PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Large numbers cause magnitude neglect: The case of government expenditures

Christina Boyce-Jacino[a,1], Ellen Peters[b], Alison P. Galvani[c], and Gretchen B. Chapman[a]

Edited by Susan Fiske, Princeton University, Princeton, NJ; received February 22, 2022; accepted May 27, 2022

Four studies demonstrate that the public's understanding of government budgetary expenditures is hampered by difficulty in representing large numerical magnitudes. Despite orders of magnitude difference between millions and billions, study participants struggle with the budgetary magnitudes of government programs. When numerical values are rescaled as smaller magnitudes (in the thousands or lower), lay understanding improves, as indicated by greater sensitivity to numerical ratios and more accurate rank ordering of expenses. A robust benefit of numerical rescaling is demonstrated across a variety of experimental designs, including policy relevant choices and incentive-compatible accuracy measures. This improved sensitivity ultimately impacts funding choices and public perception of respective budgets, indicating the importance of numerical cognition for good citizenship.

numerical cognition | policy | information presentation | numeracy

Significance

Comprehension of government expenditures requires understanding immense monetary amounts, yet numbers of such magnitude are difficult to understand. Our findings highlight the implications of basic numerical processing for participatory democracy. Basic principles of numerical cognition predict that a simple rescaling manipulation will increase nonexperts' ability to discriminate among different price tags for large government programs, a prediction that was supported in four experiments. By converting large numbers into smaller ones, regardless of their unit familiarity, people are better able to process numerical information and, subsequently, incorporate differences in budgetary magnitudes into their judgments and decisions.

In an interview about a Central American initiative, US President Joe Biden misquoted the program price tag as "almost $800 billion" when the true amount was $750 million.* Like Biden, many people confuse large budgetary amounts. In a set of four experiments, we demonstrate how difficulties in representing and reasoning about large numbers have consequences for evaluating government programs and how large numbers can be better presented so that people can use them effectively.

Numeracy and Its Implications for Citizenship

Difficulties in discriminating among large numbers stem from how numbers are cognitively represented. Following a logarithmic function, as numbers increase in magnitude, their internal representations become harder to distinguish because representations of large numbers are noisier and thus overlap more (1, 2). For example, $2 and $4 are perceived as further apart than $1,002 and $1,004 (3). Under a linear numerical representation, an absolute difference of $2 would be interpreted as the same in both cases, whereas under a logarithmic function, discrimination is a function of the ratio between two numbers. For federal expenditures, logarithmic number representations imply that people will discriminate more easily between small costs but will struggle to discriminate between large costs, displaying cost insensitivity instead. In this way, basic numerical processing has implications for participatory democracy.

The processing of numbers can be facilitated by manipulating how numerical information is presented. For instance, putting numbers into perspective or reexpressing unfamiliar numbers in familiar units can increase compression of numerical representations and thus increase discriminability among values (4, 5). For example, instead of describing an area as being "695,000 km^2," including a reference makes it easier to understand: 695,000 km^2 is about the size of Texas (5). Alternatively, numbers can be rescaled, for example, in terms of per household costs (6). Both strategies transform large numbers to smaller magnitudes, thus placing the values on an easier-to-discriminate portion of the numerical representation function and increasing sensitivity.

The Current Studies

In four preregistered studies, we assessed whether nonexperts' ability to discriminate between price tags for large government programs improves when prices are expressed as per capita values rather than national values. These studies demonstrate a simple way to remove a significant barrier to good citizenship. The current studies go beyond previous work (e.g., 6) by testing predictions derived from literature on basic processes in numerical cognition using paradigms that employ the timely context of the COVID-19

*See https://www.cnn.com/2021/03/17/politics/fact-check-biden-abc-stephanopoulos/index.html.

Author affiliations: [a]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213; [b]Center for Science Communication Research, School of Journalism and Communication, University of Oregon, Eugene, OR 97403; and [c]Center for Infectious Disease Modeling and Analysis, Yale School of Public Health, New Haven, CT 06510

Author contributions: C.B.-J. and G.B.C. designed research; C.B.-J. performed research; C.B.-J. and G.B.C. analyzed data; and C.B.-J., E.P., A.P.G., and G.B.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: christina.m.boyce-jacino.ctr@army.mil.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2203037119/-/DCSupplemental.

Published July 7, 2022.

pandemic, incentive-compatible judgments, objective assessment of judgment accuracy, and policy choices.

