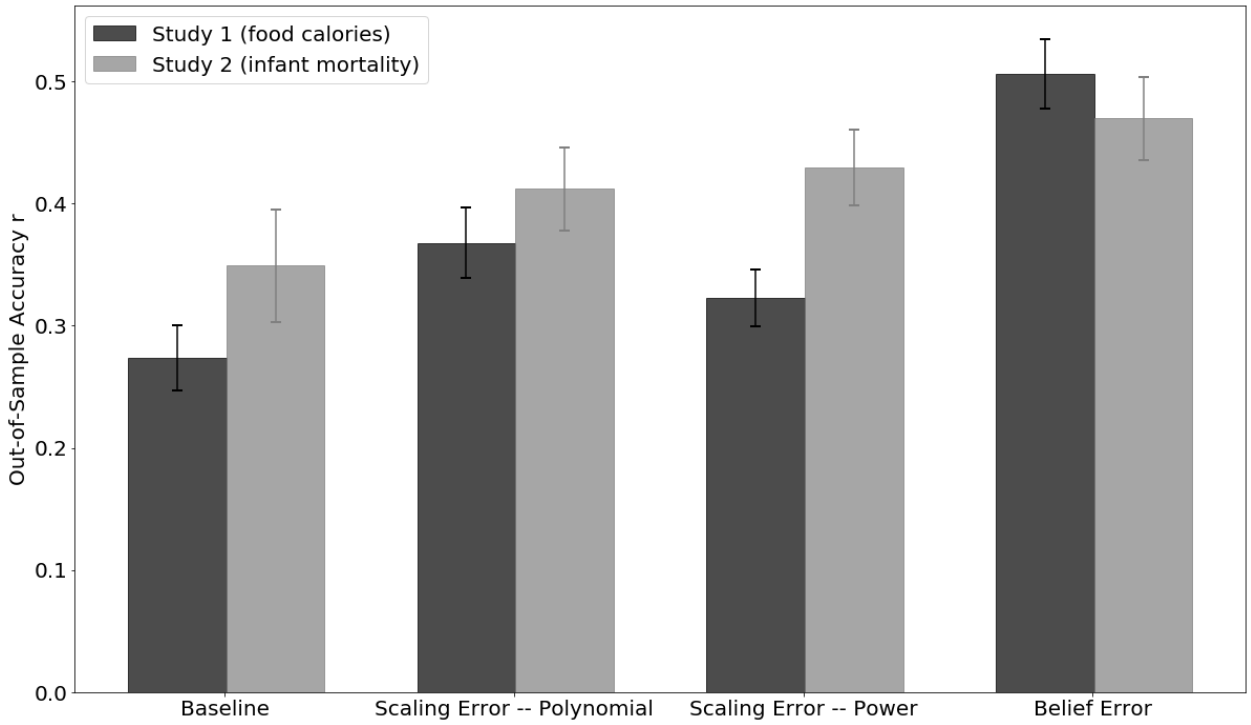


Figure 5. Bar graph of individual-level out-of-sample accuracy rates averaged across participants in Study 1 (food calories) and Study 2 (infant mortality). Error bars represent standard errors.

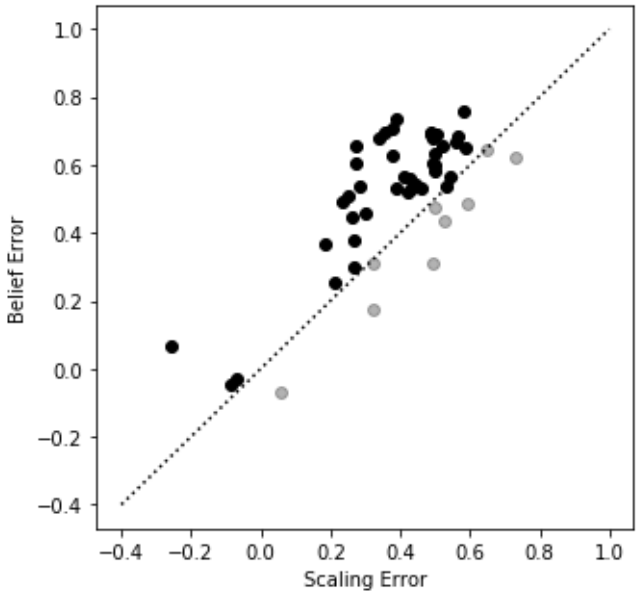
Individual-level fits. We obtained similar results on the individual level. For each participant, we tested the baseline model, the two scaling error models, and the belief error model using the LOOCV procedure. Figure 5 shows the individual level performance of all models. For calorie estimates, the best-fitting scaling error model – a third-degree polynomial – achieved an average individual-level out-of-sample accuracy rate of $r = 0.37$, while the belief error model achieved 0.51. For infant mortality rate estimates, the best-fitting scaling error model – a power function – achieved an average individual-level out-of-sample accuracy rate of $r = 0.43$, while the belief error model achieved 0.47. The baseline models performed poorly in both domains, achieving accuracy rates of 0.27 and 0.35 in the two studies. Although the average individual-level out-of-sample accuracy rates were lower than the corresponding aggregate level accuracy rates, the average individual-level accuracy rates were higher for the belief error model than the best-fitting scaling model. Two separate paired t-tests using participant-level

correlations for the best scaling error models and the belief error models showed that the difference of accuracy rates was significant for both calorie estimates ($df = 49$, $t = -6.72$, $p < 10^{-7}$) and infant mortality rate estimates ($df = 50$, $t = -3.54$, $p < 10^{-3}$). Both belief and scaling error models outperformed the baseline model in the two domains (all $p < 10^{-4}$).

Although the belief error model was the best at explaining participant estimates, there may be differences across individuals in the degrees to which they show the two types of errors. To understand individual heterogeneity in scaling errors and belief errors, Figure 6 plots the individual level out-of-sample accuracy rates (in terms of correlations) of the best-fitting belief error model against the best-fitting scaling error model, for each participant. Each point in Figure 6 represents an individual. Points lying above the diagonal represent individuals whose predictions were better fit by belief error models than scaling error models. Out of 50 participants who estimated food calories, 41 (82%) were better predicted by the belief error model. Out of 51 participants who estimated infant mortality rates, 35 (68%) were better predicted by the belief error model. It is useful to note that the accuracy rates of the two models are positively correlated across participants, indicating that participants that are well described by a scaling error model can also be well described by a belief error model.

As we fit our belief error model to individual-level data, we obtained unique weight vectors for each participant. As discussed previously in section ‘Computational Methods’, these weight vectors represent how much individual participants rely on different dimensions of the vector representations (i.e. different knowledge cues). If participants used semantic knowledge of the judgment targets similarly to each other, the individual weight vectors should correlate with each other. The average pairwise correlation was 0.32 for calorie estimates and 0.26 for infant mortality, across participants. Roughly 6% of the pairwise participant weight vector correlations

(a)



(b)

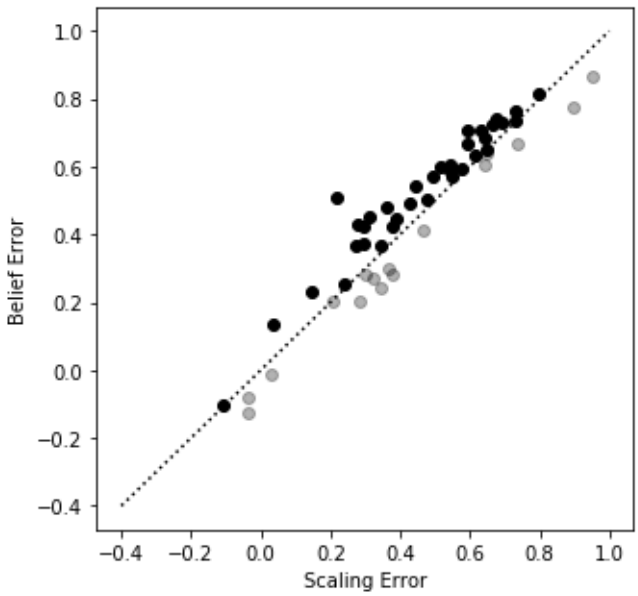


Figure 6. Scatterplots of individual-level out-of-sample correlations for belief error model vs. scaling error model in (a) Study 1 (food calories) and (b) Study 2 (infant mortality). Each dot represents a participant. Black dots represent participants whose responses are better explained by the belief error model than the scaling error model.

were negative for calorie estimates and roughly 8% were negative for infant mortality. In addition, we found that participants whose weight vectors were largely negatively correlated with others tended to give small estimates. The overall low and even occasionally negative

correlations between individual weight vectors suggest that participants varied a lot in relying on knowledge to estimate food calories and infant mortality rates.

A joint modeling approach. So far, our results show a considerable amount of improvement in predictive accuracy when using only the belief error models compared to only the scaling error models on both the aggregate and the individual level. However, it is not clear whether or not belief errors are the only cause of incorrect participant estimates. It could be that participants display both belief and scaling errors, by over or underweighting certain cues to form incorrect beliefs, and then scaling the beliefs nonlinearly to the response scale. Indeed, the scatterplots in Figures 3 and 4 suggest that the belief error model overestimated the values of low-value targets and underestimated the values of high-value targets, consistent with an inverse-S shaped weighting function used in scaling error models. Thus, it could be possible to improve the predictions of the belief error model by allowing its predictions to pass through some sort of non-linear scaling error model.

The ideal test of this would be to embed the linear function in Eq. 4 within the non-linear functions in Eq. 2 or 3 and then recover the weighting vectors and power or polynomial function parameters using best fits to participant data. Such a test is unfortunately not feasible, as it would involve fitting a high-dimensional non-linear weight vector to a relatively small (200 observation) dataset. But we did attempt a shortcut, which simplifies the estimation by combining the belief and scaling error models sequentially. Specifically, we first trained the belief error model on participant estimates of the 199 judgment targets in the training data, y_i, and obtained predictions for the training targets (call these belief-error-corrected estimates, z_i'). Then we trained a second scaling error model to predict participant estimates of these training targets, y_i, from the belief-error-corrected estimates, z_i'. Finally, we used the trained belief error

model to get a belief-error-corrected estimate for the left-out target and passed it through the trained scaling error model to get a final prediction for the left-out target. We repeated this procedure for all 200 targets in both studies to get LOOCV predictions for the joint model. Drawing on the results in the previous sections, we used the third-degree polynomial for food calories and power function for infant mortality rates for our scaling models.

