Wonderful is "More" than Terrible: Valence Influences Judgments of Magnitude

Qiawen Ella Liu (ella.qiawen.liu@princeton.edu)[1] **& Adele E. Goldberg**[2]
[1]Department of Computer Science, Princeton University
[2]Department of Psychology, Princeton University

Abstract

Everyday decisions about quantity, size, and duration require people to make judgments about magnitude. Perceived energy/intensity influences such judgments, and here we report evaluative meaning is independently predictive. For example, "wonderful" is reliably judged to be more than "terrible." Results on antonym pairs show robust agreement across participants on two tasks: an explicit rating task (*which is more?*) and a task where pairs of adjectives are positioned on a line extending from an origin toward infinity. In both, energy ratings strongly predicted responses and valence predicted additional variance. We further find both energy and valence are reflected in distributional semantic models: FastText embeddings projected onto a direction for magnitude correlate strongly with human judgments, with comparable effects of valence and energy on the model's predictions. Thus the evaluative component of magnitude is recoverable from linguistic statistics. We discuss potential mechanisms underlying this association, including metaphorical mediation, linguistic markedness asymmetries, and cognitive biases toward simulating improvement.

Keywords: magnitude; valence; language

Introduction

Language, like life, contains multitudes of opposites: *strong-weak, mild-aggressive, dull-sharp, bored-interested, stupid-smart, beautiful-ugly, intense-mild, happy-sad, active-passive, shiny-matte*. None of these particular pairs literally differs on the dimension of quantity. Yet when we asked English speakers which antonym in each pair is "more," we discovered broad agreement: shiny is "more" than matte, sharp is "more" than dull, happy is "more" than sad. This raises a foundational question about the nature of magnitude, a key dimension of human cognition.

From early in development, humans reason about *more* and *less* across domains of quantity, size, and duration, well before they learn exact number concepts. For instance, children can reliably judge who has more items without knowing how many items are present (Halberda & Feigenson, 2008). The word *more* is in fact one of the earliest abstract relational terms that children learn (Barner et al., 2009; Brown, 1973; Carey, 2009). Approximate magnitude comparison is observed across all cultures, including in societies with severely limited number vocabularies, and it is shared with non-human animals, indicating that moreness reflects a culturally and biologically widespread concept (Dehaene, 1997; Feigenson et al., 2004; Gordon, 2004).

A general theory of magnitude posits that humans rely on a partially shared representational system for comparing "more" and "less" across otherwise distinct domains (Bueti & Walsh, 2009; Dehaene et al., 1993; Henik & Tzelgov, 1982; Srinivasan & Carey, 2010; Walsh, 2003), present soon after birth (De Hevia et al., 2014; Lourenco & Longo, 2010). This magnitude system has several well-documented behavioral signatures: people decide more quickly that $9 > 2$ than that $3 > 2$ (a distance effect) and accuracy degrades as the ratio of the two magnitudes approaches one (a ratio effect). People respond faster to questions including a default term (which is bigger?) but this is tempered when the comparison involves items from the low end of the scale (e.g., *tiny and small*) (a congruity effect) (Holyoak & Walker, 1976).

Critically, these properties extend beyond perceptual quantities and numerals into language. The same distance, ratio, and congruity effects appear when speakers judge linguistic scales, from quantifiers such as *few–some–many–all* (Pezzelle et al., 2018), to expressions of probability (e.g., *unlikely-possible–likely–certain*) (Jaffe-Katz et al., 1989), to social status (Chiao et al., 2004) and gradient adjectives (e.g., *calm–annoyed–angry–furious*) (Holyoak & Walker, 1976; Varma et al., 2024). Each of these scales is not merely ordered but oriented: one pole is treated as the high end, and its opposite, the low end.