Researchers sometimes study how people process numbers in the context of numeracy. Broadly defined as mathematical and probabilistic reasoning skills, numeracy is typically measured with a series of mathematics problems (7, 8) and is critical to success in several domains (9–11). Following our preregistration, we included a measure of numeracy as a covariate in the current studies, and additional exploratory considerations of numeracy are included in *SI Appendix*.

Experiment 1. In March 2020, Amazon Mechanical Turk (MTurk) participants (n = 392) saw one of four statements about possible US COVID-19 relief packages. As shown in Fig. 1, the statements varied in their scope (national or individual) and magnitude (large or small) such that participants reading about a national stimulus package saw that "Congress passed a $100 billion [$2 trillion] relief package to address the Covid-19 national crisis." Those reading about an individual level stimulus package read that "The relief package passed by Congress to address the Covid-19 national crisis includes cash payments to individual taxpayers. Consider a payment of $1,200 [$24,000] per individual." Note that the ratio between stimulus amounts in each scale condition is held constant at 20:1. In each condition, participants rated the effectiveness of each program, defined as how well the program would address the economic impact of COVID-19.

Results. Fig. 2 and Table 1 show an interaction such that participants differentiated more between high and low individual-level payments than they did between high and low national stimulus packages ($\beta = -25.03$, SD = 5.40). These results suggest that people distinguish between two small numbers more easily than two large numbers even when the ratio is held constant. Note that we speak only to differences within national and individual conditions, not across. Because the national amounts represent the entire cost of the recovery package, which included far more than individual payments, while the individual amounts represent cash payments to individuals, the two conditions are not directly equivalent. Finally, we find that higher numeracy was correlated with lower ratings of effectiveness ($\beta = -3.21$, SD = 1.42). In *SI Appendix*, we report additional, exploratory analyses regarding the moderating effect of numeracy.

Given that both the national and individual conditions used a 20:1 ratio between magnitudes, the interaction pattern indicates that numerical representation must not have a logarithmic function as surmised by previous work. Instead, the function must be more curvilinear to account for current and prior results (6).

Experiment 2a. In our next experiment, we employ an incentive-compatible, objective measure of numerical understanding in the form of a recall ranking task. MTurk participants (n = 401) began by learning about the cost of eight

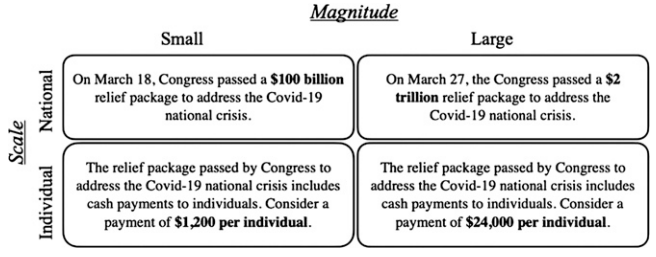

Fig. 1. Experiment design and text for Experiment 1.

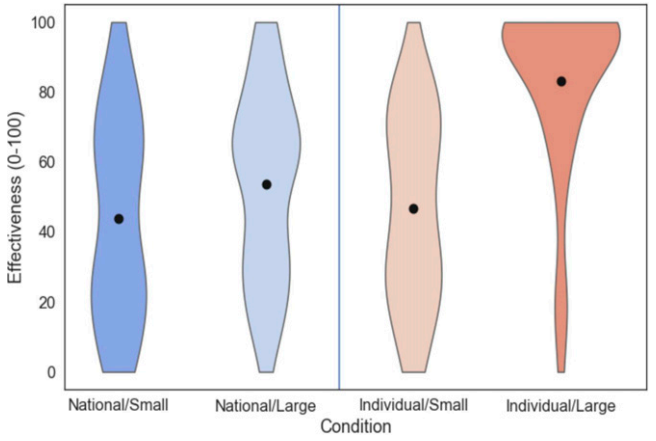

Fig. 2. Results for Experiment 1 showing rated effectiveness as a function of condition. Rating distributions are represented by violin plots where the width of the violin represents the frequency of data at each value and the center dot of each plot represents the mean effectiveness rating for that condition.

programs under time pressure (see Fig. 3*A*). Participants saw prices as national program costs (e.g., $3 billion) or in price per capita (e.g., a $3-billion program costs $10 per capita). After seeing each program's cost, participants ranked the set of eight programs by price (see *SI Appendix* for details). At the time of ranking, participants did not see the costs of the programs; rather, they had to recall them. Requiring that rank-order judgments be based on memory adds difficulty to an otherwise easy task, and while ranking costs does not require participants to preserve interval information, it does require them to properly encode program cost. Furthermore, the rank-order task was presented identically in the national and per capita conditions, and participants were paid for accuracy. We predicted higher rank-order accuracy in the per capita condition, reasoning that more overlap among large-number representations would cause greater confusion among numerical ranks.