Figure 7 shows scatterplots of participant estimates against belief-error-corrected estimates. The aggregate level out-of-sample accuracy, measured by the correlation between average participant estimates and the LOOCV predictions, was 0.84 for food calories and 0.83 for infant mortality rates. Although these are both significantly different from a correlation of 0 ($p < 10^{-50}$ for both), they are nearly identical to the belief error models alone (0.83 for both) and much higher than the best-fitting scaling error models alone (0.61 for food calories and 0.78 for infant mortality rates). Figure 7 also shows the best-fitting scaling error models used in the joint models. Although these models do permit non-linearity, they appear to be almost linear within the range of the stimuli.

Overall, these results suggest that our joint modeling of belief and scaling errors does not significantly improve predictions, largely because belief-error-corrected estimates z_i' have a linear relationship with participant estimates y_i. This in turn indicates that what seem to be scaling errors in Figures 2-4 could simply be statistical artifacts. Of course, we cannot make conclusive claims based on our fits as they involve a (potentially biased) shortcut to fitting the joint model. It could be that the joint model does improve predictive accuracy if its belief and scaling error components are fit simultaneously and not sequentially.

Explaining scaling errors with belief errors. The above section suggests that a joint belief and scaling error model does not significantly improve predictions relative to a belief-

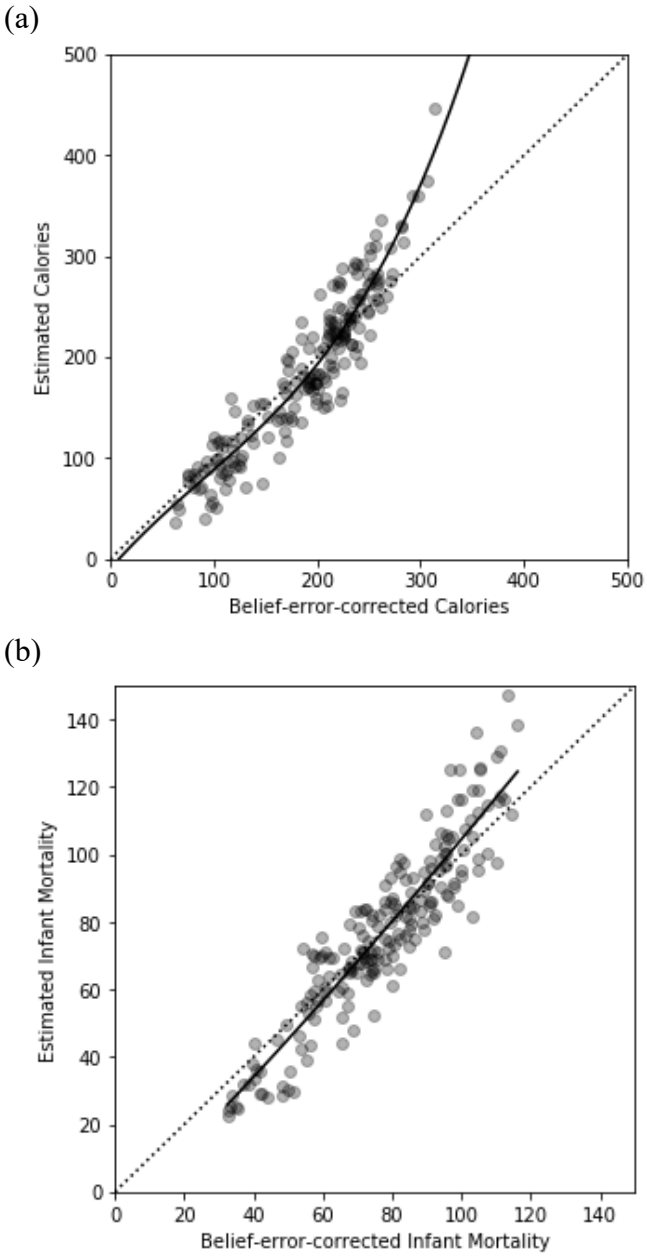


Figure 7. Scatterplots of average participant estimates (y_i) vs. belief-error-corrected estimates (z_i') in (a) Study 1 (food calories) and (b) Study 2 (infant mortality). The solid curves represent the best-fitting scaling error models used in the joint model – polynomial (Eq. 2) for food calories and power (Eq. 3) for infant mortality rates.

error-only model. Additionally, our participant-level fits show that participants that were well described by the scaling error models were also well described by the belief error models. This suggests that assuming belief errors (without assuming scaling errors) may be enough to appropriately model our behavioral data. Here we test one implication of this claim: If belief

errors are the sole cause of judgment errors it may be possible to mimic scaling errors (such as the overestimation of small values) with a belief error model alone.

In Figure 8 we plot the LOOCV prediction of the belief error model on the aggregate data for each judgment target, against the correct value for the judgment target. Recall that Figure 2, which presents similar scatterplots of participant estimates vs. correct values, shows that participants overestimated low calories, underestimated high calories, and overestimated infant mortality rates. Although these biases can be due to scaling errors, in Figure 8 we can see that our belief error model generates the same non-linearities as human participants. Indeed, in Figure 8, we can see polynomial and power scaling functions best capture the relationship between our belief error model's predictions and correct responses for Study 1 and 2 respectively. These results provide further evidence that the belief error model is capable of describing the qualitative and quantitative patterns underpinning judgment errors in naturalistic estimation tasks.

Interpretation of weight vectors. To better understand the weight vectors underlying our belief error model, we plot the dimensional values of these vectors (trained on aggregate-level data) for the two studies. Figure 9 shows the distribution of weights across 300 dimensions. The standard deviation across the dimensions of the weight vectors was 22.78 for food calories and 6.89 for infant mortality. Since the standard deviation is not a standardized measure, we also computed the Gini coefficients of the weight vectors, which were 0.2 for both tasks. These low Gini coefficients indicate a high degree of dispersion, suggesting that different dimensions contribute unevenly to numerical judgments.

Next, we correlated the two weight vectors trained on aggregate level data for the two studies. As shown in Figure 10, the correlation was minimal and insignificant (r = -0.09, p =

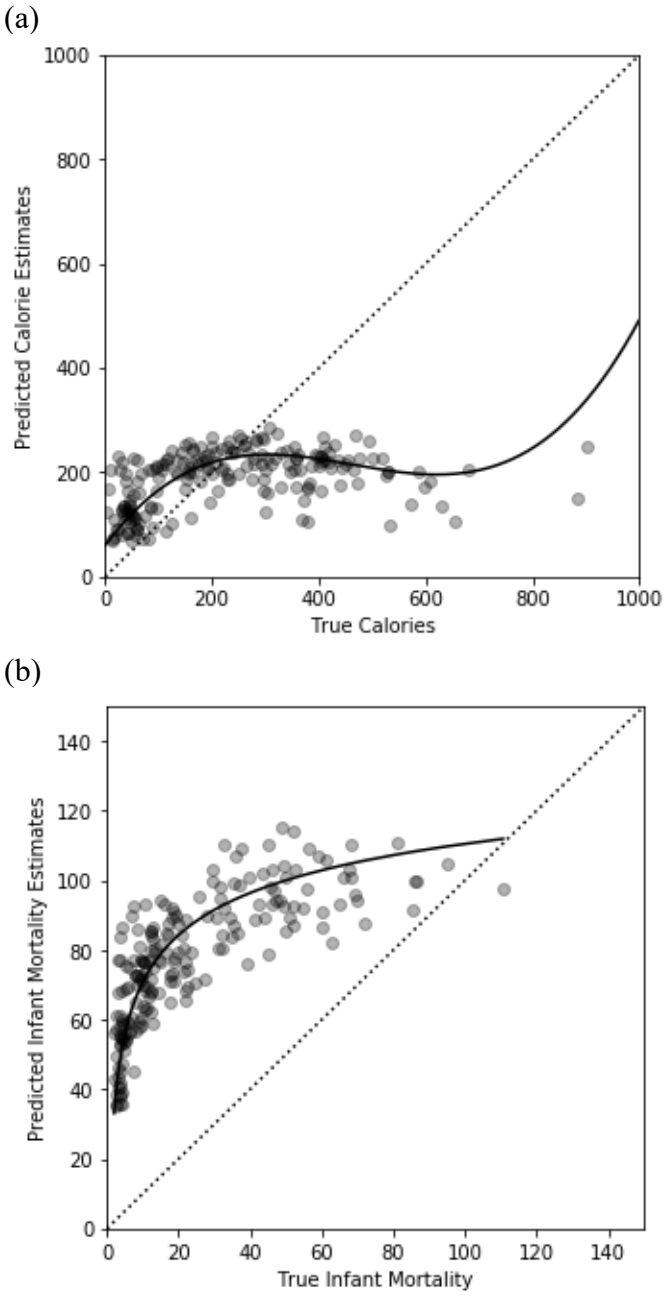
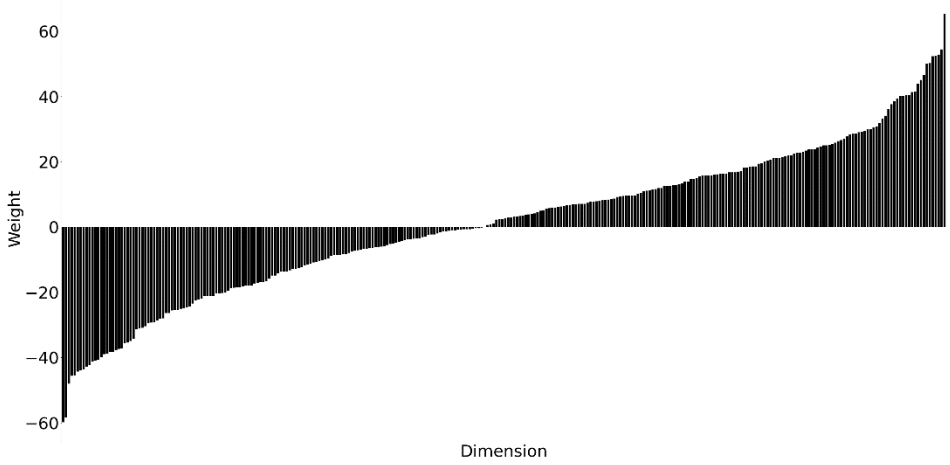


Figure 8. Scatterplots of LOOCV predicted estimates of belief error model (Eq. 4) vs. actual values in (a) Study 1 (food calories) and (b) Study 2 (infant mortality). The solid curves represent the best-fitting scaling error model trained on aggregate-level data.

