To Put That in Perspective: Generating Analogies That Make Numbers Easier to Understand

Christopher Riederer
Columbia University
mani@cs.columbia.edu

Jake M. Hofman
Microsoft Research
jmh@microsoft.com

Daniel G. Goldstein
Microsoft Research
dgg@microsoft.com

ABSTRACT

Laypeople are frequently exposed to unfamiliar numbers published by journalists, social media users, and algorithms. These figures can be difficult for readers to comprehend, especially when they are extreme in magnitude or contain unfamiliar units. Prior work has shown that adding "perspective sentences" that employ ratios, ranks, and unit changes to such measurements can improve people's ability to understand unfamiliar numbers (e.g., "695,000 square kilometers is about the size of Texas"). However, there are many ways to provide context for a measurement. In this paper we systematically test what factors influence the quality of perspective sentences through randomized experiments involving over 1,000 participants. We develop a statistical model for generating perspectives and test it against several alternatives, finding beneficial effects of perspectives on comprehension that persist for six weeks. We conclude by discussing future work in deploying and testing perspectives at scale.

ACM Classification Keywords

J.4 Social and Behavioral Sciences: Psychology; H.1.2 User/Machine Systems: Human factors, Human information processing; H.5.2 User Interfaces: Evaluation/methodology; J.7 Computers in other systems: Publishing

Author Keywords

Numeracy; statistics; education; measurement; experimentation

INTRODUCTION

Understanding numerical measurements is a crucial skill in a data-driven, online world. Inadequate reasoning about magnitudes, one manifestation of poor numerical literacy, can have negative impacts on individuals' reasoning about finances, medical care, sustainability, and the ability to differentiate between honest reporting and "fake news". From the Public Editor of the New York Times[1] to science writers like John Allen Paulos [16], there has a been a call for scientists, journalists, and policy makers to help readers make sense of published statistics by putting unfamiliar measurements in perspective. Related research in judgment and decision making has found that numerical re-expressions can improve or influence decisions, such as expressing fuel economy in "gallons per 100 miles" instead of "miles per gallon" [12], translating electricity usage into monthly costs [14], and expressing food calories as amounts of physical exercise [7]. Likewise, risk communication scholars have documented that probability re-expressions can help people appreciate medical, financial, and environmental risks [9, 15].

Recent work has generalized these examples by creating a template-based framework of *perspective sentences* that employ ratios, ranks, or unit changes to provide context around arbitrary numerical measurements [1]. Take, for instance, the area of a country such as Pakistan, which is about 307,000 square miles. Most readers have little intuition for the size of a square mile, let alone several hundred thousand of them [4], and so it can be useful to rephrase this measurement in terms of a more familiar unit such as (for Americans) the size of a U.S. state. Indeed, as found in [1], such perspectives help people remember numbers they have read, estimate unknown amounts, and detect errors in potentially erroneous statements.

But even for this relatively straightforward example, there are many different perspectives that might be used to help people comprehend a number and it is unclear *a priori* how to choose among them. For example, when communicating the size of Pakistan to a U.S. reader, its area could be phrased as "about twice the size of California", "five times larger than Georgia", or "ten times the size of South Carolina". Which of these is best, and can the features of good perspectives be identified so they can be generated at scale?

In this paper we systematically explore what makes some perspectives more helpful than others through a series of randomized, online experiments involving over 1,000 participants in which we construct and compare several automated methods for generating numerical analogies. We do so by focusing on an example domain of perspectives that illustrate country populations and areas in terms of U.S. states, varying the states and multipliers presented to participants and measuring their ability to estimate the corresponding country statistics.

Although this domain represent but a small fraction of scenarios where perspectives might be deployed, there are multiple benefits to studying it. First, country-level statistics are frequently mentioned in the news and are among the top reference questions that users ask of search engines. As such, there is

[1] http://nyti.ms/1oe6DZo

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04...$15.00

DOI: https://doi.org/10.1145/3173574.3174122

great value to be had in improving the communication of these numbers. Moreover, limiting our attention to this relatively narrow but important domain allows us to thoroughly explore the design space for constructing perspectives. By independently varying all elements of these numerical analogies, we can disentangle the effects of different ratios and reference quantities (e.g., states) to gain more general insight into what makes for useful perspectives across a range of domains.

In our first experiment we collected data on participants' familiarity with all 50 U.S. states and then had them estimate state populations or areas multiplied by a scaling factor (e.g., "If a country had an area that was 2 times larger than the area of Montana, what would its area be?"). We used these estimates to build a model that predicts the most useful perspective for a given statistic, which suggests that people perform better with well-known, accurately perceived reference quantities and simple scaling factors. This implies that if, for instance, one is estimating the area of Pakistan, it is more helpful to think of it as "twice the size of California" instead of "twice the size of Montana" or "five times larger than Georgia", even though the latter two statements are factually more accurate than the former.

Our second experiment tested this model against several other automated policies for generating perspectives. We asked a separate pool of individuals to estimate the size and population of countries when randomly assigned to perspectives based on: (1) the best perspective from our model, (2) the participant's home state, (3) the perspective with the lowest objective error, (4) a perspective with worse modeled error as a robustness check, and (5) no perspective, as a control. We find statistically and practically significant improvements in reader comprehension when perspectives are deployed for all four policies that we tested compared to the control condition. Furthermore, we find that one can choose freely from the above policies without severely compromising these benefits. This experiment also confirms that the model's predictions play out: the most accurate perspective is not always the best one, as imperfect analogies can communicate better than exact ones.

In our third and final experiment we focused on potential long-term benefits of perspectives. We contacted participants from the second experiment six weeks after the study had concluded and asked them to re-estimate the exact same statistics for the exact same countries as in the prior experiment, but this time no perspectives were shown to any participants. Remarkably, those who had previously been randomly selected to see perspectives in the original study continued to perform better than those in the control condition. As a result, we find that not only do perspectives help people understand unfamiliar measurements in the short term, but they also have a lasting impact on long-term reader comprehension.