Results. We scored participants for correctly ranking each of the 28 pairs that resulted from ordering eight programs. For example, if they ranked the most expensive program above the least expensive, they would be correct on that pair. As shown in Fig. 4*A*, participants in the per capita condition were more accurate, correctly ranking on average 19.22 (SD = 5.91) program pairs, compared to an average 17.90 (SD = 5.45) among national-cost participants [$t(399) = 2.30$, $P = 0.021$]. A regression analysis (Table 2) confirmed that accuracy increased when cost information was presented in per capita terms ($\beta = 0.262$, SD = 0.125). Numeracy correlated with greater accuracy ($\beta = 0.199$, SD = 0.068). Additional program features were explored in preregistered analyses (see *SI Appendix*).

Rescaling magnitudes into smaller units may make them more familiar; thus, it is possible that familiarity, and not rescaling per se, drives increased accuracy. To test this account, in Experiment 2b, we scaled costs by an arbitrary and unfamiliar "capitol dome" unit.

Experiment 2b. In a replication and extension of Experiment 2a, we test the robustness of using a rescaling rule to improve numerical processing by scaling down total expenditures using an unfamiliar unit. Our unit in this experiment is a capitol dome, equivalent to the estimated material costs of the US Capitol Building dome ($20 million) such that a program costing $1 billion would cost 50 capitol domes. In using this unit, we can directly test the role that familiarity or self-relevance (i.e., a per

Table 1. Linear regression results for Experiment 1

	(1) Rating	(2) Rating
Magnitude	36.26***	35.08***
	(3.89)	(3.87)
Scale	−2.95	−3.65
	(3.87)	(3.82)
Magnitude × scale	−26.41***	−25.03***
	(5.47)	(5.40)
Numeracy		−3.21*
		(1.42)
Politics		1.83
		(1.60)
White race		−7.81*
		(3.07)
Gender		2.56
		(2.92)
Age		−0.76
		(1.29)
Education		1.81
		(1.53)
N	392	392
R^2	0.248	0.283

We regressed on effectiveness rating (0 to 100, with higher numbers indicating greater effectiveness) with the primary predictors (1) and additional covariates (2) (numeracy, politics: 0 = conservative and 100 = liberal; gender: 0 = male and 1 = female). Regression coefficients are presented with standard errors in parentheses.
* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

capita rescaling rule) could play in increasing discriminability above and beyond the role of numerical processing.

MTurk participants ($n = 399$) completed the same survey as in Experiment 2a except that participants in the treatment condition saw program costs in terms of the dome unit. We predicted that participants in the dome condition would be better able to process the program cost information and would therefore be more accurate on the program ranking task.

Results. As in Experiment 2a, for each participant, we created 28 ranked pairs from our eight programs and scored participants based on whether or not they ranked the programs in each pair correctly. We find first that participants in the per capita condition were more accurate, correctly ranking on average 19.9 (SD = 5.86) program pairs, compared to on average 18.3 (SD = 5.32) for national-cost participants [$t(399) = 2.75$, $P = 0.006$]. A mixed-model logistic regression confirmed that condition significantly affected pair score, with scores higher in the dome condition than the control (β = 0.286, SD = 0.094). Numeracy played a role as well, with those scoring higher on numeracy having higher rank accuracy (β = 0.197, SD = 0.052). Additional analyses are reported in *SI Appendix*.

The results of this experiment suggest that a rescaling rule that simply transforms large numbers into smaller ones will benefit participants on tasks involving numerical processing. This effect stands regardless of familiarity with a unit.