0.13), indicating that the weight vectors for the two judgment dimensions were distinct. We also did more robust tests. First, we trained a single weight vector across both domains. The pooled model performed worse than individual models ($r = 0.81$, $p < 10^{-47}$ for food calories, and $r = 0.6$,

(a)



(b)

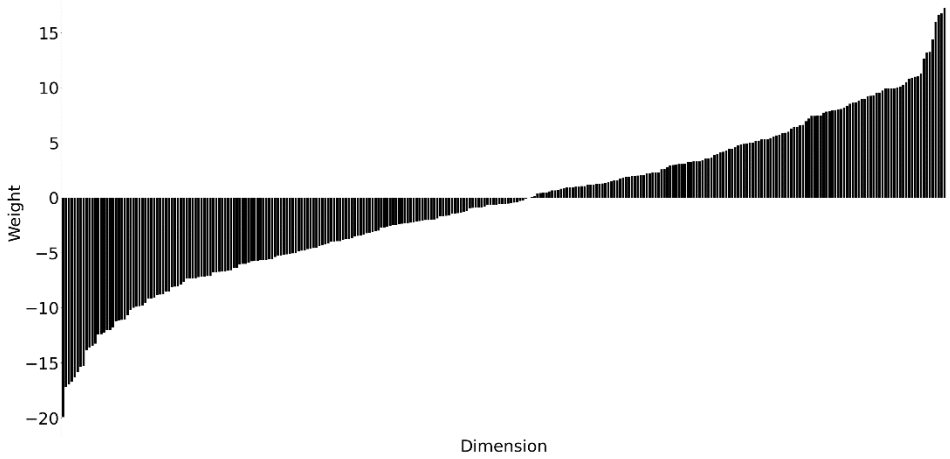


Figure 9. Dimensional values (sorted from lowest to highest) of the best-fit weight vector trained on aggregate-level data for (a) Study 1 (food calories) and (b) Study 2 (infant mortality).

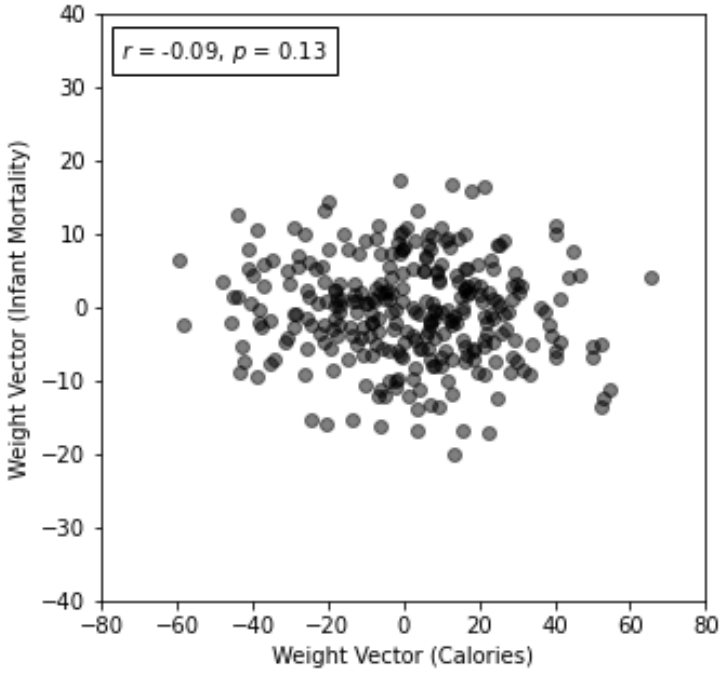


Figure 10. Scatterplot of dimensional values of the best-fit weight vector trained on aggregate infant mortality estimates vs. those trained on aggregate calorie estimates. Each dot represents a dimension.

$p < 10^{-20}$ for infant mortality rates). Second, we predicted aggregate participant response in one task with the model trained on the other task. The weight vector trained on infant mortality predicted calorie estimates with an accuracy rate (correlation between predicted and observed) of -0.06 ($p = 0.38$) and the weight vector trained on food calories predicted mortality estimates with an accuracy rate of -0.41 ($p < 10^{-8}$). Together, these results suggest that people use different weights for different domains.

Conceptual underpinnings of judgment. The weight vector ***w*** obtained from the belief error model in the above tests is also a 300-dimensional vector in the semantic space. Implicitly, this vector corresponds to the point in the semantic space that would generate the highest participant estimate, and concepts closest to this point are ones that participants associate most with the criterion value. The reason for this is that all our vectors are normalized (have a magnitude of one), and thus the dot product of these semantic vectors with the weight vector,

$\boldsymbol{w \cdot x_i}$, is a simple linear transformation of the cosine similarity of the semantic vectors and the weight vector, $\frac{\boldsymbol{w \cdot x_i}}{\|\boldsymbol{w}\|\|\boldsymbol{x_i}\|} = \frac{\boldsymbol{w \cdot x_i}}{\|\boldsymbol{w}\|}$. The dot product is the main determinant of participant estimates, and cosine similarity is a measure of relatedness and association in semantic spaces (Bhatia, 2017; Bhatia et al., 2019; Jones et al., 2015; Landauer & Dumais, 1997; Mandera et al., 2017; Mikolov et al., 2013). This relationship implies that we can use the weight vector to explore the conceptual underpinning of participant estimates. Crucially, this analysis is not limited to the food items or countries used in our studies. We can take any word or concept with a vector representation in our semantic space, measure its dot product with the best-fitting weight vector. Words or concepts with high dot products are also those with high cosine similarities, and thus will be judged by participants to be especially related to or associated with the judgment dimension. These words and concepts are also likely the ones that come to participants' minds while deliberating.

To find concepts that map strongly onto the weight vector, we took the 5,000 most frequent words from the corpus of contemporary American English (http://corpus.byu.edu/coca/) that were not judgment targets. For each word *k*, we obtained a 300-dimensional (normalized) vector, $\boldsymbol{s_k}$, from the Word2Vec model, by querying the Word2Vec vocabulary for the lower-cased representation of the word. By multiplying $\boldsymbol{s_k}$ with the weight vector $\boldsymbol{w}$ (obtained from model fits on the entire aggregate participant estimates in Studies 1 and 2) we got the predicted association of word *k* with the judgment dimension, $e_k = \boldsymbol{w \cdot s_k}$. Figures 11a and 11b present word clouds of 50 words that are most and least strongly associated with the judgment dimensions according to this measure (i.e. having the largest or smallest values of e_k). These words reveal potential associative underpinnings of judgment errors. For example, Figures 11a and 11b show that words related to feasting (e.g. restaurant, Thanksgiving, fat, and dinner) are strongly associated with high-calorie

foods; words related to nature and nutrition (e.g. crop, leaf, nutrient) are strongly associated with low-calorie foods; words related to bad living conditions (e.g. hunger, poverty, and AIDS) are strongly associated with countries with high infant mortality rates; and words related to tourism and industry (e.g. sail, cruise, and manufacturer) are strongly associated with countries with low infant mortality rates. These words and concepts will be the basis of our analysis in Studies 3-8.

Discussion

We built computational models to compare two types of errors in numerical estimation: scaling errors and belief errors. We used these models to predict naturalistic numerical estimates in two studies involving judgments of calories of food items and judgments of infant mortality rates of countries. Consistent with previous findings, the best-fitting scaling error models in both studies were nonlinear and drastically outperformed a simple linear baseline. We found the common inverse-S-shape pattern in food calorie estimation. Additionally, although almost all countries' infant mortality rates were overestimated, the magnitude of overestimation appeared to be larger for countries with low infant mortality than for countries with high infant mortality.

Although our scaling error models were able to provide a good account of our data, we obtained even higher out-of-sample predictive accuracy rates from our belief error models. This was the case on both the aggregate and individual level, indicating that we can predict participant estimates better if we assume some flexible (potentially incorrect) use of knowledge cues to form beliefs, instead of some flexible (potentially incorrect) use of correct beliefs. Moreover, using a simplified joint modeling approach, we showed that after accounting for belief errors, what originally appeared to be scaling errors largely disappeared. Indeed, the belief error models were also able to mimic the inverse-S shaped patterns observe in our behavioral data, which are often considered to be caused by scaling errors. Thus, it appears that belief errors can mimic scaling

errors if the belief error model is appropriately specified. This in turn suggests that judgment errors may stem primarily from the incorrect use of knowledge, rather than the incorrect scaling of the true beliefs. Of course, this does not mean that there is no explanatory value to scaling errors, as both belief errors and scaling errors could be operating simultaneously. Additionally, two types of errors may influence numerical estimation differently depending on the task at hand. For example, the improvement of the belief error models over the best-fitting error models was much smaller in infant mortality estimation (0.05 difference in out-of-sample correlation on the aggregate level, 0.04 on the individual level) than in food calorie estimation (0.22 difference in out-of-sample correlation on the aggregate level, 0.16 on the individual level), suggesting that belief errors are less prominent (relative to scaling errors) in infant mortality estimation than in food calorie estimation.

Finally, the best-fitting weight vectors to our aggregate participant estimates were used to identify the conceptual underpinnings of participant judgments. As would be expected, we found that feast and fat related words were associated with high calories; nature and health related words were associated with low calories; hunger and disease related words were associated with high mortality rates; and tourism and industry related words were associated with low mortality rates. These results offer valuable insights regarding the psychological substrates of numerical estimation, and suggest one reason why participants may be making judgment errors in our two studies. They also offer precise quantitative predictions regarding the strength of participant associations with the judgment dimension in these studies. These predictions can be tested by eliciting relatedness ratings of these concepts with the high and low ends of the food calorie and infant mortality dimensions. This is the goal of Studies 3 and 4.

Figure 11. Word clouds of words that are predicted by belief error models as strongly associated with high and low calories based on Study 1 data (a) and high and low infant mortality rates based on Study 2 data (b). Word font corresponds to the magnitude of predicted association (i.e. dot product of word vector and weight vector, e_k). Word clouds of word associates that listed most frequently under high and low conditions in (c) Study 5 (food calories) and (d) Study 6 (infant mortality). Word font corresponds to the frequency of a word being listed.