What gives a scale its orientation? A natural candidate is energy, psychologically experienced as increased arousal, and broadly construed as the capacity to exert force, produce effects, or bring about change. [1] In everyday experience, higher positions afford greater potential, larger and heavier objects tend to exert greater force, stronger agents generate greater impact, and as a result, increases along these dimensions systematically co-vary. These correlations between magnitude and energetic qualities are reflected in several conceptual metaphors: e.g., MORE as UP (prices can be sky-high or come crashing down), MORE as BIG (numbers can be giant or tiny), MORE as STRONG (signals can be strong or weak) (Lakoff, 1987; Lakoff & Johnson, 1980). Converging behavioral and neural evidence suggests that these correspondences are not merely linguistic conventions but reflect a partially shared magnitude code: comparisons along one magnitude dimension reliably interfere with or prime, depending on the task,

[1]We use the term *energy* rather than *arousal* for consistency with our behavioral task and because *energy* avoids the narrower sexual connotations of *arousal* in English.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

judgments along others: Larger numbers align with greater spatial extent, brighter stimuli, louder sounds, and more forceful actions (Bueti & Walsh, 2009; Hartmann & Mast, 2017; Lindemann et al., 2007; Pinel et al., 2004; Vierck & Kiesel, 2010; Walsh, 2003).

Intriguingly, some of the same magnitude-based mappings (e.g., More as Up, Big, Strong, Intense) do not merely represent changes in quantity or energy, but are also routinely recruited to express value. Across languages and cultures, upward spatial position is associated with goodness as well as more quantity, as when positive qualities are described using words and phrases expressing verticality (e.g., *take the high road, lofty ideals, top quality*). Verticality is associated with positive valence in behavioral tasks as well. For example, positive stimuli are preferentially associated with upper spatial locations in speeded classification tasks (Cian, 2017; Meier & Robinson, 2004). Verticality is not the only dimension associated with positive valence: larger stimuli are evaluated more positively than smaller ones even when size is irrelevant to the task (Meier et al., 2008); heavier objects are judged more important and consequential (Jostmann et al., 2009), and greater strength is often associated with power, competence, and positive evaluation (Casasanto, 2009).

These findings suggest a relationship between magnitude and valence that raises certain questions. Is the valence component observed in magnitude judgments merely a byproduct of an alignment with energy, since higher energy states are frequently more positive? Or does valence contribute independently to decisions about magnitude? That is, do valence effects remain once the energy factor is controlled for?

Another largely untested question is whether valence is systematically encoded in language in ways that parallel the way magnitude is encoded. In addition to the pervasive use of shared dimensions in conceptual metaphors to express both quantity and value (e.g., *up*, *big*, *strong*), language exhibits more general asymmetries that may bias how valence and magnitude align. Linguistic theories of markedness suggest that positive adjectives often function as unmarked defaults, while negative adjectives are more likely to be pragmatically or morphologically marked. For example, asking *how good* a restaurant is does not presuppose it is good, whereas asking *how bad* it is does (Clark, 1969). Perhaps relatedly, a *Pollyanna Hypothesis* has been proposed to explain why positive words are more frequent than negative words (Boucher & Osgood, 1969; Dodds et al., 2015). Together, these asymmetries suggest that linguistic experience may systematically align positive evaluation with the dominant pole of magnitude dimensions.

In this paper, we provide a systematic test of two hypotheses about magnitude representations. First, we examine whether valence makes an additional, independent contribution to judgments of "what is more", beyond an expected effect of energy. Second, we test whether any such evaluative contribution to magnitude judgments is systematically reflected in language. We report strong agreement across participants about which of two adjectives is "more," whether the question is posed verbally or by asking participants to position the terms on a number line. As expected, we find that perceived energy robustly predicts people's magnitude judgments. Importantly, we also find that perceived valence explains unique variance as an additional significant predictor of moreness.

We further show that the implicit valence of adjective pairs is recoverable from linguistic statistics. Using a distributional semantic model, projections onto a constructed moreness direction strongly correlate with human judgments and with valence predicted model-based moreness estimates, to a degree comparable to energy. Together, results indicate an intrinsic evaluative component to people's magnitude judgment, a component that appears to be reflected in and learnable from language.

Methods

To assess the robustness of moreness judgments, we used two independent behavioral measures of magnitude: an explicit comparative rating task and a spatial line placement task. These measures were administered to separate groups of participants. The set of adjective antonym pairs was the same.