Experiment 3. Finally, we investigated whether presenting smaller magnitudes also alters support for federal programs. In this experiment, our measure of numerical discrimination focused on whether participants chose to fund the less expensive of two purportedly equivalent-impact programs. MTurk participants ($n = 399$) responded to eight pairs drawn from eight federal programs by choosing to fund one program in each pair. Four of the eight pairs were matched on all qualities (e.g., scope, target domain, effectiveness) except for price, in an effort to highlight cost as the singular important feature. The remaining pairs were mixed such that we would expect participants to use cues beyond price in their decisions. Counterbalanced assignment of price to program ensured that program details other than price could not result in a net preference.

As before, we presented cost information to participants either as the national program costs (e.g., $300 million) or in per capita costs (e.g., $1 per capita). We predicted that participants would select the less expensive program more frequently in the per capita condition compared to the national-price condition.

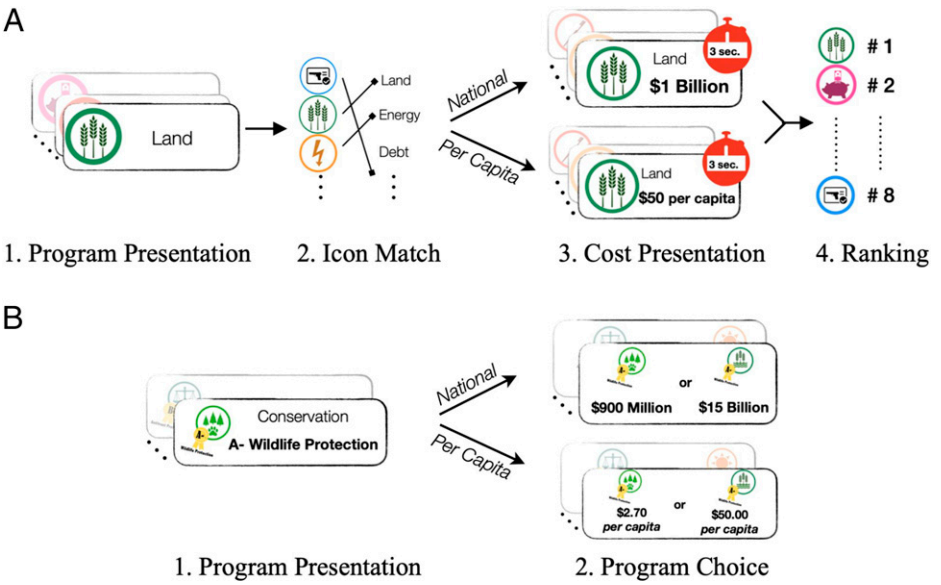

Fig. 3. Survey progression for Experiment 2a (*A*) (the procedure for Experiment 2b is not shown but is equivalent to that of 2a) and Experiment 3 (*B*). (*A*) The progression for Experiment 2a is as follows: in part 1, participants begin by seeing the description and program icon for each program. They then complete an attention measure where they match each program icon to the program name. In part 3, they are presented with cost information under time pressure. In part 4, they rank each program according to its cost, as they remember it. In Experiment 2b, program costs are presented in a domes condition instead of per capita. (*B*) Experiment 3 proceeds similarly where participants begin (part 1) by learning the program name and description. They then proceed directly to learning the cost information in part 2, where they are also asked to choose which of the two programs to fund.