Studies 3 and 4: Semantic Relatedness Judgements

Experimental Methods

Participants. We recruited a total 50 participants (mean age = 18 years, 34% female) in Study 3 and 51 participants (mean age = 18 years, 39% female) in Study 4 from Prolific Academic. All participants were from the United States and had an approval rate of 80% or above.

Stimuli. For Study 3, we selected concepts predicted by our best-fitting belief error model (fit to aggregate data from Study 1) to be strongly related to high- and low-calorie foods. Specifically, these were the ten concepts with the largest values of e_k (leaf, garden, flower, bulb, cotton, lawn, sunlight, agricultural, nutrient) and the ten concepts with the smallest values of e_k (fat, famous, recipe, restaurant, monster, chef, heaven, birthday, cooking, beast). Similarly, for Study 4, we selected concepts predicted by our best-fitting belief error model (fit to aggregate data from Study 2) to be strongly related to high and low infant mortality rates. These were the ten concepts with the largest values of e_k (bush, neighborhood, rice, pastor, homeless, poverty, hunger, AIDS, African, hungry) and ten concepts with the smallest values e_k (European, extensive, sail, manufacturer, cruise, throughout, maker, strict, leading, bow).

Procedure. Participants were asked to rate the relatedness of a concept to the judgment dimension on a slider from -100 (strongly related to the low end of the judgment dimension) to 100 strongly related to the high end of the judgment dimension). In Study 3, the prompt read "How related do you feel [*concept*] is to low- or high-calorie food?" In Study 4, the prompt read "How related do you feel [*concept*] is to countries with low or high infant mortality rates?". Each participant rated all 20 concepts and saw only one concept on each screen. The order of the 20 concepts was randomized.

Results

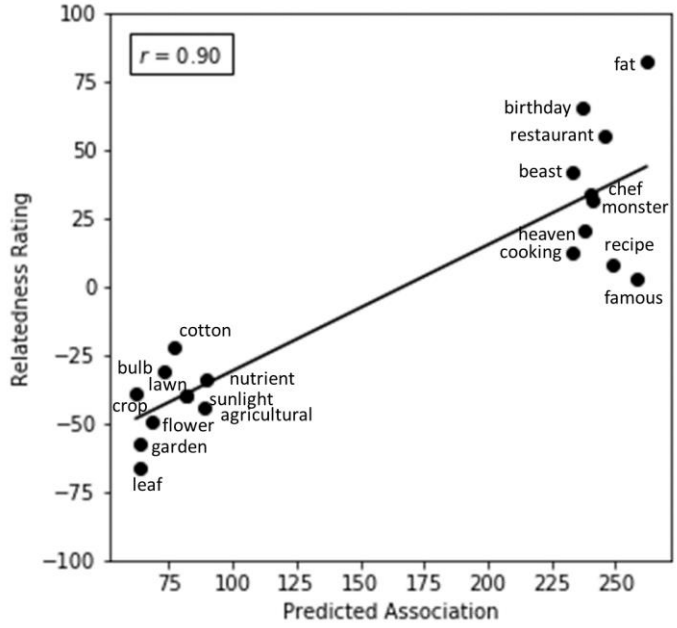
If our belief error model captures the conceptual underpinnings of judgment errors, then the concepts identified by the model as having large values of e_k should be given higher ratings by participants than the concepts identified by the model as having small values of e_k. Figure 12 shows scatterplots of average participant relatedness ratings and the predicted associations, e_k. Each dot represents a concept. As can be seen in this figure concepts with high predicted associations also typically have high relatedness ratings. Overall, the correlations between average relatedness ratings and e_k are 0.90 for the 20 food calorie concepts and 0.73 for the 20 infant mortality concepts. These are both significantly different to zero ($p < 10^{-7}$ and $p < 10^{-3}$ respectively). The Spearman correlations are 0.78 ($p < 10^{-4}$) for food calories and 0.76 ($p < 10^{-4}$) for infant mortality rate.

Additionally, in Study 3, the ten concepts with high values of e_k had an average relatedness rating of 35.40, and the ten concepts with low values of e_k had an average relatedness rating of -42.23. Similarly, in Study 4 concepts with high values of e_k had an average relatedness rating of 30.70, and concepts with low values of e_k had an average relatedness rating of -19.69. These ratings are significantly different when evaluated using a t-test ($df = 8$, $t = 8.43$, $p < 10^{-5}$ for Study 3; $df = 8$, $t = 3.92$, $p < 10^{-2}$ for Study 4). Finally, 76% of participants in Study 3 and 62% of participants in Study 4 gave positive relatedness ratings to all concepts with high values of e_k and negative ratings to all concepts with low values of e_k, providing additional evidence for our predictions.

Discussion

We tested the ability of the belief error model to predict the concepts that are most related to or associated with the judgment dimension. For this purpose, we used the association measure, e_k, obtained by taking the dot product between a concept vector and the best-fitting weight

(a)



(b)

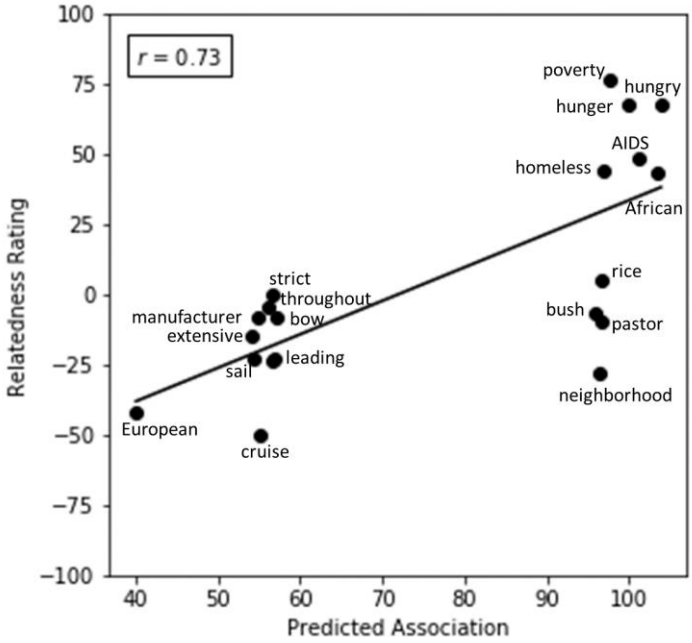


Figure 12. Scatterplots of average participant relatedness rating vs. predicted association in (a) Study 3 (food calories) and (b) Study 4 (infant mortality). Each dot represents a unique word

associate predicted by belief error models from Study 1 (food calories) and Study 2 (infant mortality).

vectors from participant data in Studies 1 and 2. On the aggregate level, we found that e_k is highly correlated with participant ratings, so that concepts considered by our model to be positively or negatively related to food calories (in Study 3) and infant mortality rates (in Study 4) are indeed given high or low relatedness ratings by participants.

This result suggests that concepts with very high values of our association measure e_k could be highly activated in the minds of participants making judgments involving high-calorie foods or high infant mortality rates. Similarly, concepts with very low values of e_k could be highly activated in the minds of participants making judgments involving low-calorie foods or high infant mortality rates. These are predictions about the memory processes involved in numerical judgment, which can be tested using a free association task. We do so in Studies 5 and 6.

Studies 5 and 6: Free Association

Experimental Methods

Participants. We recruited a total of 200 participants (mean age = 32 years, 34% female) in Study 5 and 201 participants (mean age = 31 years, 50% female) in Study 6 from Prolific Academic. All participants were from the United States and had an approval rate of 80% or above.

Procedure. As the predicted association, e_k, specifies both the direction and the magnitude of association, in each study, we examined two conditions – high and low. In the high condition, participants were asked to list ten different words that first come to their mind when they think of high-calorie foods (Study 5) or countries with high infant mortality rates (Study 6). In the low condition, participants were asked to do the same thing but when they think of low-

calorie foods (Study 5) or countries with low infant mortality rates (Study 6). Participants were randomly assigned to either a high or low condition and entered all ten words at once in a list of ten textboxes. The sample prompt in the low condition in Study 5 read "Please list 10 different words that first come to your mind when you think of low-calorie food. These could be any words, including words describing actual food, words describing sensations, emotions and feelings, words corresponding to real-world objects, people and places, or words describing abstract concepts." The prompts for the high condition in Study 5 and the prompts in Study 6 were similar to this.

Results

With ten listed words per participant per condition, we obtained a set of 1,010 (nonunique) word associates for high-calorie foods, 990 for low-calorie foods, 1,010 for countries with high infant mortality rates, and 1,000 for countries with low infant mortality rates. Figures 11c and 11d show word clouds of the 50 most frequently listed word associates in the high and low conditions for the two domains.

138 words in Study 5 and 142 words in Study 6 were listed at least three times and also had a vector representation in the Word2Vec vocabulary. For each of these words k, we obtained a normalized 300-dimensional vector, $\boldsymbol{s_k}$, from the Word2Vec model (note that we used a word frequency cut-off of three as highly infrequent words are likely to give very extreme and spurious association measures). Then using the approach outlined above, we calculated the value of $e_k = \boldsymbol{w} \cdot \boldsymbol{s_k}$, where $\boldsymbol{w}$ was the weight vector from the belief error model trained on the aggregate participant data in Study 1 or Study 2. Again, e_k measures how much word k is associated with the criterion. If our belief error models are able to identify word associations, then we would

expect e_k to correlate with relative word frequencies in the high vs. low conditions in our free association studies.

We measured these relative word frequencies using Eq. 6, where H_k is the number of times word k was listed in the high condition, L_k is the number of times word k was listed in the low condition, H_N is the total number of different words listed in the high condition, and L_N is the total number of different words listed in the low condition. Implicitly, this equation gives us a measure of the relative strength of association of a given word with high-calorie foods or countries with high infant mortality rates, as revealed by our participant free association data. We write the relative word frequency of word k as r_k.