Finally, we deployed the perspectives generated by our model to the Bing search engine to improve the quality of "instant answer" search results. Prior to our work a query such as "area of Pakistan" would show an answer of "307,373 square miles". We modified results to add a simple perspective to all area-related queries, so that this answer, for example, now includes the phrase "about twice the size of California", seen by all U.S. users.

In the remainder of the paper we first review related work and then provide more details on each of these experiments and their results.

RELATED WORK

Our work builds upon past research that has explored methods for generating and evaluating perspectives.

Past work by Barrio, Goldstein, and Hofman [1] provided broad support for the usefulness of perspectives across a range of domains, but relied on a crowdsourced incentive scheme wherein humans were payed small amounts to manually generate perspectives. Their experimental studies were centered around a dozen different examples with one perspective for each, offering relatively little in the way of explaining why some perspectives are more useful than others or designing policies for automatically generating perspectives.

In parallel, Kim, Hullman, and Agrawala [11] developed a browser plugin to generate visual, personalized numerical analogies for distances and areas mentioned in news stories. They constructed a model that selects popular landmarks as reference entities, preferring those that are close to the user's stated location and are related to the number in question by a multiplier value between zero and one to those between one and ten, and so on. A randomized trial showed that users preferred viewing news stories that contain these personalized spatial analogies to a control condition without them. This policy also compared favorably to a global one in which all users are assumed to be located at the Empire State building in New York City. A more general framework for using visual representations to improve numerical comprehension is specified in Chevalier, Vuillemot and Gali [6].

Subsequently, Chaganty and Liang [5] focused on automatically generating naturally phrased perspective sentences for a broader range of domains. They did so by combining a small set of reference statistics across different dimensions to rephrase arbitrary statistics (e.g., "131 million dollars $\approx$ annual employee salary $\times$ population of Texas $\times$ 30 minutes"). A recursive neural network was built to translate the formulas for these "compositional perspectives" to more naturally phrased sentences (e.g., "131 million dollars is about the cost to employ everyone in Texas over a lunch period"). A user study found that people preferred the perspectives generated by this system to those from simpler baselines, and that the neural network scored well on standard metrics used in evaluating machine translations.

Our research deviates from these studies in important respects. Our primary goal is not only to automate the generation of perspectives, but also to learn and explain what factors can be empirically shown to improve comprehension (as opposed to liking). Existing research tackles the former but places less emphasis on the latter. For instance, [11] used a manually specified function to rank potential perspectives, whereas we learn this function from user feedback. And while [5] learned a ranking function from user feedback, this function is based upon what perspectives people *prefer* as opposed to how perspectives *impact comprehension*. Towards this end, rather than asking users what perspectives they like best, we extend the approach taken in [1] and evaluate policies for designing perspectives based on the extent to which they improve people's estimates of unfamiliar numbers.

DESIGNING PERSPECTIVES

To make the tasks of understanding and designing perspectives tractable, we focused our attention on explaining unfamiliar geographic entities in terms of more familiar ones. Specifically, we looked at policies for constructing numerical analogies that rephrase the population and area of different countries in terms of U.S. states. Although this is just one example domain, it is both easy to work with and highlights a number of potentially important decisions when designing perspectives more generally.

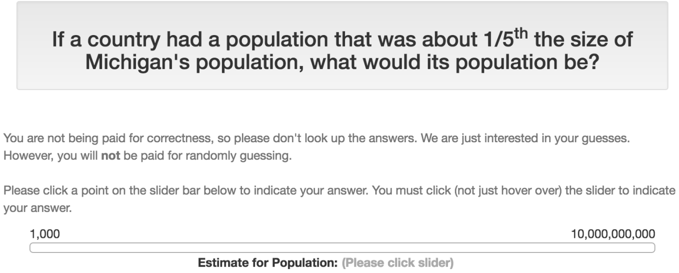
For example, suppose an author wishes to put Pakistan's size into perspective by expressing it as a multiple of a U.S. state. In theory the answer is easy: given a reference state (e.g., California), we simply divide to calculate a multiplier (e.g., 1.87 times larger) that correctly relates the area of country to the state. But in practice there are several factors to consider, including the choice of a reference state, the relative ease of working with its corresponding multiplier, and the objective error introduced by any approximations made in the comparison. Ideally one would select a state which is both very familiar and an exact match for the country's population using a multiplier of one, but in practice such matches are uncommon and trade-offs must be made between these three factors.

When such ideal comparisons are not possible, we could of course use the exact multiplier in a perspective (e.g., "Pakistan is 1.87 times larger than California"), but past work in cognitive psychology shows this to be a poor choice as people have a substantially easier time reasoning with round numbers between one and ten [2,3,8,17]. Following this work, we restrict our multipliers to the set $\{1,2,5,10\}$ and round to the nearest allowed multiplier to arrive at an approximate perspective (e.g., "Pakistan is about two times larger than California"). When making comparisons to small countries, reciprocals of these multipliers are employed (e.g., "the population of Montenegro is about 1/10th of Missouri's population"). This approach results in a total of 700 possible perspectives with three different elements: 2 dimensions (area or population), 7 different multipliers ($\{\frac{1}{10},\frac{1}{5},\frac{1}{2},1,2,5,10\}$), and 50 U.S. states. Given the area or population of a country, we restrict this set to the perspectives that are within 10% error of the country's actual area or population.

In the experiments that follow we test different policies for selecting perspectives from these alternatives to automatically rephrase any given country statistic.

EXPERIMENT 1:
IMPACT OF PERSPECTIVE COMPONENTS

The purpose of this experiment was to understand what makes for a good (or bad) perspective, in this case by looking at the choice of U.S. states and multipliers used to create numerical analogies for country statistics. Are there particularly good or bad reference entities (states) and which multipliers are easiest for people to understand and work with when estimating an unknown quantity?



If a country had a population that was about 1/5th the size of Michigan's population, what would its population be?

You are not being paid for correctness, so please don't look up the answers. We are just interested in your guesses. However, you will **not** be paid for randomly guessing.