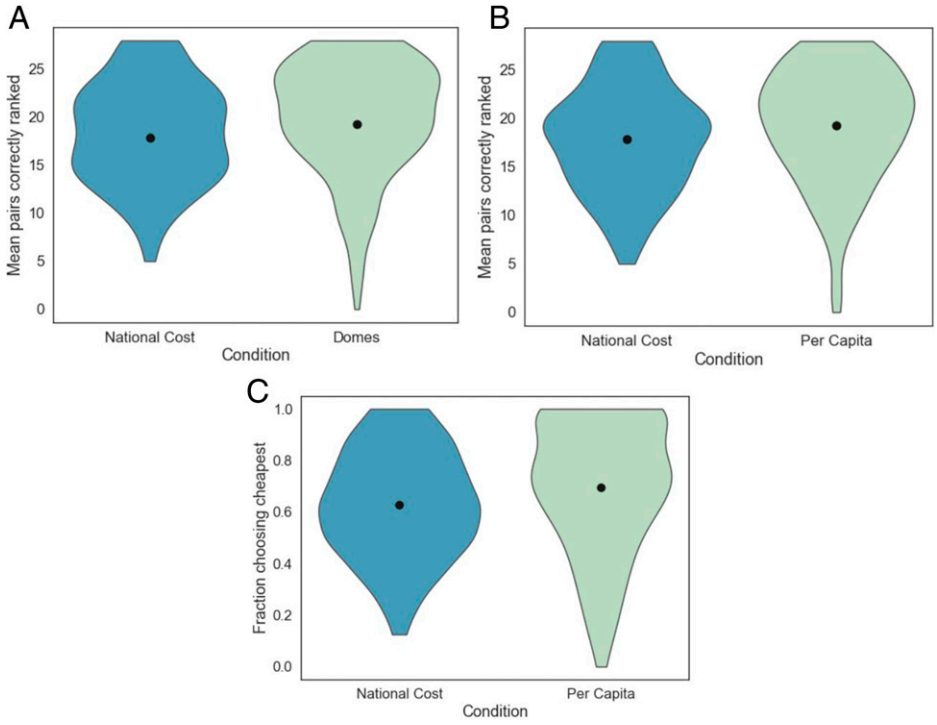

Fig. 4. Results for Experiment 2a (*A*), Experiment 2b (*B*), and Experiment 3 (*C*). *A* and *B* show participants' rank-order responses depicted as the number of pairs (out of 28) ordered correctly. *C* shows the fraction of pairs (out of eight) where the participant chose the less expensive program. All panels present violin plots in which the width of the curve corresponds with the frequency of data points and in which the center dot represents the mean of the distribution.

Results. As shown in Fig. 4*B*, participants chose the cheaper program (our measure of numerical discrimination) more often in the per capita condition (M = 0.694, SD = 0.255) than the national-cost condition (M = 0.629, SD = 0.209; β = 0.397, SD = 0.103; see Table 2). Greater numeracy also predicted more choices of the less expensive program (β = 0.198, SD = 0.059), as did better performance on an attention-check measure (β = 0.217, SD = 0.044). Although the present results include both choice pairs matched on all nonprice qualities as well as choice pairs that were unmatched, in *SI Appendix* we present additional analyses that explore this specific feature of the program pairs.

Discussion

Numerical comprehension is a basic building block of good decision-making, and our work demonstrates the limitations of that comprehension, due to fundamental properties of numerical cognition. Basic principles of numerical cognition lead directly to predictions about a manipulation that improves discrimination among large magnitudes. Across four studies, we found robust evidence for a rescaling manipulation which improves discrimination among price tags for large government programs. Specifically, scaling down large numbers caused greater price discrimination in program-effectiveness ratings, improved rank ordering of program magnitudes, and led to greater preference for less expensive policies intended to be equivalent otherwise. This relatively simple change to how information is presented ameliorates misunderstandings, thereby improving the potential for participatory democracy.

The current work offers the following insights. First, Experiment 1 replicated the central findings of Saiewitz and Piercey (6) and demonstrated that rescaling's effect persisted for smaller ratio differences. Our study had the added contributions of employing a diverse US sample, judgments of policy impact, and the highly important context of COVID-19 aid, where motivated processing could have increased sensitivity, regardless of magnitude. Experiment 1 also demonstrated an important boundary condition on numerical representation effects. Despite a constant ratio between numbers, participants were more sensitive to cost magnitude at the individual than national scale, a finding that has implications for the functional form of numerical cognitive representation.

In Experiments 2a and 2b, we employed an incentive-compatible recall ranking task with an objectively correct answer to explore how information presentation can impact numerical discrimination. Incentive-compatible tasks have not previously been employed in numerical cognition studies, and the impact of the rescaling manipulation in this task despite the incentive for accuracy suggests that noisy processing of very large numbers is due to cognitive mechanisms, not low motivation to engage in the task. Even though ranking should have preserved ordinality in both conditions, we see lower accuracy in the national-level condition, suggesting that noisy representations of larger numbers produce greater confusion between budget numbers (2). If a noisy numerical representation is conceptualized as a distribution centered around the precise value, our results can be explained by positing that the variance in the distributions increases more than proportionally with magnitude. An analogous account is that people encode a fuzzy trace, or gist, rather than the precise numerical value (12). Experiment 2b demonstrates that the influence of rescaling is due to numerical magnitude rather than effects of self-relevance or familiarity.