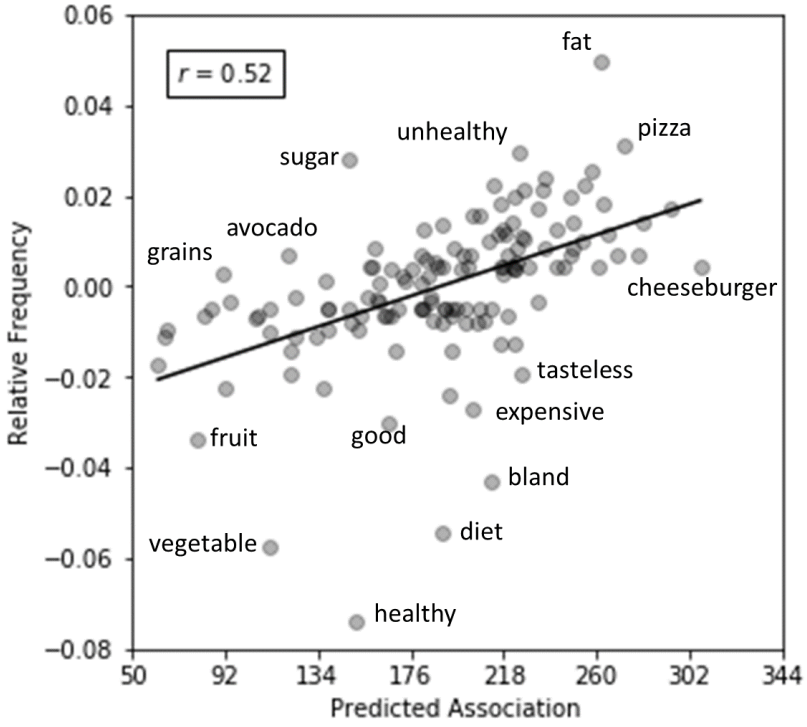
$$r_k = \frac{H_k}{H_N} - \frac{L_k}{L_N} \tag{6}$$

If the associations revealed by the best-fitting weight vectors in Study 1 and 2 capture the associations that participants have with the criterion variable, we would expect to observe a positive relationship between the relative frequency variable, r_k, and the predicted association, e_k, for the words listed by participants in Studies 5 and 6. We do observe this positive relationship. As shown in Figure 13, the Pearson correlations between relative frequency and predicted association are 0.52 ($p < 10^{-10}$) for food calories and 0.36 ($p < 10^{-4}$) for infant mortality rates. Spearman's rank correlations are likewise 0.62 ($p < 10^{-15}$) for food calories and 0.39 ($p < 10^{-5}$) for infant mortality rates.

To understand the dynamics of participant word association, we also calculated the predicted associations e_k for the ten listed word associates for each participant, and for each serial position, we averaged the predicted association across participants. If participants listed word associates in descending order of semantic relatedness to the judgment dimension, i.e. high-/low-calorie foods or countries with high/low infant mortality rates, then we should expect

to see the average predicted association with the judgment dimension (e_k) decreases with serial position in the high condition and increases in the low condition. Figure 14 illustrates these trends. Consistent with this figure, a linear regression reveals that predicted association e_k has a significant positive relationship with serial position in the low condition ($\beta = 1.40$, $t = 3.17$, $p =$

(a)



(b)

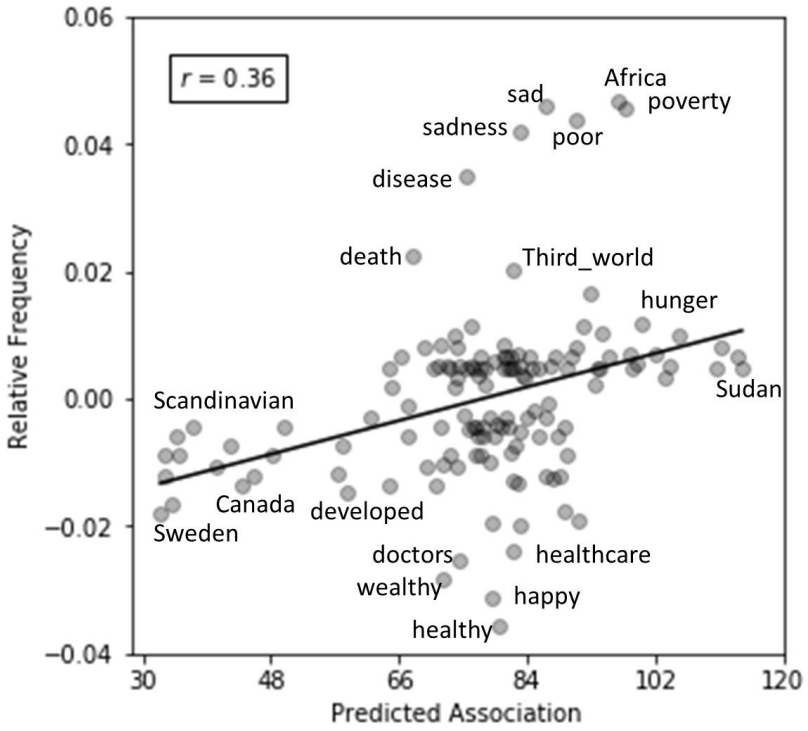
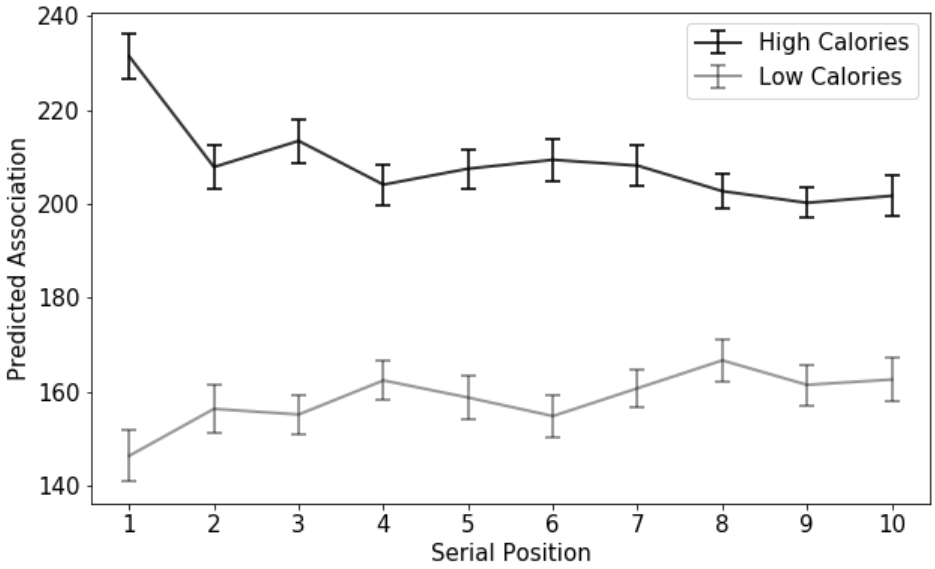


Figure 13. Scatterplots of relative frequency vs. predicted association in (a) Study 5 (food calories) and (b) Study 6 (infant mortality). Each dot represents a unique word associate listed by participants. Some word associates are labeled for illustration purpose.

(a)



(b)

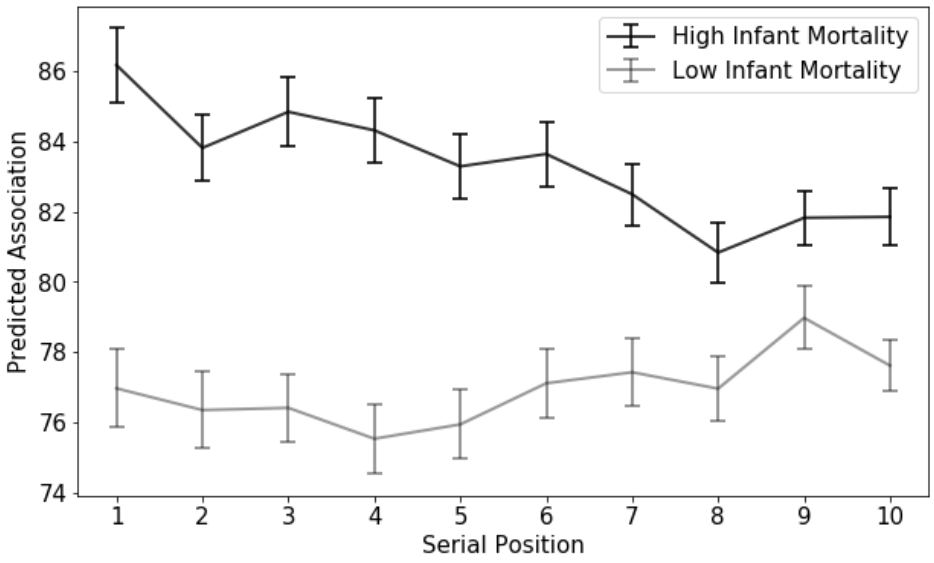


Figure 14. Serial position curves for (a) Study 5 (food calories) and (b) Study 6 (infant mortality). The y-axis is the average predicted association, e_k, for all the words being listed at the given serial position (1 for listed first to 10 for listed last). Error bars represent standard errors.

0.013, 95% CI [0.38, 2.42] for Study 5; $\beta = 0.21$, $t = 2.35$, $p = 0.047$, 95% CI [0.004, 0.406] for Study 6) and a significant negative relationship with serial position in the high condition ($\beta = -2.18$, $t = -3.11$, $p = 0.015$, 95% CI [-3.81, -0.56] for Study 5; $\beta = -0.47$, $t = -5.646$, $p < 10^{-3}$, 95% CI [-0.67, -0.28] for Study 6).

Discussion

Our belief error models are able to make predictions regarding the associative underpinnings of judgment. Implicitly, these models identify regions of the semantic space that are most associated with the criterion variable. Words and concepts in these regions should be the ones that come to people's minds as they deliberate. In Studies 5 and 6 we tested these predictions using a free association task. We found that the frequency with which participants listed a given word in the high vs. low judgment prompt conditions correlated with the association of that word with the judgment prompt inferred using the best-fitting belief error models in Studies 1 and 2. Additionally, similar to past memory research that suggests recall

outputs follow descending order of memory strength (Barnhardt, Choi, Gerkens, & Smith, 2006; Gillund & Shiffrin, 1984; Jou, 2008), we found that participants listed words in descending order of semantic relatedness to the weight vector of the belief error models. Thus our belief error models could predict not only the relative frequencies of words in our free association tasks, but also memory dynamics in these free association tasks. This provides further support in favor of belief error models for studying the cognitive processes and mental representations underlying numerical estimation.

However, since there was no numerical estimation task involved in Studies 5 and 6, it is an open question whether participants would actually think of these words as they generate estimates. Therefore, in Studies 7 and 8, we used Ericsson-Simon verbal protocols (Ericsson & Simon, 1980) that ask participants to report thought process after an estimation task to verify the results of Studies 5 and 6.