Please click a point on the slider bar below to indicate your answer. You must click (not just hover over) the slider to indicate your answer.

Estimate for Population: (Please click slider)

Figure 1. An example of the interface shown in our first experiment where participants where asked to estimate the areas or populations of hypothetical countries in terms of U.S. states.

In order to isolate the effects of different states and multipliers, we designed an experiment that randomly varied these elements when asking participants to estimate the populations or areas of *hypothetical* countries. For example, "If a country had a population that was about 1/5th the size of Michigan's population, what would it's population be?", as pictured in Figure 1. This has two desirable properties. First, it allowed us to examine arbitrary combinations of states and multipliers without the constraints imposed by considering a specific target country, or the biases introduced by interactions between the country and the perspective elements. Second, randomly varying these two elements removes any potential confounds in estimating individual state or multiplier effects. Before or after this task, participants were also asked to make *direct estimates* the population or area of states (e.g., "What is the population of Michigan"?) to establish a baseline for how familiar people are with these reference quantities. To discourage people from simply looking up answers, each question reminded participants that we were interested in their best educated guess and that they would be paid based on their effort to honestly complete the task rather than on the correctness of their responses.

Participants. Participants were 341 workers on Amazon's Mechanical Turk platform [13] who were paid \$1.50 for participation. Participants were restricted to individuals living in the United States with high ($> 95\%$) approval ratings on previous Mechanical Turk tasks as reported by the platform.

Design. Participants were randomly assigned to either answer questions about area (n=174) or population (n=167). Within the area group, 88 were randomly assigned to provide direct state estimates first, while 86 were asked to estimate hypothetical country areas first. Within the population group, these figures were 84 and 83, respectively. Stimuli comprise all 50 U.S. states and the multipliers 1/10, 1/5, 1/2, 1, 2, 5, and 10.

Procedure. At the start of the experiment, each participant reported their age, gender, zip code, and the measurement system (imperial or metric) that they were most familiar with.

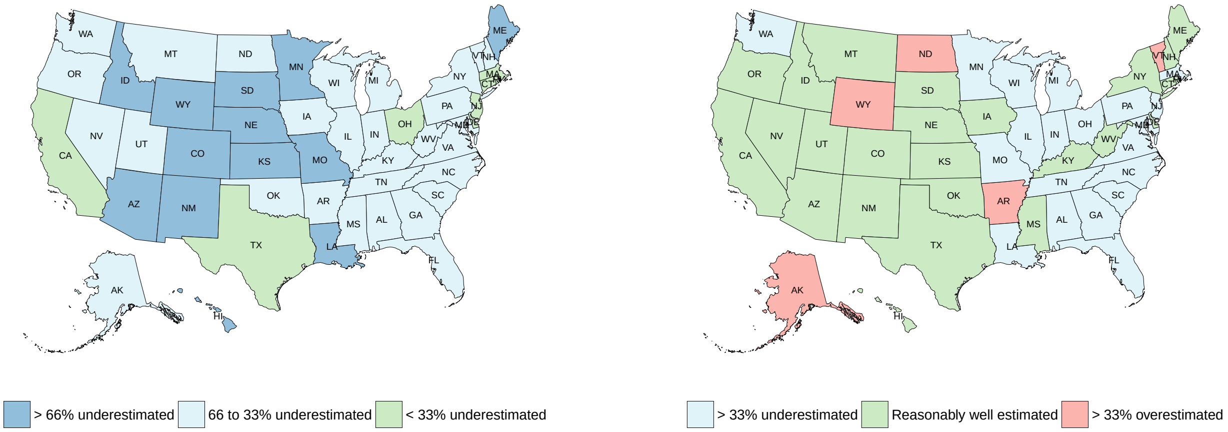


Figure 2. Perceived area (left) and population (right) by state. Cooler colors indicate underestimates and warmer colors show overestimates. Area tends to be underestimated for most states, with a few notable exceptions such as California and Texas. Population estimates are more evenly distributed, with many of the more sparsely populated states such as Vermont and Wyoming showing overestimation.

Next they rated their familiarity with each of the 50 U.S. states on a three point scale ("not familiar", "somewhat familiar", "very familiar"), where states were presented randomly to mitigate systematic ordering or fatigue effects.

Participants were then assigned a total of 21 different states and asked to make estimates based on the area or population of each, depending on the condition to which they were randomly assigned. Each participant received a customized set of 21 states based on the familiarity ratings they provided in the first part of the experiment, where the set of 21 aimed to include seven "not familiar" states, seven "somewhat familiar" ones, and seven "very familiar" states, sampling from the other categories when needed.

Estimation consisted of two counter-balanced phases. In the hypothetical estimation phase, each of these 21 states was paired with a randomly chosen multiplier from the set of seven multipliers mentioned above, and participants were asked to estimate the area or population of the corresponding hypothetical country. In the direct estimation phase, participants saw the same 21 states as in the hypothetical estimation phase without any multipliers and were simply asked for estimates of state areas or populations. All questions were randomly ordered within each phase.

Results

Having independently varied reference entities (states) and multipliers enables us to measure how both of these factors affect the accuracy of hypothetical country statistic estimates. As others have found, people's estimates of such statistics often vary by several orders of magnitude—both from each other and from the truth—rendering the usual process of calculating raw averages ineffective. To deal with the wide range of responses, we follow [4] and compute all summary statistics in log-space, exponentiating after computing means and standard deviations. For instance, to compute the average perceived population of a state, we first take the log of each participant's estimate, then calculate the average across all responses, and, finally, exponentiate the result. Likewise, we measure the difference between estimated and actual values in log-space to quantify how many orders of magnitude off participants are in estimating any given quantity, looking at both signed and unsigned versions of this measure. When displaying results in the text or figures we transform all summary statistics back to an interpretable scale by exponentiating. We follow the same procedure throughout our analyses of this and each subsequent experiment.