The findings from Experiments 2a and 2b indicated a robust rescaling effect even though participants were able to compare programs to one another. Extant research on separate versus joint evaluations (13, 14) suggests that we would expect greater

Table 2. Mixed-effects logistic regression results for Experiments 2a, 2b, and 3

	Experiment 2a	Experiment 2b	Experiment 3
	(1)	(2)	(2)
	Score	Score	Choice of cheaper
Condition	0.259*	0.286**	0.471***
	(0.107)	(0.094)	(0.117)
Numeracy	0.235***	0.198***	0.222**
	(0.063)	(0.052)	(0.066)
Attention check	0.121	0.063	0.245***
	(0.062)	(0.053)	(0.063)
Politics	0.086	0.056	−0.001
	(0.054)	(0.045)	(0.058)
Age	−0.057	0.017	0.114*
	(0.045)	(0.033)	(0.048)
Gender	−0.201*	−0.099	−0.010
	(0.098)	(0.089)	(0.113)
N	11,228	11,172	3,192
Clusters	401	399	399

In Experiments 2a and 2b, we predicted score (1 = correct and 0 = incorrect) at the level of each ranked pair of programs (28 per participant). In Experiment 3, we examined participant choices on the level of each program pair (eight observations per participant; 1 = chose cheaper program and 0 = chose more expensive program). In all three models, covariates included condition (total cost vs. per capita or dome presentation), numeracy, an attention-check measure, and a set of demographics (political orientation: 0 = conservative and 100 = liberal; gender: 0 = male and 1 = female). The analysis is clustered on the level of participant. Regression coefficients are presented with standard errors in parentheses. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

numerical sensitivity in a comparative setting regardless of numerical presentation. Nonetheless, we found a significant effect of rescaling, suggesting an important role of scaling in numerical sensitivity.

Finally, Experiment 3 again adopted a joint evaluation setting but assessed how per capita presentation affects policy choices of which government programs to fund. A counterbalanced study design allowed us to draw inferences about numerical discrimination from choice of the less expensive program. Presenting large monetary amounts in a way better understood by participants enabled them to better discriminate between programs.

Together, this work both illuminates basic cognitive processes and also contributes to our understanding of barriers to good citizenship, demonstrating the importance of presenting information in a manner amenable to the workings of the human mind. We demonstrate that rescaling large numbers facilitates improved decision-making by allowing people to better recognize the difference between two numbers, thus allowing for improved discrimination. Extensions of the present work should explore scaling effects in other domains requiring communication of large numbers, such as public health. Future work could also explore other ways to improve discriminability, such as experience. Whereas the numerical cognition literature suggests that the logarithmic representation of numbers is fixed in adults, research on the role of experience, such as decisions by sampling (15), suggests that comprehension of large numbers might be improved by simply experiencing more large numbers. More broadly, future work could investigate the ancillary effects of scaling large numbers. For instance, although we propose that transforming a program cost in the billions to one in the tens improves how well numerical information is understood and used, the use of smaller numbers may also alter perceptions of program benefits. Results from such studies would enhance our ability to provide prescriptive advice for how to help lay people be more actively engaged citizens.

Materials and Methods

All experiments were carried out with the approval of Carnegie Mellon University's Institutional Review Board (Protocol ID STUDY2017_00000392). Participants gave fully informed consent prior to taking part in each survey.

Experiment 1.

Participants. US-based participants ($n = 404$) were recruited on MTurk in April 2020 and completed the survey for monetary compensation. As per preregistration (see https://aspredicted.org/blind.php?x=9r9wm6) (16), 12 participants were removed from analysis for failure to pass an attention check, leaving 392 in the analysis. Of the participants, 56.6% were male and 41.6% were 30 to 39 y old.

Design and procedure. Participants were randomly assigned to one of four conditions in a between-subjects design, 2 (scale: national vs. individual) × 2 (magnitude: large vs. small), and read a description of COVID-19 stimulus package legislation (see Fig. 1). All participants answered the question: "How effective do you think this relief package will be in addressing the economic impact of the Covid-19 national emergency?" (0 to 100 scale). They also completed a 13-item numeracy test, consisting of an 11-item test (7) and two original questions using numbers in the millions and billions, as well as a demographic questionnaire (political affiliation, gender, age, education, race).

Experiment 2a.

Participants. US-based MTurk participants ($n = 401$) completed the survey for monetary compensation. As per preregistration (see https://osf.io/nbwsy/?view_only=02e8b6b177d44acea5be39ed52fc3e12) (17), no exclusion criteria were applied. Of the participants, 51.8% were male and 43.1% were 30 to 39 y old.