Studies 7 and 8: Verbal Protocols

Experimental Methods

Participants. We recruited a total of 48 participants (mean age = 32 years, 58% female) in Study 7 and 52 participants (mean age = 33 years, 46% female) in Study 8 from Prolific Academic. All participants were from the United States and had an approval rate of 80% or above.

Procedure. Similar to Studies 5 and 6, in each study, we examined two conditions – high and low. In the high condition, participants were asked to estimate the calories of two foods – lard and cheeseburger (Study 7) or the infant mortality rates in two countries – Ethiopia and Haiti (Study 8). In the low condition, participants were asked to do the same but for tea and lettuce (Study 7) or Sweden and Japan (Study 8). These items were selected because they either received

the highest or lowest average participants estimates in Studies 1 and 2. In the first part of the study, participants were asked to perform the same estimation task as in Studies 1 and 2. Each participant gave estimates for all four stimuli. The order of the two conditions was randomized and so was the order of the two stimuli within the same condition. In the second part of the study, participants were shown their previous answers and were asked to describe their thought process when they gave estimates in each of the two conditions. The sample prompt in the low condition in Study 7 read "You estimated tea to have [*value*] calories per 100g and lettuce to have [*value*] calories per 100g. Please briefly describe how you estimated these foods." The prompts for the high condition in Study 7 and the prompts in Study 8 were similar to this. The order of the two conditions was also randomized.

Results

The estimates in Studies 7 and 8 were not significantly different to those in Studies 1 and 2. In Study 7, the average number of words reported was 18.50 (s = 13.02) in the low condition and 21.73 (s = 17.81) in the high condition. In Study 8, the average number of words reported was 13.98 (s = 8.18) in the low condition and 14.77 (s = 8.70) in the high condition. There was no significant difference in the word count between the two conditions in either study (df = 94, t = -1.00, p = 0.32 for Study 7; df = 102, t = -0.47, p = 0.64 for Study 8). After removing stop words[2] and excluding words mentioned less than three times in either condition, we obtained a total of 56 unique words in Study 7 and 53 in Study 8. Following the same analysis done in Studies 5 and 6, for each of these words k, we obtained a normalized 300-dimensional vector, $\boldsymbol{s_k}$, from the Word2Vec model. Next, we computed the two measures – predicted association, e_k =

[2] We used the stop word library from NLTK toolkit in Python. Other data cleaning procedures (e.g. stemming, lemmatizing, auto-correction) were also done by NLTK package.

$\boldsymbol{w}\cdot\boldsymbol{s_k}$, where $\boldsymbol{w}$ was the weight vector from the belief error model trained on the aggregate participant data in Study 1 or Study 2; and the relative word frequencies, r_k, obtained from Eq. 6 (where H_k was the number of times word k appeared in the high condition, L_k was the number of times word k appeared in the low condition, H_N was the total number of different words that appeared in the high condition, and L_N was the total number of different words that appeared in the low condition).

Again, e_k measures how much word k is associated with the criterion and r_k measures the relative strength of association of a given word with high-calorie foods or countries with high infant mortality rates. Given the promising results in Studies 3-6, we would still expect to observe a positive relationship between the relative frequency variable, r_k, and the predicted association, e_k. This is indeed the case. As shown in Figure 15, the Pearson correlations between relative frequency and predicted association are 0.74 ($p < 10^{-10}$) for food calories and 0.34 ($p = 0.01$) for infant mortality rates. Spearman's rank correlations are likewise 0.62 ($p < 10^{-6}$) for food calories and 0.20 ($p = 0.16$) for infant mortality rates. These high correlations persist even when we remove the test items (e.g. lettuce, tea, etc.) from the data – in this case, the correlations are

(a)

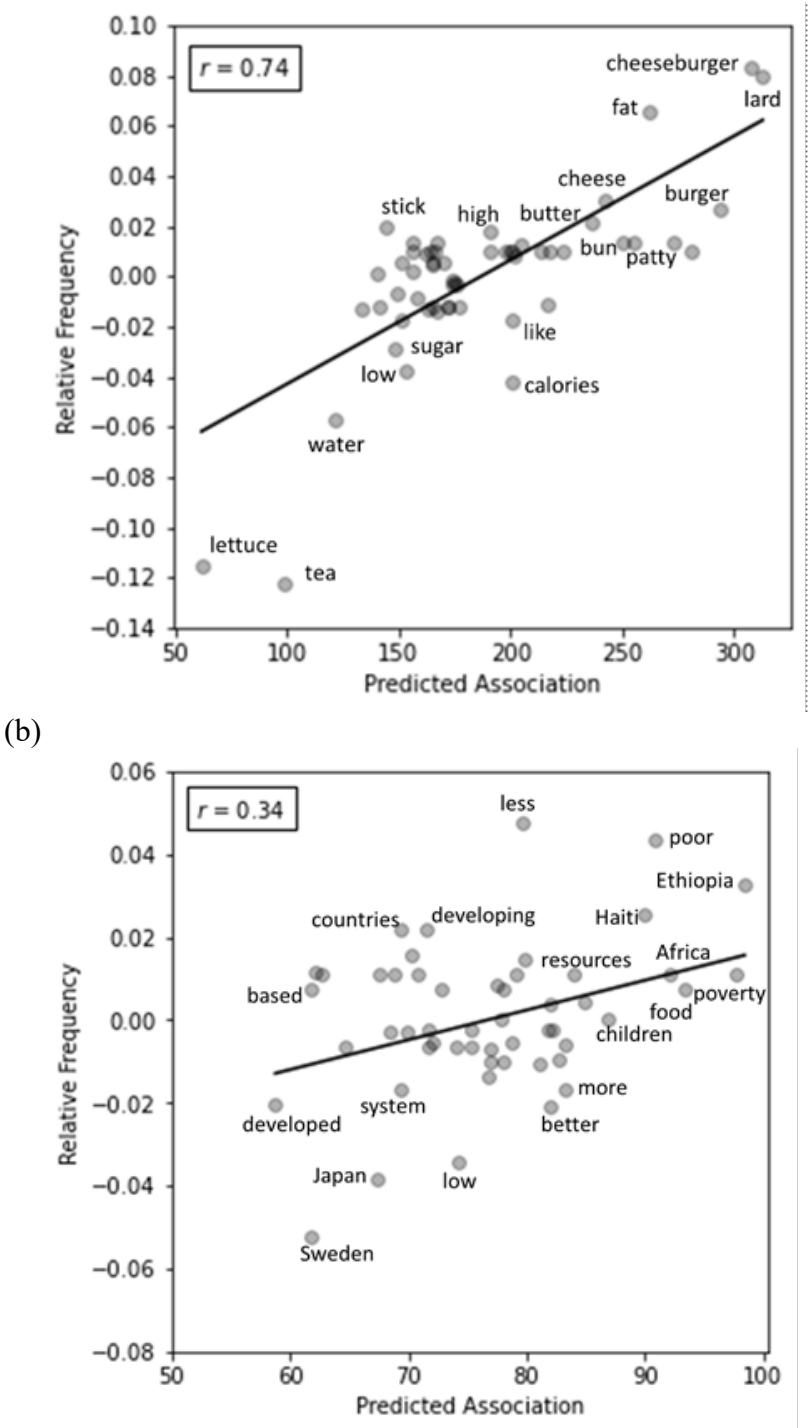


Figure 15. Scatterplots of relative frequency vs. predicted association in (a) Study 7 (food calories) and (b) Study 8 (infant mortality). Each dot represents a unique word associate listed by participants. Some word associates are labeled for illustration purpose.

0.53 ($p < 10^{-4}$) for food calories and 0.34 ($p = 0.05$) for infant mortality rates. Spearman's rank correlations are 0.51 ($p < 10^{-3}$) for food calories and 0.21 ($p = 0.23$) for infant mortality rates.

Discussion

By asking participants to recall thought process right after performing an estimation task involving stimuli with extreme estimates, we were able to recover the mental process underlying numerical estimation. We measured the common words that participants thought of as they estimated numerical quantities for stimuli with typically high or low estimates and tested the ability of the belief error model to predict these words. On the aggregate level, we found that the relative frequency of these words was highly correlated with the association strength predicted by the belief error model. This result suggests that concepts with high predicted association by our model are more likely to be activated during the estimation process.

It is worth noting that the correlations are much higher in the food calories domain compared to infant mortality rates. This is consistent with findings from Studies 1 and 2 that the difference between the accuracy of the best scaling error model and the best belief error model was larger in calorie estimation than in infant mortality estimation. Together, these results indicate that belief errors may be more prominent in calorie estimation than in mortality estimation.

It is also worth noting that the current studies only collected general thought processes for limited stimuli with extreme estimates (two stimuli with high average estimates and two stimuli with low average estimates). Therefore, the words in participants' verbal reports may not fully represent the concepts that would be activated by other foods or countries. However, our results show a promising direction to investigate more nuanced thought processes for different stimuli and different individuals. As shown in Studies 1 and 2, the belief error model provides good individual-level fits, which suggests that the belief error model can generate concepts that are

likely to be activated by different people. If the thought process is collected immediately after each individual estimation task for each participant, the belief error model can further predict individual-level and item-specific psychological substrates of numerical estimation.

General Discussion

Predictive Accuracy

We modeled two types of errors in numerical estimation – scaling errors and belief errors. Scaling errors emerge when participants incorrectly report (otherwise correct) beliefs about the judgment target, whereas the belief errors arise from the incorrect use of knowledge to form these beliefs. In Studies 1 and 2 we collected participant estimates of calories of common food items and infant mortality rates of countries and built scaling error models and belief error models to predict participant estimates. We found that the predictions of belief error models were highly accurate (achieving out-of-sample correlations of 0.83 with aggregate participant estimates), and additionally, significantly outperformed the predictions of scaling error models. We obtained similar results on the individual-level showing that our belief error models are capable of describing the structure of individual heterogeneity in belief formation and numerical judgment. Overall, these results suggest that numerical estimation can be better described as the outcome of an error-prone belief formation process (that correctly maps incorrect beliefs to the response scale) than an error-prone response process (that utilizes correct beliefs but maps them incorrectly onto the response scale).