We begin by examining how accurate people are in directly estimating the area and population of the 50 U.S. states. We first compute the average perceived population and area of each state as described above, and then compare each quantity to its true value, calculating the percent error as $(perceived - actual)/actual$. Figure 2 shows the results in a map of the 50 states and the degree to which its dimensions (area or population) were over- or under-estimated by participants. We see substantial differences between the true and perceived areas and populations across states. For example, states such as California, New York, and Texas are relatively well estimated in terms of both area and population. They also rank in the top four most familiar states as rated by our participants, which could make them good candidates as reference entities to use in perspectives. On the other hand, states such as North Dakota, Idaho, and Wyoming are both far underestimated in terms of area (estimates were about 30% of the true value) and at the same time far overestimated in terms of population (about 130% of the true value). One potential explanation for these results is that these states have population densities that are well below that of most U.S. states. Assuming participants base population estimates off of state size and typical population densities, this will lead to overestimates of the number of people who live in these sparse states. This, in addition to their scoring low on familiarity—fewer than 1 in 20 participants rate them as highly familiar—suggests that such states would make poor choices for reference entities in perspectives that communicate areas or populations.

Next we look at people's ability to estimate the areas and populations of hypothetical countries as ratios of U.S. states. As

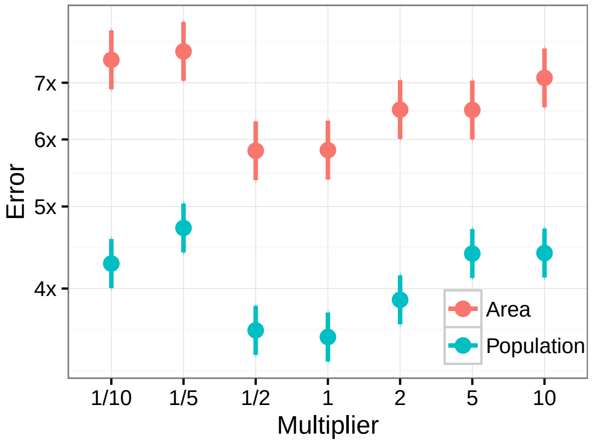


Figure 3. The effect of different multipliers on estimation error. Multipliers of 1 or 1/2 result in minimum error, whereas other multipliers result in less accurate estimates.

mentioned above, we measure error in log-space, computing the average order-of-magnitude deviation between each participant's response and the true value. We examine the unsigned result, for instance treating estimates that are 10 times too high or 10 times too low as both having an error of "10x". In general, we find that hypothetical area estimation is a harder task (with an average error of about 6.6x) compared to population estimation (having an average error of 4.1x). To put this in perspective, an average error of 4.1x in estimating Maryland's population of 6 million people would place responses between 1.5 and 25 million people on average.

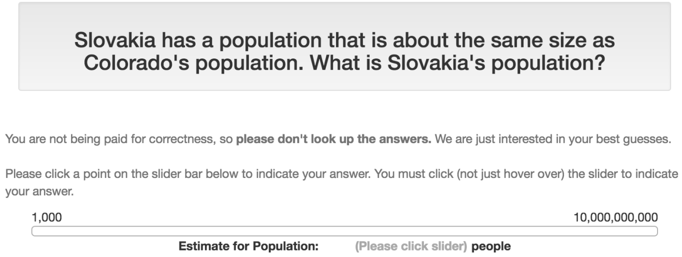
Figure 3 breaks down these errors by the multiplier used in the hypothetical perspective. In this figure, the x axis corresponds to all perspectives shown with a particular multiplier and the y axis corresponds to the average error for those perspectives. Area and population are denoted by color (red and blue, respectively). A multiplier of one produces the most accurate estimates in both dimensions, although, perhaps surprisingly, other multipliers like $\frac{1}{2}$ do not show a significant difference in the accuracy of resulting estimates. In general, the smaller multipliers of $\frac{1}{10}$ and $\frac{1}{5}$ result in higher error than 1 or $\frac{1}{2}$, with this observation holding across both dimensions of area and population.

These independent analyses of state and multiplier level effects reveal important insights about how people perceive different reference quantities and how well they are able to work with various multipliers. At the same time, they do not offer a clear prescription for how to automatically generate perspectives for any given country. To this end, we use the results of this experiment to build a simple model that jointly considers the impact of different states and multipliers to automatically construct perspectives for any given country area or population. Specifically, we fit a linear model to predict the estimate that a given perspective will elicit from participants given the multiplier and state used in the perspective:

$$\log(\text{estimate}) \sim \beta_{state} + \beta_{multiplier} \qquad (1)$$

where β_{state} and $\beta_{multiplier}$ capture state- and multiplier-specific fixed effects, respectively. Given the area or population of a country of interest, we use the fitted model to rank all 350 possible perspectives (50 states × 7 multipliers) and select the one that is most likely to produce approximately accurate responses.

We test this model in the next experiment, where we evaluate it alongside several other policies for automatically generating perspectives.



Slovakia has a population that is about the same size as Colorado's population. What is Slovakia's population?

You are not being paid for correctness, so **please don't look up the answers.** We are just interested in your best guesses.

Please click a point on the slider bar below to indicate your answer. You must click (not just hover over) the slider to indicate your answer.

1,000 10,000,000,000

Estimate for Population: (Please click slider) **people**

Figure 4. A screenshot from our second experiment, designed to compare several different policies for generating perspectives for country areas and populations.

EXPERIMENT 2: EVALUATING POLICIES TO CREATE PERSPECTIVES

We designed this experiment to evaluate and compare the effectiveness of different policies for automatically generating perspectives, including the model developed above. Past work has mainly assessed what people prefer to see, as opposed to how these perspectives impact comprehension. We focus on the latter, and compare four different perspective-generation policies with varying levels of personalization and data requirements. In one extreme we look at a highly personalized policy that phrases any country statistic in terms of the U.S. state that a given participant is most familiar with. In another we simply select the perspective with the lowest objective error, without considering any perceptual biases they may induce. We compare the quality of estimates under these policies to each other as well as to a control condition where no perspectives are shown. Each of these policies is described in more detail below and example perspectives under each policy for the population of Angola are shown in Table 1, where we have assumed the participant is from the state of Minnesota.