Design and procedure. Participants were randomly assigned to one of two conditions (national cost vs. per capita) in a between-subjects design. Their task was to rank order a set of government programs according to cost. In the national-cost condition, participants saw the cost in numerical form (e.g., $600 million), and in the per capita condition, cost was in terms of a per capita cost. We approximated the US population to be 300 million, so a program costing $600 million would cost $2 per capita.

The experiment proceeded in four parts (see Fig. 3*A*). In part 1, participants were familiarized with the names of eight programs (see *SI Appendix* for details) and a representative icon for each. Each program was presented individually, and to advance, participants had to select the true name of the program from a list. In part 2, the participants matched each program with its corresponding icon. This task served as our attention-check measure. In part 3, participants learned how much each program cost. This information was given individually for each program and under time pressure: participants had 3 s to view the name, icon, and price tag. In part 4, participants then completed an incentivized task in which they ranked the programs according to cost from memory. Participants received a $5 bonus if their rank order was completely correct (34 received this bonus). Finally, they completed the same numeracy and demographic items as in Experiment 1.

Experiment 2b.

Participants. US-based MTurk participants ($n = 399$) completed the survey for monetary compensation. No exclusion criteria were applied, as per preregistration (see https://osf.io/kmdzu/?view_only=f3e8da71bacc4655bade7f07f7b8cf0a) (18).

Design and procedure. The design and the procedure of Experiment 2b were the same as those for Experiment 2a except that a different rescaling rule was used. In this experiment, we scaled federal budget items by a capitol dome unit equal to $20 million. We applied this rule such that a program costing $40 million would cost 2 capitol domes. Across all eight programs, dome costs ranged from 1 to 750 capitol domes.

Experiment 3.

Participants. US-based MTurk participants ($n = 399$) completed the survey for monetary compensation. No exclusion criteria were applied, as per preregistration (see https://osf.io/tacdk/?view_only=d72695169ed6495a931ccc7414f3e630) (19). Of the participants, 46.6% were male and 33.3% were 30 to 39 y old.

Design and procedure. Participants were randomly assigned to one of two conditions (national cost vs. per capita) in a between-subjects design. As in Experiment 2a, participants saw cost information for eight federal programs where cost was either in total numerical form (national-cost condition: e.g., $600 million) or in terms of per capita costs (per capita condition: e.g., $2 per capita). Programs were designed such that each million-dollar program had a similar program costing billions. These matched programs were similar on all dimensions except cost (see *SI Appendix* for details). To communicate this equivalence, we assigned each program an efficacy grade within a domain. For example, the wind power program had an "A−" in the domain of "renewable energy." Its partner program, solar energy, had the same letter grade and domain.

In part 1, participants saw a program name, description, icon, and information about its effectiveness. For each program, they were asked to select the correct performance "letter grade" and evaluation domain from a list. Their answers to those questions serve as our attention-check measure. In the second part, participants saw eight pairs of programs with corresponding costs. Each pair consisted of programs either matched (same domain, one price tag in the millions, one in the billions, and same program efficacy grade) or unmatched (different domains, one cost millions and one billions, but they had different efficacy grades). All participants saw four matched programs and four pairs of unmatched programs which were chosen randomly from the set of all possible 12 unmatched pairings (see *SI Appendix* for analysis of matched vs. unmatched pairs). For each pair, they selected which of the two programs they supported funding. After making their eight choices, participants completed a numeracy scale and a demographic questionnaire.

Data Availability. Data for experiments 1 through 4 data have been deposited in the Open Science Framework (OSF; https://osf.io/a3pe9/?view_only=011d0014f6234f739b5a4d502eb73b76) (20). Preregistration template data are available at As Predicted for Experiment 1 (https://aspredicted.org/blind.php?x=9r9wm6) (16) and OSF for Experiments 2a (https://osf.io/nbwsy/?view_only=02e8b6b177d44acea5be39ed52fc3e12) (17), 2b (https://osf.io/kmdzu/?view_only=f3e8da71bacc4655bade7f07f7b8cf0a) (18), and 3 (https://osf.io/tacdk/?view_only=d72695169ed6495a931ccc7414f3e630).

ACKNOWLEDGMENTS. Financial support for this study was provided in part by NSF Grants SES-1948887 awarded to C.B.-J., SES-2017651 and SES-1558230 awarded to E.P., and SES-1851702 awarded to G.B.C. The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. Timothy Liu assisted with Experiment 1.