Of course, we are not suggesting that scaling errors do not have a role in numerical estimation, and it may be the case that certain tasks are more susceptible to scaling errors. For example, the difference between the accuracy of the best scaling error model and the best belief error model was larger in calorie estimation than in infant mortality estimation. Intuitively, model

accuracy reflects the degree to which judgment errors can be predicted by either the incorrect use of knowledge or the incorrect scaling of true beliefs. Therefore, our results suggest that people may exhibit more scaling errors when estimating infant mortality rates than when estimating calories. This could be due to the possibility that the knowledge representations of countries are less biased than those of food, as in a supplementary analysis we found that vector space semantics were able to explain 88% of the variance in true infant mortality rates and 78% of the variance in true calories of food. On the other hand, it could also be that psychological magnitudes of food calories are less distorted (or more closely match the objective values) than those of infant mortality rates, since in general, people may have more knowledge about food calories than about infant mortality. Unfortunately, we do not have any measures of the amount of knowledge people have in either domain. Future work could investigate how previous exposure to objective criterion values affects the difference between two types of models.

Even if scaling errors are involved in our tasks, the very high accuracy rates obtained by the belief error models are especially striking given the fact that these models did not have access to the true criterion value being judged. Note that although our models did have access to the participant's estimates in the training data, unlike the scaling error model, the belief error model did not know the true criterion value – that is the true calorie of the food and the true infant mortality of the country. Additionally, due to the LOOCV procedure, the models did not have access to the participant's estimates for the test data. This is in contrast to the scaling error models that were explicitly given the true calorie values of the food items and the true infant mortality rates of the countries. We believe that these results are intriguing as they suggest that the belief error model can be used to model participant judgments even when the researcher does not have access to the ground truth values for the judgment dimension. So, for example, we

could use the belief error model to predict people's estimates for the probability of an earthquake in different countries, even if we don't know the objective probability of earthquakes.

The reason why we were able to achieve such high accuracy rates despite this limitation is that we equipped our belief error models with rich knowledge representations in the form of high-dimensional vectors for targets obtained from semantic space models (Bhatia, 2017; Bhatia et al., 2019; Jones et al., 2015; Landauer & Dumais, 1997; Mandera et al., 2017; Mikolov et al., 2013). Semantic space models use the structure of word distribution in natural language to derive vectors for words that represent word meaning, so that words that are closely related or highly associated with each other are given similar vectors. These vectors have been shown to describe the structure of knowledge representation in humans, and predict similarity judgments and high-level (associative) judgments obtained through standard cognitive tasks (Bhatia et al., 2019; Jones et al., 2015; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mandera et al., 2017; Mikolov et al., 2013; Pennington et al., 2014). By using word vectors derived from semantic space models to describe the set of knowledge cues possessed by participants, we were able to add to our belief models a degree of naturalism, richness, and generality, thereby greatly improving their ability to describe nuances in belief formation and numerical estimation. This is in contrast to prior work on belief errors in numerical estimation, which often specifies participant knowledge using a small set of predefined experimenter generated cues (Brown, 2002; Juslin et al., 2003; Litchtenstein et al., 1978; MacGregor, Lichtenstein, & Slovic, 1988; von Helversen & Rieskamp, 2008). Although this work has provided a number of important insights regarding the psychology of numerical estimation, it has been unable to specify accurate predictive models capable of predicting participant estimates for arbitrary natural judgment prompts. It is only by representing participant knowledge with rich semantic representations can

we build powerful quantitative models that are able to give accurate out-of-sample predictions of everyday numerical estimates.

Mimicking Scaling Errors

One important property of our belief error models is their ability to mimic the types of judgment errors generated by scaling error models. Specifically, as with our participants, and the scaling error models fit to participant data, the best-fitting belief error models overestimated small calories and underestimated large calories, and overestimated infant mortality rates (with greater overestimation for small mortality rates than large mortality rates). This suggests that a sufficiently sophisticated belief error model may be all that is needed to account for the qualitative patterns typically attributed to scaling errors.

With that in mind, it is important to note that prior work has found strong evidence for both scaling and belief errors in numerical estimation (Brown & Siegler, 1993; Hertwig et al., 2005; Litchtenstein et al., 1978; von Helversen & Rieskamp, 2008), and some work has even argued that apparent belief errors are actually caused by scaling errors (Landy et al., 2018). Although belief error models equipped with semantic vector representations may be able to explain some of the data captured by scaling error models, it is quite likely that the ideal model will need to combine both sources of error in order to account for the full scope of participant data. Such a model will over or underweight semantic vector dimensions when forming beliefs, and then map the beliefs non-linearly to the criterion variable. Unfortunately, fitting this kind of high-dimensional non-linear model to participant data is not computationally trivial. We attempted to fit such model using various computational shortcuts, and found that the joint scaling and belief error model did not outperform the predictions of the belief error-only model, suggesting that scaling errors do not add any additional predictive power on top of belief errors.

However, our shortcuts may have led to suboptimal fits, and so we consider their results to be imprecise and speculative. Clearly, building better models of naturalistic numerical estimation is an important research topic, one that will be able to shed light on the respective roles of belief and scaling errors in judgment.

Conceptual Underpinnings of Judgment

One unique benefit of our belief error model is that it provides a theoretical connection between research on numerical judgment and research on semantic cognition. Crucially, the semantic vectors in our belief error model are trained on actual natural language that people use to communicate knowledge. Therefore, these vectors provide comprehensive knowledge representations for nearly all naturalistic objects and concepts. As mentioned above, semantic vectors are commonly used to specify the contents of semantic memory, and in turn predict similarity judgment, as well as list recall, memory search, and free association (Bhatia et al., 2019; Jones et al., 2015; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mandera et al., 2017; Mikolov et al., 2013; Pennington et al., 2014). By drawing on these important insights, we were able to make predictions regarding the concepts most related to and associated with criterion variables in our experiments. These concepts are those that are located closest to the best-fitting weight vector of our belief error model: If these concepts were judgment targets, they would be given the highest ratings. In Studies 3 and 4, we tested the ability of our belief error models (fit on participant data in Studies 1 and 2) to identify the conceptual underpinning of numerical judgment, by asking participants to rate various concepts in terms of their relatedness to the judgment dimensions. As predicted, concepts identified by our models as being associated with foods with high calories and countries with high infant mortality rates were indeed given

higher relatedness ratings than concepts associated with foods with low calories and countries with low infant mortality rates.

Although Studies 3 and 4 involve a semantic judgment task, the semantic space models used in our analysis also make predictions regarding the words and concepts that come to mind when participants are asked to think about foods with high or low calories and countries with high or low infant mortality rates. We tested these predictions in Studies 5 and 6 and found that our best-fitting belief error models (from Studies 1 and 2) predicted word frequencies in free association tasks involving calories and mortality rates. Importantly, these belief error models were also able to predict serial position effects, as concepts identified by our model as being highly related to the judgment dimensions were listed earlier in the free association tasks. These results provide novel insights regarding the dynamics of memory in naturalistic numerical estimation.

Finally, Studies 7 and 8 utilized verbal protocols that allowed for a more robust test for the belief error models in predicting thought processes that take place during numerical estimation. Our results showed the concepts predicted by our model to be highly associated with the judgment dimensions were more likely to be activated when participants gave high estimates, as revealed by post hoc verbal protocols. This provides further evidence that our belief error models are able to account for the conceptual underpinnings of numerical judgment.

Cognitive Modeling of Numerical Judgment

Our use of high-dimensional semantic representations for modeling knowledge shows that it is possible to build cognitive models of naturalistic numerical judgment that mimic individuals in both what they know about the judgment targets and how they transform this knowledge onto the response scale. For this reason, these models able to predict responses on

related cognitive tasks, such as those involving semantic relatedness judgments, free association, or verbal protocols. Of course, the models considered in this paper are by no means the most sophisticated cognitive models possible. For example, we have chosen to describe the response process as involving a linear combination of the dimensions of our semantic vectors. This is inspired by classical research on numerical judgment according to which judgment is the outcome of a linear cue-aggregation process that weights cues in proportion to their statistical values (Brunswik, 1952; Gigerenzer & Kurz, 2001; Karelaia & Hogarth, 2008). However, people may be using other cognitive mechanisms for mapping semantic vectors onto the criterion value. It is possible that the judgment prompt guides responses through a simple associative activation process (e.g. Bhatia, 2017; Kahneman, 2002; Sloman, 1996), so that a judgment target that activates the set of concepts related to the criterion variable is judged to have a high value on that variable. For example, a food item that is associated with and activates the concept "fat" will be judged to have high calories, whereas a food item that activates the concept "nutrient" will be judged to have low calories. Such a model cannot be distinguished from the linear model used in this paper on the basis of numerical judgment data (the linear model implicitly identifies a region in the semantic space that is highly associated with the criterion, and thus rates judgment targets closest to concepts in that region as being the highest on the criterion variable). However, the associative activation model does predict that free association or think-out-loud responses made during deliberation should correlate with judgment on the participant-level. Future work could try to elicit free associations in conjunction with the numerical judgment (as in e.g. Johnson, Haubl, & Keinan, 2007; Hardisty, Johnson, & Weber, 2010), thereby providing a more rigorous process-level analysis of the cognitive mechanisms at play in judgment. Our successes at

predicting free associations in Studies 5 and 6 and at predicting common words in recalled thought processes in Studies 7 and 8 suggest that such tests are feasible.

Another possibility is that the activation of word associates is incidental and does not influence estimation. Under this hypothesis, people may use the weight vector to make the judgment (perhaps with some kind of algebraic judgment rule, as in our belief error model), but the weight vector nonetheless activates concepts that are closely associated in the semantic space. Again, this process would be identical to the belief error model (if the algebraic rule is the vector dot product of the weight and concept vector). To test the different memory processes, it may be possible to do a priming study in which participants are primed with different concepts (e.g. fat, organic, tasty) and see if the prime affects the eventual judgment. This is another promising avenue for future work.

Other, more complex ways of transforming semantic vectors onto the judgment dimension involve exemplar models. Such models are capable of learning various complex cue-criterion relations, and can, for this reason, describe non-linear transformations from the semantic space onto the criterion. Exemplar models have previously been applied to numerical judgment (e.g. Juslin et al., 2003; Juslin, Karlsson, & Olsson, 2008; also see Dougherty, Gettys & Ogden, 1999; Juslin & Persson, 2002; Smith & Zárate, 1992). Although this work has mostly used artificial cue values, a similar exercise can easily be repeated with semantic vectors. It is also possible to apply the mapping model proposed by Von Helversen & Rieskamp (2008; also see Brown & Siegler, 1993). This model assumes that cue values (or, in our case, semantic vectors) are used by participants to specify ordinal relations between judgment targets. These ordinal relations are then projected onto the response scale using typical criterion values for

various categories of targets. Again, these are valuable directions for future research on naturalistic numerical estimation.