- **Best Modeled:** the perspective with the lowest error predicted by our model, which accounts for perceptual effects of the seven multipliers and 50 states revealed by our first experiment.
- **Home State:** participants entered the U.S. state with which they were most familiar. Perspectives were phrased in terms of this state, with up to one decimal place in the multiplier. This ensured that perspectives from this method respected the same constraint as other policies, falling within 10% error of the area or population of the country participants were asked to estimate.
- **Minimum Objective Error:** the perspective with the lowest objective error (in contrast to the lowest modeled error of the *best modeled* policy) using one of the seven chosen

Policy	*Perspective Text*
None	–
Best Modeled	Angola has a population that is about the same size as New York's population.
Home State	Angola has a population that is about 3.9 times the size of Minnesota's population.
Minimum Objective Error	Angola has a population that is about ten times the size of New Mexico's population.
Robustness	Angola has a population that is about five times the size of Oregon's population.

Table 1. Examples of perspective sentences for the four perspective-generation policies (and the control condition).

multipliers and 50 states. Despite being factually accurate, we expected this policy to perform worse than *best modeled*, as it fails to account for perceptual effects.

- **Robustness:** to test the robustness of our model, we chose a perspective with similar objective error to the *best modeled* perspective, but worse modeled error. More specifically, of the perspectives within 5% objective error of the *best modeled* policy, we selected the perspective with the *worst* modeled error. These perspectives were also constrained to use the seven multipliers and 50 states mentioned above. We expected this to show the worst performance of any of the non-control policies, as it intentionally chooses unfamiliar states or difficult multipliers while still maintaining factual accuracy.

- **None:** no perspective was shown, as a control.

Participants. Again using Mechanical Turk, 1,017 participants began the study, with 977 fully completing it. Participants were paid $1 for participation. The survey was open for two days.

Design. Every participant was randomly assigned to a group asked about area (n=508) or population (n=469) of 20 countries. These countries were chosen to have a diversity of populations, physical sizes, and geographic regions, and were presented to participants in a random sequence to mitigate systematic ordering or fatigue effects. Participants had a 20% chance of being assigned to a control group that would see no perspectives (n=93 for area, n=77 for population). Treatment group participants saw perspectives alongside the questions, with perspectives chosen randomly from the four non-control policies. To test transfer learning effects, no participant was shown perspectives on their last two questions.

Procedure. As in Experiment 1, at the start of the experiment participants reported demographic information including age, gender, zip code, and preferred measurement system (imperial or metric). Additionally, participants were asked to specify the U.S. state with which they were most familiar, such as a state where they live or grew up, referred to as their "home state".

Participants were then presented 20 questions as mentioned above, asking them to estimate either the size or area of countries. One fifth of our participants were randomly assigned to a control group where no perspective statements were shown alongside any questions. The remaining participants were shown perspective statements alongside each question, which were randomly selected from the four remaining perspective-generation policies. Figure 4 contains a screenshot displaying a question alongside a perspective statement.

No participant saw perspective statements in the last two questions, in order to test if participants in the treatment group would show improved performance on questions without perspectives compared to the control. That is, would the process of thinking about country statistics in terms of perspectives aid participants for questions when there was no guiding perspective?

Results

The high-level result of this experiment is consistent with past work [1], showing that participants were much more accurate when they saw a perspective than when they did not. Specifically, comparing responses in the control condition to those where participants were shown perspectives, we find both a statistically significant and practically meaningful reduction in error (two-sided t-test, $t(105) = 3.9, p < 0.001$ for area and $t(86) = 3.6, p < 0.001$ for population). This difference is highlighted in Figure 5, which shows results for each perspective-generation policy we tested. The x axis corresponds to the policy used and the y axis measures the average within-participant error for perspectives generated by that policy, with error bars showing one standard error above and below the mean. Looking at responses on area estimates, for instance, we see that participants in the control condition were, on average, off by a factor of 15 from the true areas of the countries they were presented with, whereas average error is cut in half (to 7.5x) when participants were provided with perspectives from our model. We see similar results for population estimates, where participants go from an error of 6.5x in the control condition to 3.7x when aided by our model.

Comparing our model to the other perspective generation policies shows a number of interesting effects. First, we see that the *robustness* condition has significantly higher error than the *best modeled* perspective policy (two-sided t-test, $t(804) = 3.4, p < 0.001$ for area and $t(755) = 4.1, p < 0.001$ for population). This confirms the idea that while some perspectives are objectively equivalent to each other, they are not all equally effective in aiding comprehension. Second, and somewhat surprisingly, we see that the *home state* policy has higher error than the *best modeled* policy, despite the former being personalized to the participant (two-sided t-test, $t(809) = 2.7, p = 0.007$ for area and $t(721) = 2.7, p < 0.006$ for population). We hypothesize that this is due to the difficulty of working with the less familiar multipliers that accompany personalized perspectives (e.g., "about the same size as New York's population" may be easier to comprehend and recall