Rating of moreness, energy, and valence

Participants. We recruited $N = 150$ online English-speaking American adults from Prolific (63 male, 83 female, 2 non-binary, and 2 preferred not to say). Participants' mean age was 42.4 years ($SD = 12.17$, range = 21-70).

Antonym Selection. We prompted Claude-Sonnet-4.5 to generate 100 adjective antonym pairs chosen to vary in the degree to which valence and energy aligned or diverged. The set included pairs emphasizing valence differences with minimal energy contrast (e.g., *honest/deceptive*), pairs emphasizing energy differences with relatively neutral valence (e.g., *dynamic/static*), as well as pairs in which valence and energy pointed in the same direction (e.g., *bold/timid*) or in opposite directions (e.g., *peaceful/violent*). The model was instructed to avoid morphologically related pairs (e.g., *fair/unfair*) and quantity-based pairs (e.g., *tall/short*). The first author manually reviewed each generated pair and iteratively re-prompted the model to replace pairs that were morphologically related, quantity-based, near-synonymous, or that otherwise failed to meet the design criteria.

Procedure. Each participant was randomly assigned 20 antonym pairs, presented in random orders. For each pair, participants were asked to respond to, "Which is more, *A* or *B*?" by responding on a 7-point scale ranging from "*A* is definitely more" to "*B* is definitely more" with "Unsure" as the midpoint. The order in which the two antonyms appeared was counterbalanced across participants.

Following the moreness judgments, participants rated each adjective in random order on two dimensions: valence

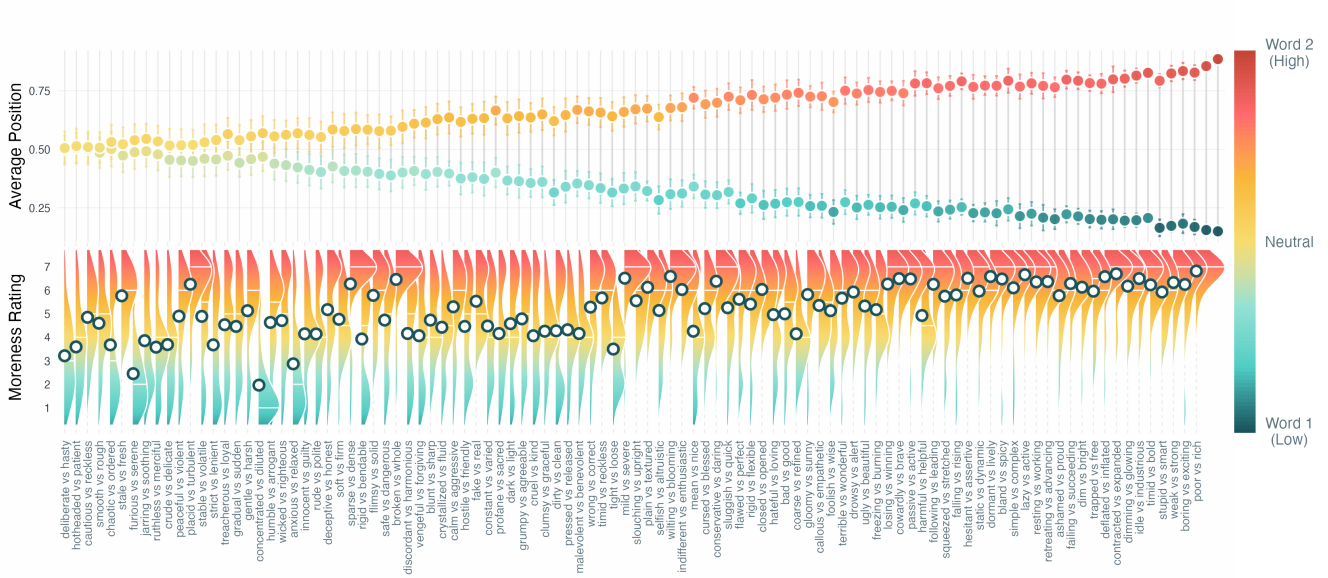

Figure 1: Distribution of moreness ratings and spatial placements across 100 adjective antonym pairs. Top panel shows average spatial position on the horizontal line (0 = origin, 1 = toward infinity), with circles indicating mean values and error bars showing standard errors. Bottom panel shows moreness ratings on a 7-point scale (1 = Word 1 is definitely more, 7 = Word 2 is definitely more, 4 = unsure). Antonym pairs are ordered along the x-axis from left to right by their average spatial position difference. Histograms show the distribution of individual participants' moreness rating. Color gradient represents the dominant response direction, transitioning from teal (Word 1 judged as "more") through neutral yellow to red (Word 2 judged as "more").