Although overall, our results support the dominant role of belief errors, we do not rule out the possibility of scaling errors. Even though the words generated in verbal reports are predictable by the belief error model and most of the reports suggest that participants based their estimates on associations, a few reports show that participants also relied on other non-verbal cues to generate estimates. For example, a participant reported using one death per 60 child births as a benchmark to estimate infant mortality rates; a participant whose estimates were significantly lower than others believed that the infant mortality is low in general; some participants estimated calories by imagining the portion size of 100 grams of foods or by comparing them with other food items whose calorie they remembered. These individual thought processes reveal knowledge cues that are not captured by the belief error model. In those cases, a more complex model that incorporates belief errors, scaling errors, and perceptual traits would provide a better picture of the cognitive process underlying numerical estimation. Moreover, in some rare cases, the thought process was too quick that participants themselves could not even verbalize their deliberation and later self-corrected their previous estimates in their verbal reports. This suggests that even an immediate recall of the thought process may not necessarily recover the actual estimation mechanisms (see e.g. Nisbett & Wilson, 1977 for a well-known critique). A complementary and more robust approach to examine the underlying cognitive process could involve using neuroimaging to measure real-time brain activation during numerical estimation, which is also a valuable direction for future research.

Learning Processes

Both the exemplar and mapping models discussed in the previous section require prior information about criterion values of some judgment targets in order to make predictions. This information can be given to participants in experiments with a learning task preceding numerical judgment. Learning-based experiments are powerful tools for analyzing the complex mental processes at play in numerical judgment, and we hope to apply them to the types of settings considered in this paper in our future work.

A related question involves how and when people learn the weight vectors posited by our belief error models. We believe that in settings without prior experience, participants may merely use the word vector corresponding to the judgment dimension. For example, when estimating food calories without any prior knowledge of actual calories, people may simply generate an answer based on the similarity between the word vector for the target concept (e.g. apple) and the word vector for the word "calories". In this case, the weight vector is identical to the word vector of the judgment dimension – "calories". When people have some prior knowledge of actual calorie amounts of various foods, we believe that people gradually adapt their weight vectors for food calories to better reflect the calorie distribution in real life. Moreover, we assume that weight vectors are learned as a by-product of everyday language experience where some feedback learning could occur but not necessarily. Again, learning tasks could be useful for investigating these hypotheses.

Naturalistic Judgments

Finally, we would like to emphasize the naturalism of the two domains examined in this paper. Our approach is unique in that it can be used to study numerical estimates for complex natural entities, such as food items and countries. This opens up new avenues for applying psychological theories of semantic cognition and judgment and decision making, to policy-

relevant applications, such as those pertaining to health-related and humanitarian issues. For example, our belief error models can be used to pick out food items that tend to have large estimation errors and accordingly develop behavioral interventions to help mitigate these errors. Similar models can be used to diagnose biases in judgments of infant mortality rates. Ultimately, with the theoretically grounded insights obtained through our analysis, it should be possible to design behavioral interventions to more closely align participant calorie estimates and infant mortality rates with true values, and more generally, to nudge healthy eating and charitable giving (Butera & Houser, 2018; Glanz & Mullis, 1988; Goswami & Urminsky, 2016; Michie et al., 2009; Oppenheimer & Olivola, 2011)

Models capable of describing judgments for natural entities are also desirable due to their very large domain of applicability. Unlike judgment models trained on artificial experimental stimuli, our models can predict, with high accuracy, the calories that people will assign to the thousands of food items that have natural language labels. Similar predictions can be made for infant mortality rates for hundreds of countries and geographic regions. More generally, the semantic spaces used to equip our models with knowledge representations have vast vocabularies, and thus can be applied to any domain with natural language judgment targets. These judgment targets include food items and countries, but also consumer goods and brands, job occupations, public figures and celebrities, and the many other concept categories at play financial, consumer, social, political, and economic judgment (see Richie et al., 2019 for a discussion). We look forward to future work that extends our approach to model the types of psychological mechanisms at play in the many important judgments that people make on a day-to-day basis.

References

Barnhardt, T. M., Choi, H., Gerkens, D. R., & Smith, S. M. (2006). Output position and word relatedness effects in a DRM paradigm: Support for a dual-retrieval process theory of free recall and false memories. *Journal of Memory and Language*, *55*(2), 213–231.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014 (52nd Annual Meeting of the Association for Computational Linguistics)* (pp. 238–247). East Stroudsburg, PA: ACL.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Bhatia, S. (2019a). Predicting risk perception: New insights from data science. *Management Science, 65*(8), 3800–3823.

Bhatia, S. (2019b). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 627–640.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36.

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition*, *179*, 71–88.

Bhatia, S., & Walasek, L. (2019). Association and response accuracy in the wild. *Memory & Cognition*, *47*(2), 292–298.

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. *Psychology of Learning and Motivation*, *41*, 321–359.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511–534.

Brunswik, E. (1952). The conceptual framework of psychology. *Psychological Bulletin*, *49*(6), 654–656.

Butera, L., & Houser, D. (2018). Delegating altruism: Toward an understanding of agency in charitable giving. *Journal of Economic Behavior & Organization*, *155*, 99–109.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Chandon, P., & Wansink, B. (2007). The biasing health halos of fast-food restaurant health claims: Lower calorie estimates and higher side-dish consumption intentions. *Journal of Consumer Research*, *34*(3), 301–314.

Chernev, A., & Chandon, P. (2011). Calorie estimation biases in consumer choice. In R. Batra, P. Keller, & V. Strecher (Eds.), *Leveraging consumer psychology for effective health communications: The obesity challenge* (pp. 104–121). New York, NY: M.E. Sharpe.

Curtis, D. W., Attmeave, F., & Harrington, T. L. (1968). A test of a two-stage model of magnitude judgment. *Perception & Psychophysics*, *3*(1), 25–31.

Dougherty, M. R. P., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review*, *115*(1), 199–213.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180–209.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251.

Erlick, D. E. (1964). Absolute judgments of discrete quantities randomly distributed over time. *Journal of Experimental Psychology*, *67*(5), 475–482.

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology*, *12*, 263–282.

Fennell, J., & Baddeley, R. (2012). Uncertainty plus prior equals rational bias: An intuitive bayesian probability weighting function. *Psychological Review*, *119*(4), 878–887.

Gallagher, C. A. (2003). Miscounting race: Explaining whites' misperceptions of racial group size. *Sociological Perspectives*, *46*(3), 381–396.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gigerenzer, G., & Kurz, E. M. (2001). Vicarious functioning reconsidered: A fast and frugal lens model. *The essential Brunswik: Beginnings, explications, applications*, 342–347.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67.

Glanz, K., & Mullis, R. M. (1988). Environmental interventions to promote healthy eating: A review of models, programs, and evidence. *Health Education Quarterly*, *15*(4), 395–415.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166.

Goswami, I., & Urminsky, O. (2016). When should the ask be a nudge? The effect of default amounts on charitable donations. *Journal of Marketing Research*, *53*(5), 829-846.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*, 1006–1033.

Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). A dirty word or a dirty world? Attribute framing, political affiliation, and query theory. *Psychological Science*, *21*(1), 86–92.

Herda, D. (2013). Too many immigrants? Examining alternative forms of immigrant population innumeracy. *Sociological Perspectives*, *56*(2), 213–240.

Herda, D. (2015). Beyond innumeracy: Heuristic decision-making and qualitative misperceptions about immigrants in Finland. *Ethnic and Racial Studies*, *38*(9), 1627–1645.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 621–642.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440.

Hollands, J., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, *107*(3), 500–524.

Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory & Cognition, 45*, 1350–1370.

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619.

Johnson, E.J., Haubl, G., Keinan, A. (2007). Aspects of endowment: A query theory of value. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 461–474.

Johns, B. T., Jones, M. N., & Mewhort, D. J. (2016). Experience as a free parameter in the cognitive modeling of language. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2291–2296). Austin, TX: Cognitive Science Society.

Johns, B. T., Jones, M. N., & Mewhort, D. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review, 26*, 103-126.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics, 20*(1), 1-31.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of memory and language*, *55*(4), 534-552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1–37.

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). New York, NY: Oxford University Press.

Jou, J. (2008). Recall latencies, confidence, and output positions of true and false memories: Implications for recall and metamemory theories. *Journal of Memory and Language*, *58*(4), 1049–1064.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive science*, *26*(5), 563–607.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel Prize Lecture*, *8*, 351–401.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological bulletin*, *134*(3), 404–426.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211–240.

Landy, D., Guay, B., & Marghetis, T. (2018). Bias and ignorance in demographic perception. *Psychonomic Bulletin & Review*, *25*(5), 1606–1618.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551–578.

MacGregor, D., Lichtenstein, S., & Slovic, P. (1988). Structuring knowledge retrieval: An analysis of decomposed quantitative judgments. *Organizational Behavior and Human Decision Processes*, *42*(3), 303–323.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and

counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Manning, J. R., & Kahana, M. J. (2012). Interpreting semantic clustering effects in free recall. *Memory*, *20*(5), 511–517.

Michie, S., Abraham, C., Whittington, C., McAteer, J., & Gupta, S. (2009). Effective techniques in healthy eating and physical activity interventions: A meta-regression. *Health Psychology*, *28*(6), 690–701.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Oppenheimer, D. M., & Olivola, C. Y. (2011). *The science of giving: Experimental approaches to the study of charity*. Psychology Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . ., Vanderplas, J. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12* (Oct), 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology, 5*(1): 50. doi: https://doi.org/10.1525/collabra.282.

Shepard, R. N. (1981). Psychological relations and psychophysical scales: On the status of "direct" psychophysical measurement. *Journal of Mathematical Psychology*, *24*(1), 21–57.

Simon, H. A. (1969). The sciences of the artificial. Cambridge, MA: MIT Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*(1), 3–21.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1975). *Psychophysics*. New York: Wiley.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*(6), 377–411.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, 237–249.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

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

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 613–625.

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, *137*(1), 73–96.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1–14.