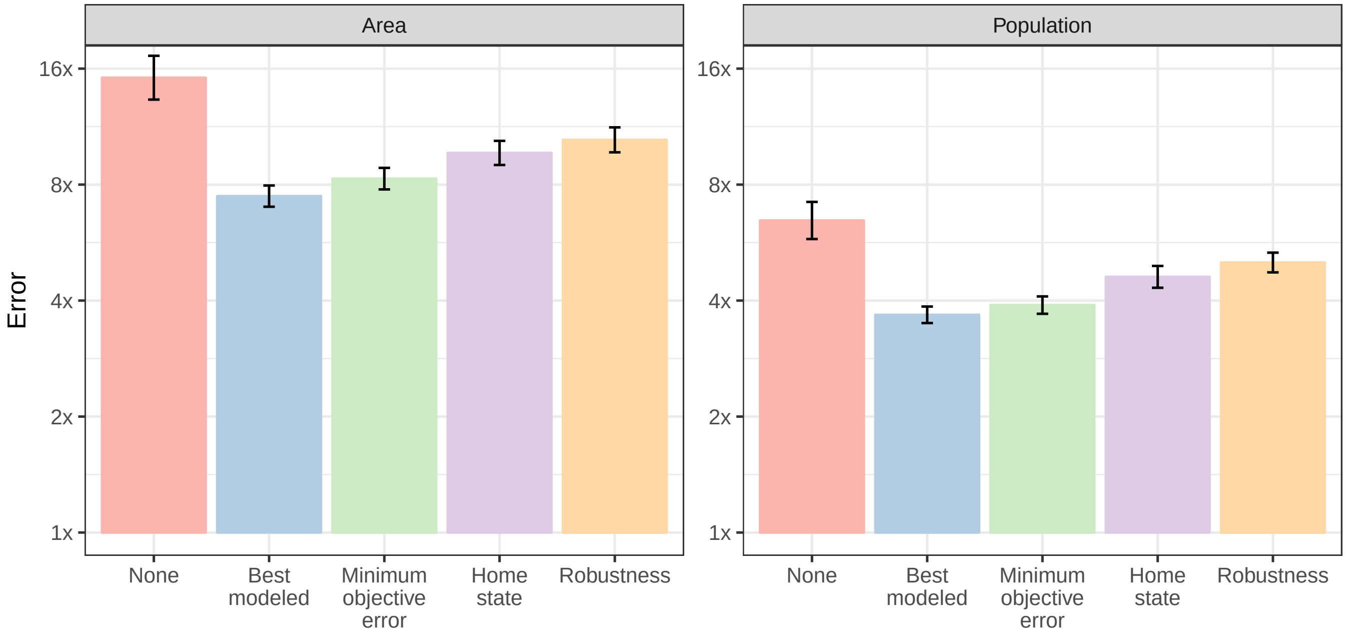


Figure 5. A comparison of participant accuracy based on perspective policy. Showing perspectives from any of the four tested policies decreases participants' error significantly over showing no perspective.

than "about 3.9 times the size of Minnesota's population", even for a participant who lives in Minnesota). While both of these effects are statistically significant, overall we find that the exact choice of perspective policy does not have a large impact on comprehension compared to the alternative of estimation unaided by perspectives.

Figure 6 provides more insight into the impact of perspectives, showing how estimates of country areas (left) and populations (right) shifted between the control and *best modeled* conditions. The x axis corresponds to the true statistic for each country (square miles or population) and the y axis shows the average response. The dashed line on the diagonal shows where perfectly calibrated results would lie if the average estimate for each country statistic matched its true value; points above this line indicate overestimates, while below correspond to underestimates. Estimates in the control condition are shown in red, whereas blue points show the average response for participants who were shown the best modeled perspective, with lines of best fit displayed for the two respective policies. An arrow is shown for each country to highlight the change in estimates between the two conditions.

For almost all perspectives in the area condition, we see higher quality estimates when participants are shown perspectives, as the blue dots are substantially closer than red to the dotted line. As indicated by the direction of the arrows, this is mainly due to the fact that perspectives somewhat correct for the systematic underestimation of area. For population, we see more of a mix, but with overall accuracy greatly increasing. The red lines in both plots have a slope substantially lower than one, indicating participants' tendency to overestimate small values and underestimate large values. In both plots, the blue line has a slope much closer to one, demonstrating that perspectives not only decrease the error of estimates but also improve calibration, reducing systematic over- or underestimation as a function of size. Some overall underestimation remains, possibly due to participants underestimating most reference quantities and our generation policies being constrained to show truthful perspectives.

Finally, to study if exposure to perspectives in the past changed the way people performed on new estimation tasks, we looked at the last two questions in our study in which no participants were shown perspectives. On average for these two questions, participants assigned to make area estimates had an error of 12.6x in the control group compared to 10.6x in the treatment group. Participants assigned to make population estimates had an average error of 5.4x in the control group and 4.7x in the treatment group. Neither of these differences was statistically significant, however, leaving this question open for future work.

EXPERIMENT 3: LONG-TERM EFFECTIVENESS OF PERSPECTIVES

The results of our first two experiments provide strong evidence that automated perspectives improve comprehension, specifically by helping people estimate unfamiliar quantities. In this experiment we ask whether this effect lasts only as long as a participant reads a perspective sentence, or if exposure to perspectives provide benefits over longer periods of time. To assess the long-term impact of perspectives, we conducted a third experiment several weeks after the conclusion of the previous one where we called back the exact same participants and asked them the exact same questions, but this time no one saw any perspectives.