("Which one feels more positive?") and energy ("Which one conveys more energy?"), on 7-point scales ranging from "A clearly feels more positive/clearly conveys more energy" to "B clearly feels more positive/conveys more energy". The order of valence and energy rating blocks was counterbalanced between participants, and the order of antonyms within each block was randomized.

Six catch trials were included to assess basic attention and comprehension of the judgment scales, using unambiguous antonym pairs with an obvious correct direction (e.g., *small/large* for moreness, *bad/good* for valence, *strong/weak* for energy). Seven participants who failed more than two catch trials were excluded from analyses.

Line placement task

Participants. We recruited a new group of $N = 150$ online English-speaking American adults from Prolific (61 male, 87 female, 2 non-binary). Participants' mean age was 43.25 years ($SD = 11.86$, range = 22-78).

Stimuli. We used the same 100 antonym pairs generated for the moreness rating task.

Procedure. Each participant was randomly assigned 20 antonym pairs, presented in random order. For each pair, both adjectives appeared on screen above a horizontal line marked with an origin point on the left and an infinity symbol (∞) on the right. Participants were instructed to drag each word to any position on the line that felt most natural, based on their intuitive sense of how the concepts should be arranged spatially. Participants were not instructed to indicate which adjective represented "more" at any point during the task; magnitude judgments were inferred from participants' spatial placements. The left–right order of the two antonyms on the screen was counterbalanced across participants.

Four catch trials using literal magnitude comparisons were embedded in the task (*small/large*, *short/long*, *less/more*, *narrow/wide*). To pass a catch trial, participants had to place the conventionally "more" word farther from the origin than the "less" word. Four participants who failed more than one catch trial were excluded from analyses.

Distributional semantic analysis

To test whether the structure of moreness judgments is reflected in linguistic distributional statistics, we conducted a computational analysis using pretrained word embeddings following the semantic projection approach (Grand et al., 2022). We used FastText embeddings trained on Wikipedia and Common Crawl (fasttext-wiki-news-subwords-300; Bojanowski et al. 2016), which provide high-quality distributional represen-

tations derived from large-scale natural language use. [2]

Deriving the magnitude projection. Following Grand et al. (2022), we defined a magnitude direction by averaging across four antonym pairs chosen as among the most prototypical lexical instantiations of the more/less contrast in English: *more/less*, *increase/decrease*, *many/few*, and *greater/lesser*. These pairs were selected before the analysis, not post-hoc.

For each seed pair (w_i^+, w_i^-), we computed a difference vector $\vec{d}_i = \vec{w}_i^+ - \vec{w}_i^-$ and normalized it to unit length. The magnitude axis $\vec{M}$ was the mean of the four normalized difference vectors:

$$\vec{M} = \frac{1}{4} \sum_{i=1}^{4} \frac{\vec{d}_i}{\|\vec{d}_i\|} \tag{1}$$

For each of the 100 test antonym pairs (a, b) in our stimulus set, we computed the magnitude projection as the cosine similarity between the pair's difference vector and the magnitude axis:

$$\text{proj}(a, b) = \frac{(\vec{a} - \vec{b}) \cdot \vec{M}}{\|\vec{a} - \vec{b}\| \cdot \|\vec{M}\|} \tag{2}$$

A positive value indicates that b aligns more closely with *more* than a does, a negative value indicates the reverse, and the absolute value reflects the strength of the association.

Results

Moreness ratings track spatial placement on line

As shown in Figure 1, magnitude judgments derived from explicit moreness ratings closely aligned with spatial placement on the line. That is, across adjective pairs, the word that tended to be judged as "more" in the rating task was also typically placed farther from the origin (toward infinity) in the spatial task. For example, if participants tended to rate *rich* as more, other participants likewise tended to place *rich* farther to the right than *poor*.