1. S. Dehaene, V. Izard, E. Spelke, P. Pica, Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science* **320**, 1217–1220 (2008).
2. C. R. Gallistel, R. Gelman, *Mathematical Cognition* (Cambridge University Press, 2005).
3. E. Peters, P. Slovic, D. Västfjäll, C. K. Mertz, Intuitive numbers guide decisions. *Judgm. Decis. Mak.* **3**, 619–635 (2008).
4. P. J. Barrio, D. G. Goldstein, J. M. Hofman, "Improving comprehension of numbers in the news" in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2016), pp. 2729–2739.
5. J. Hullman, Y. S. Kim, F. Nguyen, L. Speers, M. Agrawala, "Improving comprehension of measurements using concrete re-expression strategies" in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2018), pp. 1–12.
6. A. Saiewitz, M. D. Piercey, Too big to comprehend? A research note on how large number disclosure format affects voter support for government spending bills. *Behav. Res. Account.* **32**, 149–158 (2020).
7. I. M. Lipkus, G. Samsa, B. K. Rimer, General performance on a numeracy scale among highly educated samples. *Med. Decis. Making* **21**, 37–44 (2001).
8. L. M. Schwartz, S. Woloshin, W. C. Black, H. G. Welch, The role of numeracy in understanding the benefit of screening mammography. *Ann. Intern. Med.* **127**, 966–972 (1997).
9. T. Davis, E. M. Kennen, J. A. Gazmararian, M. V. Williams, "Literacy testing in health care research" in *Understanding Health Literacy: Implications for Medicine and Public Health*, J. G. Schwartzberg, J. B. VanGeest, C. C. Wang, Eds. (American Medical Association Press, Chicago, IL, 2005). p. 15779.
10. V. F. Reyna, W. L. Nelson, P. K. Han, N. F. Dieckmann, How numeracy influences risk comprehension and medical decision making. *Psychol. Bull.* **135**, 943–973 (2009).
11. T. Låg, L. Bauger, M. Lindberg, O. Friborg, The role of numeracy and intelligence in health-risk estimation and medical data interpretation. *J. Behav. Decis. Making* **27**, 95–108 (2014).
12. V. F. Reyna, P. G. Brust-Renck, How representations of number and numeracy predict decision paradoxes: A fuzzy-trace theory approach. *J. Behav. Decis. Making* **33**, 606–628 (2020).
13. C. K. Hsee, J. Zhang, General evaluability theory. *Perspect. Psychol. Sci.* **5**, 343–355 (2010).
14. C. K. Hsee, The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organ. Behav. Hum. Decis. Process.* **67**, 247–257 (1996).
15. N. Stewart, N. Chater, G. D. Brown, Decision by sampling. *Cognit. Psychol.* **53**, 1–26 (2006).
16. C. Boyce-Jacino, E. Peters, A. P. Galvani, G. B. Chapman, As Predicted: Numeric processing of the 2020 stimulus package (#38306). AsPredicted, Wharton Credibility Lab, University of Pennsylvania. https://aspredicted.org/blind.php?x=9r9wm6. Deposited 31 March 2020.
17. Millions Billions: Federal Program Ranking Task, Preregistration Template from AsPredicted.org, doi: 10.17605/OSF.IO/NBWSY. Open Science Framework (OSF). https://osf.io/nbwsy?view_only=02e8b6b177d44acea5be39ed52fc3e12. Deposited 3 December 2020.
18. Millions Billions: Federal Program Ranking Task, Preregistration Template from AsPredicted.org, doi: 10.17605/OSF.IO/KMDZU. Open Science Framework (OSF). https://osf.io/kmdzu?view_only=f3e8da71bacc4655bade7f07f7b8cf0a. Deposited 8 October 2020.
19. Millions Billions: Federal Program Ranking Task, Preregistration Template from AsPredicted.org, doi: 10.17605/OSF.IO/TACDK. Open Science Framework (OSF). https://osf.io/tacdk?view_only=d72695169ed6495a931ccc7414f3e630. Deposited 3 December 2020.
20. Millions Billions: Federal Program Ranking Task, Open Science Framework (OSF). https://osf.io/a3pe9/?view_only=011d0014f6234f739b5a4d502eb73b76. Deposited 21 August 2020.