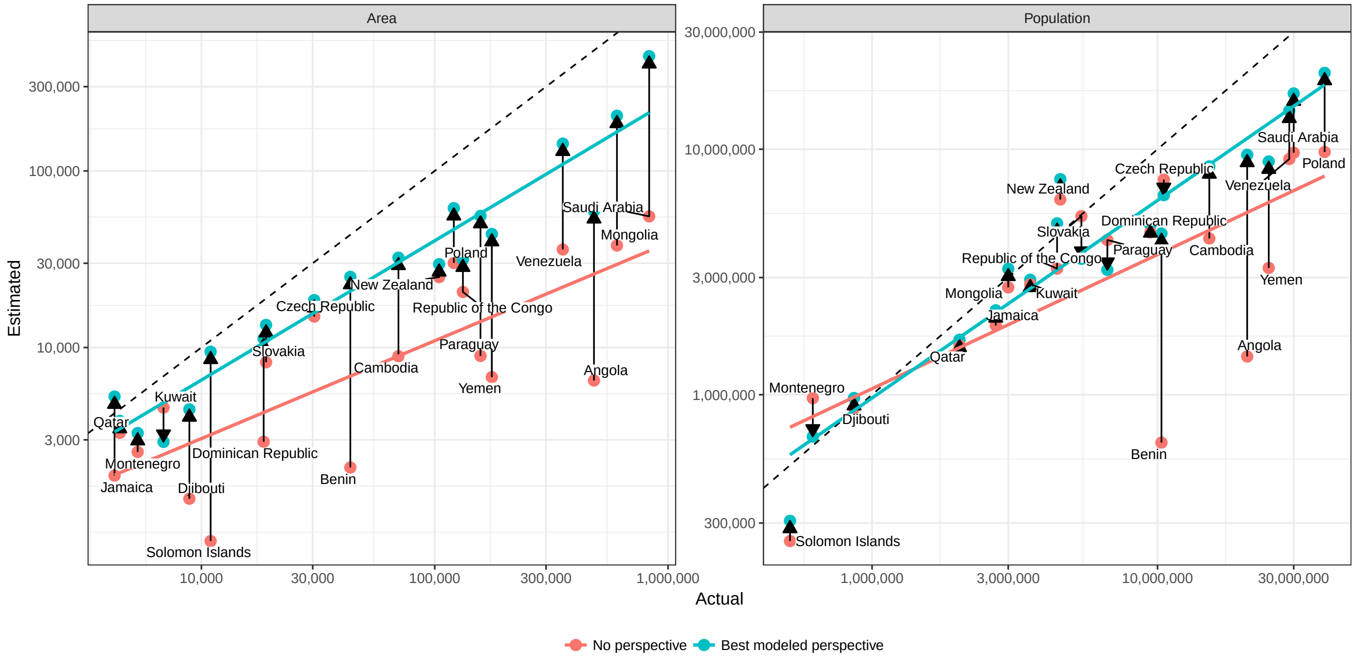


Figure 6. The effect of perspectives on the area (left) and population (right) for each country in experiment 3. Average estimates without perspectives are shown in red, with an arrow pointing to average estimates with perspectives in blue. A line of best fit is shown for each policy, demonstrating that perspectives not only improve the accuracy of estimates but also calibration across a wide range of values.

Participants. Out of the 977 participants who completed the second experiment, 637 participants returned for this experiment, a return rate of 65%. Participants were paid $1 for completing the study, which was available for two days.

Design. Each participant was asked the same questions for the same countries as in the prior experiment, with 324 participants asked questions about area and 313 about population. Questions were shown in the same order as in experiment 2. No perspective statements were shown to any participants regardless of the condition they were assigned to in the previous experiment.

Procedure. Six weeks after our second experiment, we used Mechanical Turk's API to contact prior experimental participants who had successfully completed our second experiment. Each test participant repeated the exact same task they completed in experiment 2, but, as mentioned above, no perspectives were shown. For example, a participant who in experiment 2 received as their third question "Poland has a population about the same as California's population. What is the population of Poland?" would in experiment 3 see as their third question "What is the population of Poland?"

Results

The results of this experiment are shown in Figure 7 (under "Followup Study") alongside results from experiment 2 ("Initial Study") for both area (left panel) and population (right panel). Each bar shows the average within-participant estimation error for a different experiment and condition. Results for those who were in the control condition in experiment 2—and therefore never saw any perspectives—are colored red, whereas results for participants who were exposed to perspectives in experiment 2 are shown in blue. Error bars indicate one standard error above and below the estimated mean.

As shown in the rightmost column of the figure, exposure to population perspectives in the initial study caused a substantial long-term reduction in estimation error in the followup experiment. The error difference between the two groups in the followup study is 6.6x for those who were exposed to population perspectives six weeks earlier compared to 4.9x for those who never benefited from perspectives, a statistically significant difference (two-sided t-test, $t(68) = 2.2, p = 0.03$). Given that the treatment group from experiment 2 did not have the benefit of perspectives in experiment 3, it is not surprising that their error increased in the followup study. Nonetheless, participants in the treatment group from experiment 2 continued to have much better performance than their peers in the control group, even six weeks after seeing perspectives. In contrast to population and despite significant results in a previous set of recall experiments, we failed to observe significant differences in area estimates between the two groups at the six week mark, as indicated by the rightmost column of the area panel in Figure 7.

The mechanism by which initially seeing perspectives improves later performance is an open question. One possibility is that participants simply remember the perspective. If, for example, someone can remember that Poland has the same population as California, they can use this evidence when making estimates in the future. A second possibility is that participants do not remember the exact perspective, but instead learn the strategy of comparing an unfamiliar object to a more

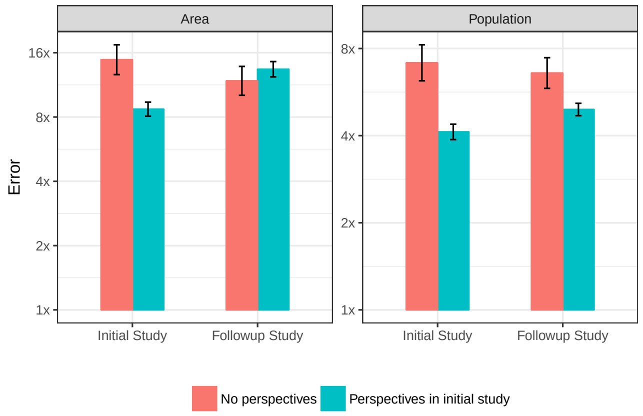


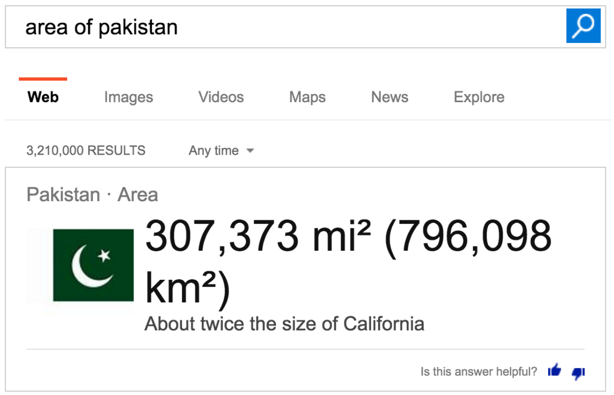
Figure 7. A comparison of error on an initial experiment and a followup experiment conducted 6 weeks later. Participants who saw population perspectives in the earlier experiment had statistically significantly better estimates than those who did not, despite neither group seeing perspectives in the later time period.

familiar one in order to make an estimate. If someone is asked to estimate the size of Poland, even if they don't know that the country is the same size as California, perhaps they will think to compare it to another state with which they are familiar. We leave the study of this mechanism as an open question for future research.