To quantify the correspondence between the two measures, we computed the correlation between average moreness ratings and average spatial position differences across adjective pairs. Because antonym pairs have no inherent ordering—and word order was randomized during data collection (e.g., participants might see *soft vs. firm* or *firm vs. soft*), we imposed a fixed alphabetical ordering within each pair for analysis (e.g., *firm vs. soft*). This sorting method is content-neutral, serving only to define the direction of signed differences. Averaged across participants, moreness ratings and spatial position differences were strongly correlated ($r = .88$), indicating substantial convergence between the two measures.

Nonetheless, some pairs exhibited partial misalignment across tasks. For instance, *concentrated* was more often judged as "more" in the rating task, while spatial placements for *concentrated/diluted* were less differentiated. Conversely, some pairs showed clearer spatial separation but bimodal moreness ratings: e.g., for *dirty/clean*, participants tended to place *clean* further to the right, yet for moreness ratings, half of the people judged *clean* as more, half judged *dirty* as more. The last point highlights the fact that on a minority of antonym pairs, participant responses tended to be bimodal on one or both measures. For example, for *treacherous/loyal* and *gradual/sudden* participants' responses on both tasks tended toward bimodality, resulting in intermediate mean ratings and average central spatial placements for both words in the pairs.

Valence predicts moreness judgments controlling for perceived energy

We next examined whether moreness judgments could be explained solely by perceived energy or whether valence contributed independently. For each adjective pair, we computed its mean moreness rating, mean spatial position difference, mean perceived energy, and mean perceived valence. Valence and energy were only weakly correlated ($r = .24$), indicating limited multicollinearity in our dataset. We then regress both magnitude measures on energy and valence in multiple linear regression models.

As shown in Figure 2, perceived energy robustly predicted both moreness ratings ($\beta = .83$, $t = 26.59$, $p < .001$) and spatial position differences ($\beta = .61$, $t = 13.52$, $p < .001$). Valence remained a significant predictor of both measures after controlling for energy. For moreness ratings, valence showed a smaller but reliable effect ($\beta = .32$, $t = 10.14$, $p < .001$). For spatial placement, the contribution of valence was comparable in magnitude to that of energy ($\beta = .54$, $t = 12.06$, $p < .001$). In both cases, adding valence to the energy-only model significantly improved model fit (moreness rating: $F(1, 97) = 102.92$, $p < .001$; placement: $F(1, 97) = 145.53$, $p < .001$).

If valence and energy each contribute independently to moreness, pairs in which the two dimensions point to opposite poles should yield less consistent judgments across participants than pairs in which they agree. To test this, we regressed the within-pair standard deviation of moreness ratings on a continuous divergence score (the negative product of standardized valence and energy differences, with higher values indicating stronger disagreement between the two dimensions). Divergence strongly predicted within-pair variability ($\beta = 0.29$, $SE = 0.04$, $t = 7.88$, $p < .001$, $R^2 = 0.38$). That is, when valence and energy pointed to opposite poles, participants' moreness ratings spread out substantially more than when the two dimensions agreed.

Word embeddings reliably reproduce human magnitude judgments and are equally predicted by energy and valence

Finally, we asked whether magnitude structure extracted from language alone mirrors these human judgments. We projected

[2] We intentionally used static distributional embeddings rather than large language models, whose representations are shaped not only by linguistic co-occurrence but also by instruction tuning and value-alignment objectives. Because our goal was to isolate structure present in language statistics themselves, distributional embeddings provide a more faithful test bed.

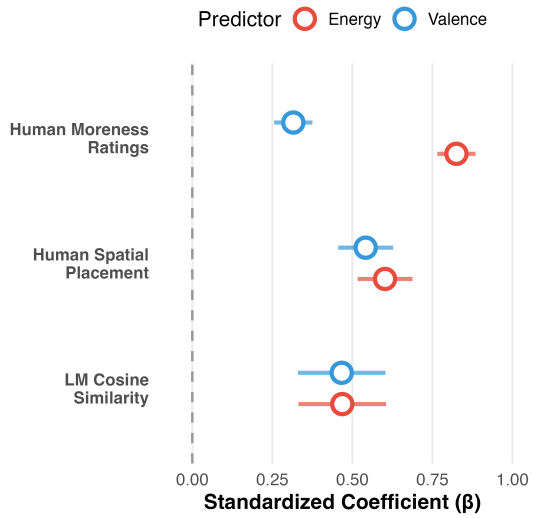

Figure 2: Standardized regression coefficients (β) from multiple linear regressions predicting three magnitude-related measures (Human moreness ratings, human spatial placement, and language-model cosine similarity) from perceived energy and valence. Points indicate estimated coefficients and horizontal bars denote 95% confidence intervals.

each adjective pair into a one-dimensional magnitude axis derived from FastText embeddings and compared these projections to the human measures. Language model projections were strongly correlated with both human moreness ratings ($r = .71$) and spatial position differences ($r = .73$), indicating that distributional semantic representations capture much of the structure underlying human magnitude judgments.

To examine what predicts these language-based projections, we regressed the embedding-derived magnitude values on normative ratings of perceived energy and valence. Both dimensions significantly predicted the language model projections, with standardized coefficients of comparable magnitude (Energy: $\beta = .46, t = 6.50, p < .001$); valence ($\beta = .46, t = 6.45, p < .001$). [3]

Discussion

Magnitude is a foundational dimension of human cognition, yet much of what we know about it concerns numerical quantity or energy/intensity. In this work, we asked whether evaluations of valence (positive vs. negative) also contribute to our conception of magnitude. We reported strong convergence between two measures aimed to determine which adjective in an antonym pair was judged to have greater magnitude: A direct judgment task asked participants which antonym was "more" and a second task asked other participants to position both terms on a number line without any inclusion of the word, *more*. As expected, perceived energy robustly predicted each measure. Notably, valence accounted for significant additional variance.

The same pattern was evident in distributional semantic models: projections onto a constructed moreness direction correlated strongly with human judgments, and valence ratings predicted model-based moreness estimates to a degree comparable to energy ratings. Together, these results demonstrate that if one quality is judged to be "more" than another quality, it also tends to be judged *better* than the other quality. Moreover, analyses revealed that the contribution of valence to magnitude is reflected in and therefore learnable from linguistic statistics.

Why do different measures weight valence and energy differently?

Although energy and valence both predicted moreness across all three measures of magnitude, their relative contributions differed by task (Figure 2). Energy was the strongest predictor of explicit judgments ("which is more?"), while valence and energy played roughly equivalent roles in the spatial placement task and in language-model representations. One explanation is that these differences arise from how the tasks themselves frame what it means to judge "more." While the explicit linguistic cue "more" appears to bias participants toward interpreting magnitude in terms of energy, mapping words onto space may recruit additional sensorimotor heuristics, e.g., a "rightward-is-better" bias that tracks people's dominant hand (Casasanto, 2009). Yet we found that the language model, despite no embodied experience, weighted valence almost as heavily as energy, suggesting that language substantially interweaves the concepts of magnitude and valence. We discuss this below.

Why is positive valence *more*?

A central question raised by these findings is *why* positive evaluation should systematically align with moreness. Below we discuss three non-mutually exclusive mechanisms that could give rise to this alignment.

First, as alluded to in the introduction, many domains that structure metaphors of magnitude (*up*, *big*, *strong*) are themselves positively valenced, with behavioral evidence confirming that these dimensions systematically align with valence (Crawford, 2009). Thus the "more as good" association might be mediated by these other magnitude-related dimensions. On this account, valence may not directly attach to moreness, but instead take effect through mediating dimensions that are jointly recruited for magnitude perception.