DISCUSSION

In this work we have shown that perspective statements help people estimate unknown quantities and that the beneficial effects of perspectives can remain significant for at least six weeks after the time of exposure. We tested several automated policies for generating perspectives and found that each provides substantial benefits over a control condition without perspectives, demonstrating that it is both possible and relatively easy to improve reader comprehension at scale. Interestingly, we found that a simple, global model for generating perspectives is competitive with a personalized policy. This is not to say that global approaches are always superior to personalized ones, but rather that one can construct effective domain-specific perspectives for a wide audience without elaborate optimization.

Encouraged by these results, we have since deployed perspectives in "instant answer" numbers returned by the Bing search engine. The model developed in this paper was used to create a library of perspectives for country areas in terms of U.S. states, which was first evaluated by a third-party panel of human judges in a side-by-side comparison before being incorporated into the search engine. Now, when the search engine receives a query from a user in the U.S. about the geographic area of a country, it displays a small piece of text comparing the country to a U.S. state. For example, as shown in Figure 8, when responding to a query for the "area of Pakistan", the search engine puts the answer of "307,373 square miles" into perspective as "about twice the size of California". With this first scenario successfully launched, we are actively working to add perspectives to other instant answers provided by the search engine.



area of pakistan

3,210,000 RESULTS

Pakistan · Area

307,373 mi^2 (796,098 km^2)

About twice the size of California

Is this answer helpful?

Figure 8. An example perspectives generated by our model, rendered live on the Bing search engine, which phrases the area of Pakistan as twice the size of California.

Though the work presented here focuses on the prominent but relatively narrow domain of country-level statistics, the insights it reveals apply much more broadly. Having established the importance of simple multipliers and familiar reference objects in generating effective analogies, the main challenge going forward is generalizing these ideas to arbitrary domains so that perspectives can be deployed and tested at scale. One approach to solving this problem is to replace domain-specific human feedback on the effectiveness of different perspectives with machine-learned models that can be automated across a variety of domains. For instance, in ongoing work similar to that of Hullman et al. [10], we are creating a database of reference objects that covers a wide range of measurements and contains proxy features for gauging the familiarity and analogical suitability of these reference objects. These features include how often reference objects are mentioned in different text corpora, queried in search engines, and visited on Wikipedia, all of which can be gathered automatically and easily localized to different subpopulations.

We see these as important steps in utilizing online platforms to improve numerical comprehension among both authors and their audiences. We hope that the perspectives framework will not only aid producers and consumers of information, but also stimulate research in education, journalism, and cognitive psychology.

REFERENCES

1. Pablo J. Barrio, Daniel G. Goldstein, and Jake M. Hofman. 2016. Improving Comprehension of Numbers in the News. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 2729–2739. http://doi.acm.org/10.1145/2858036.2858510

2. S Bautista, R Hervás, P Gervás, Richard R Power, and Sandra Williams. 2011. How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies. In *INTERACT 2011*. Vol. 6946. Springer Berlin Heidelberg, 57–64. http://dx.doi.org/10.1007/978-3-642-23774-4_7

3. Elizabeth M Brannon. 2006. The representation of numerical magnitude. *Current opinion in neurobiology* 16, 2 (2006), 222–229. `http://doi.org/10.1016/j.conb.2006.03.002`

4. Norman R Brown and Robert S Siegler. 1993. Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological review* 100, 3 (1993), 511. `http://dx.doi.org/10.1037/0033-295X.100.3.511`

5. Arun Chaganty and Percy Liang. 2016. How Much is 131 Million Dollars? Putting Numbers in Perspective with Compositional Descriptions. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 578–587. `http://www.aclweb.org/anthology/P16-1055`

6. Fanny Chevalier, Romain Vuillemot, and Guia Gali. 2013. Using concrete scales: A practical framework for effective visual depiction of complex measures. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2426–2435. `http://doi.org/10.1109/TVCG.2013.210`

7. Sunaina Dowray, Jonas J. Swartz, Danielle Braxton, and Anthony J. Viera. 2013. Potential effect of physical activity based menu labels on the calorie content of selected fast food meals. *Appetite* 62, 0 (2013), 173–181. `http://doi.org/10.1016/j.appet.2012.11.013`

8. Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in cognitive sciences* 8, 7 (2004), 307–314. `http://doi.org/10.1016/j.tics.2004.05.002`

9. Gerd Gigerenzer. 2014. *Risk savvy: how to make good decisions*. Viking Books, New York, NY, USA.

10. Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. 2018. Improving Comprehension of Measurements with Concrete Re-expression Strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systemsa (CHI '18)*. ACM.

11. Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating Personalized Spatial Analogies for Distances and Areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 38–48. `http://doi.acm.org/10.1145/2858036.2858440`

12. Richard P Larrick and Jack B Soll. 2008. The MPG Illusion. *Science* 320, 5883 (2008), 1593–1594. `http://doi.org/10.1126/science.1154983`

13. W Mason and S Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* (Jan 2012). `http://doi.org/10.3758/s13428-011-0124-6`

14. Dennis L McNeill and William L Wilkie. 1979. Public policy and consumer information: Impact of the new energy labels. *Journal of Consumer Research* (1979), 1–11. `http://doi.org/10.1086/208743`

15. L Neuhauser, K Paul, B Fischhoff, NT Brewer, and J Downs. 2011. Communicating Risks and Benefits: An Evidence-Based User's Guide. (2011).

16. John Allen Paulos. 1988. *Innumeracy: Mathematical illiteracy and its consequences*.

17. Bruce M Ross and Trygg Engen. 1959. Effects of round number preferences in a guessing task. *Journal of experimental psychology* 58, 6 (1959), 462. `http://dx.doi.org/10.1037/h0049112`