Another possibility is that the observed association between moreness and positivity reflects asymmetrical linguistic use of antonyms. Linguistic theories of markedness propose that negative adjectives are more likely to be morphologically

[3] A potential alternative explanation is that participants simply judged the more frequent (and thus more cognitively accessible) antonym as "more." To address this, we re-fit all three models (moreness ratings, spatial placement, and embedding-derived projections) with the Zipf-frequency difference between the two words as an additional predictor (from the WordFreq Python package; Speer (2022)). Frequency was non-significant in every model (all $|t| < 0.36$, all $p > .72$). The energy and valence coefficients were essentially unchanged.

marked than positive adjectives (e.g., *unhappy* vs. *happy*) (Ingram et al., 2016). Why the unmarked pole should tend to be positive itself requires an explanation, but on this account, the alignment of "more" with the positive pole may arise because linguistic conventions systematically privilege the unmarked term, as the default reference point for comparison. Our data cannot determine whether such markedness asymmetries play a causal role in shaping magnitude judgments or merely correlate with them. A direct test could experimentally manipulate markedness, e.g., introducing novel adjectives whose markedness is experimentally assigned. If markedness causally biases moreness, then making one pole unmarked in language should shift it toward being more likely to be judged "more."

The finding that judgments of moreness as well as the unmarked pole tend to align with positivity suggests the possiblity that both stem from a cognitive bias toward simulating improvement. Research on counterfactual thinking and mental simulation suggests that people disproportionately generate more positive alternatives—representations of how things could be better rather than worse—even when alternatives are unconstrained or arbitrary (Byrne, 2002). For instance, people spontaneously imagine counterfactuals such as "If I had studied, I would have gotten an A," more than worse outcomes (Roese et al., 2005). Likewise, when English, Polish, or Mandarin speakers were asked to list ways in which an ordinary situation could be different (e.g., how an apartment, a job, or one's daily routine might change), people reliably generated more improvements than degradations—such as a nicer kitchen, a higher salary, or a more flexible schedule—even when explicitly encouraged to consider both directions (Mastroianni & Ludwin-Peery, 2022). Under this account, "more" may be naturally aligned with positive evaluation because it corresponds to a default direction of imagined improvement.

Limitations and open questions

Several limitations of the present work point to important directions for future research. First, our measures capture explicit judgments and placements, rather than rapid or implicit responses. It remains unclear whether the association between 'more' and 'good' is spontaneous and automatic, or whether it emerges only when we force people to make a judgment. Measures such as implicit association tasks could help determine whether valence is activated spontaneously during magnitude processing.

Second, our participant samples were drawn from a single cultural and linguistic context. Although the linguistic markedness and positivity biases we appeal to have been documented across languages, the strength and direction of the valence–moreness association may vary across cultures. Some cultures may place greater value on moderation, balance, or restraint, potentially weakening or reversing the tendency for "more" being evaluated as better. Cross-linguistic and cross-cultural studies will be critical for determining whether the evaluative component of moreness is universal or culturally contingent.

A third limitation is that we did not classify antonym pairs by their scalar properties. Antonym pairs differ along several dimensions that may shape people's magnitude judgments: e.g., whether they project open scales unbounded at both ends (*good/bad*), scales bounded at one end (*loud/quiet*), or scales bounded at both ends *(full/empty*); whether their standard of comparison is relative and contextually fixed (e.g., *tall*) or absolute and tied to a scale endpoint (e.g., *dead*) (Kennedy & McNally, 2005); or whether the two poles sit on the same side of a natural zero or straddle a neutral midpoint (e.g., *hot/cold* or *dirty/clean*) (Varma et al., 2024). These properties may help deepen the cognition of magnitude, but we leave them for future work such as whether energy and valence are carried by different kinds of scales, e.g., energy may be captured more by one-sided scales with a floor but no ceiling (because things can always be more intense), while valence may involve scales that straddle a neutral midpoint (e.g., *kind/cruel, beautiful/ugly*).

Acknowledgments

We are grateful to Princeton University and the AI Lab for funding and to Arielle Belluck for early discussions related to quantity and metaphor.

References

Barner, D., Chow, K., & Yang, S.-J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, *58*(2), 195–219.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Boucher, J., & Osgood, C. E. (1969). The pollyanna hypothesis. *Journal of verbal learning and verbal behavior*, *8*(1), 1–8.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B*, *364*(1525), 1831–1840.

Byrne, R. M. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in cognitive sciences*, *6*(10), 426–431.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right-and left-handers. *Journal of Experimental Psychology: General*, *138*(3), 351.

Chiao, J. Y., Bordeaux, A. R., & Ambady, N. (2004). Mental representations of social status. *Cognition*, *93*(2), B49–B57. https://doi.org/10.1016/j.cognition.2003.07.008

Cian, L. (2017). Verticality and conceptual metaphors: A systematic review. *Journal of the Association for Consumer Research*, *2*(4), 444–459.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological review*, *76*(4), 387.

Crawford, L. E. (2009). Conceptual metaphors of affect. *Emotion review*, *1*(2), 129–139.

De Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., & Streri, A. (2014). Representations of space, time, and number in neonates. *Proceedings of the National Academy of Sciences*, *111*(13), 4809–4813.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of experimental psychology: General*, *122*(3), 371.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., et al. (2015). Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, *112*(8), 2389–2394.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, *306*(5695), 496–499. https://doi.org/10.1126/science.1094492

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, *6*(7), 975–987.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the" number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology*, *44*(5), 1457.

Hartmann, M., & Mast, F. W. (2017). Loudness counts: Interactions between loudness, number magnitude, and space. *Quarterly Journal of Experimental Psychology*, *70*(7), 1305–1322.

Henik, A., & Tzelgov, J. (1982). Is three greater than five? the relation between physical and semantic size in comparison tasks. *Memory & Cognition*, *10*(4), 389–395.

Holyoak, K. J., & Walker, J. H. (1976). Subjective magnitude information in semantic orderings. *Journal of Verbal Learning and Verbal Behavior*, *15*(3), 287–299. https://doi.org/10.1016/S0022-5371(76)90026-8

Ingram, J., Hand, C. J., & Maciejewski, G. (2016). Exploring the measurement of markedness and its relationship with other linguistic variables. *PLoS One*, *11*(6), e0157141.

Jaffe-Katz, A., Budescu, D. V., & Wallsten, T. S. (1989). Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty. *Memory & Cognition*, *17*(3), 249–264. https://doi.org/10.3758/BF03198463

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological science*, *20*(9), 1169–1174.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*(2), 345–381. https://doi.org/10.1353/lan.2005.0071

Lakoff, G. (1987). *Women, fire, and dangerous things*. University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lindemann, O., Abolafia, J. M., Girardi, G., & Bekkering, H. (2007). Getting a grip on numbers: Numerical magnitude priming in object grasping. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1400.

Lourenco, S. F., & Longo, M. R. (2010). General magnitude representation in human infants. *Psychological Science*, *21*(6), 873–881.

Mastroianni, A., & Ludwin-Peery, E. (2022). Things could be better.

Meier, B. P., & Robinson, M. D. (2004). Why the sunny side is up: Associations between affect and vertical position. *Psychological science*, *15*(4), 243–247.

Meier, B. P., Robinson, M. D., & Caven, A. J. (2008). Why a big mac is a good mac: Associations between affect and size. *Basic and Applied Social Psychology*, *30*(1), 46–55.

Pezzelle, S., Bernardi, R., & Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition*, *181*, 117–126. https://doi.org/10.1016/j.cognition.2018.08.009

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*(6), 983–993.

Roese, N. J., Sanna, L. J., & Galinsky, A. D. (2005). The mechanics of imagination: Automaticity and control in counterfactual thinking. *The new unconscious*, 138–170.

Speer, R. (2022, September). *Rspeer/wordfreq: V3.0* (Version v3.0.2). Zenodo. https://doi.org/10.5281/zenodo.7199437

Srinivasan, M., & Carey, S. (2010). The long and the short of it: On the nature and origin of functional overlap between representations of space and time. *Cognition*, *116*(2), 217–241.

Varma, S., Sanford, E. M., Marupudi, V., Shaffer, O., & Lea, R. B. (2024). Recruitment of magnitude representations to understand graded words. *Cognitive Psychology*, *153*, 101673. https://doi.org/10.1016/j.cogpsych.2024.101673

Vierck, E., & Kiesel, A. (2010). Congruency effects between number magnitude and response force. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 204.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488.