Improving Comprehension of Numbers in the News

Pablo J. Barrio
Columbia University
New York, NY
pjbarrio@cs.columbia.edu

Daniel G. Goldstein
Microsoft Research
New York, NY
dgg@microsoft.com

Jake M. Hofman
Microsoft Research
New York, NY
jmh@microsoft.com

ABSTRACT
How many guns are there in the United States? What is the incidence of breast cancer? Is a billion dollar budget cut large or small? Advocates of scientific and civic literacy are concerned with improving how people estimate and comprehend risks, measurements, and frequencies, but relatively little progress has been made in this direction. In this article we describe and test a framework to help people comprehend numerical measurements in everyday settings through simple sentences, termed *perspectives*, that employ ratios, ranks, and unit changes to make them easier to understand. We use a crowdsourced system to generate perspectives for a wide range of numbers taken from online news articles. We then test the effectiveness of these perspectives in three randomized, online experiments involving over 3,200 participants. We find that perspective clauses substantially improve people's ability to recall measurements they have read, estimate ones they have not, and detect errors in manipulated measurements. We see this as the first of many steps in leveraging digital platforms to improve numeracy among online readers.

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
Consider a billion dollar cut to the federal budget or a million liter decrease in global carbon dioxide emissions. Are these large or small numbers? Unfamiliar measurements make up much of what we read, but unfortunately carry little or no meaning to typical readers, as they can be difficult to interpret without the appropriate context. As others have found [4, 14, 11], and we shall show, people have difficulty remembering, estimating, and detecting errors in measurements sampled from everyday reading material.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI'16, May 07 - 12, 2016, San Jose, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858510

Improving numerical literacy among the general population has been a long-standing challenge, with popular books [21] and programs [3] devoted to the cause. The problem is so pervasive that the public editor of the New York Times recently issued a statement calling for Times writers to "put large numbers in context."[1] In this paper, we propose and test a method for improving numerical communication. In particular, we introduce simple sentences, termed *perspectives*, that employ percentages, ratios, rankings or other comparisons to provide context around numerical measurements in online content. We show that the perspective framework is flexible enough to provide context for a wide range of numerical measurements, but simple enough to be understood and used by everyday readers. We develop a crowdsourced system to generate perspectives and conduct randomized experiments to demonstrate their impact on numerical comprehension. Somewhat surprisingly, we find that through the use of perspectives, the very same users who often have difficulty understanding measurements can in fact help clarify these numbers for other readers.

To illustrate our approach, consider the dozen quotes taken from the New York Times shown in Table 1. Each sentence contains a numerical measurement (in bold) and is followed by a perspective (generated by crowd workers, in italics), designed to make the measurement easier to understand. One of these quotes, for instance, mentions the number of registered firearms in the United States, which is about 300 million. It can be challenging to estimate this statistic if one has never seen it before, and difficult to recall even if one has seen it in the past. Likewise, it can be challenging to detect whether a printed number is correct or contains an error (e.g., if 30 million were written instead of 300 million). Our experiments show that each of these tasks (recall, estimation, and error detection) is substantially easier with the help of a perspective sentence that rephrases the measurement as "about equal to 1 firearm for every person in the United States." To preview one of our results, while only 40% of people shown only the original quote were able to recall this number exactly, nearly 55% of participants who where randomly selected to see it phrased as firearms per person were able to do so. Although the exact effect size varies depending on the quote, measurement, and perspective, we find similar support for the benefits of perspectives across all of our experiments.

In the remainder of the paper we discuss how the quotes in Table 1 were generated and test the impact they have on numerical comprehension. First, we briefly describe the perspective

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

Quote and top-rated perspective
"The prosthetic seems **15** centimeters longer than the other leg," Sebastian Bayer, who finished fifth, said. *To put this into perspective, 15 centimeters is about equal to the length of half a foot.*
He reached double digits in sacks in six seasons, including 2001, when he set the single-season record of **22.5**. *To put this into perspective, 22.5 sacks in a season is about 1.4 sacks for every game played.*
The F.A.A. permits amateurs to fly unmanned aviation systems – the agency's term for drones and other devices governed by remote control – if the aircraft stay under **400** feet and well away from any airports. *To put this into perspective, 400 feet of elevation is about equal to the height of a 40 story skyscraper.*
Mr. Obama also repeated his concerns about Russian compliance with the 1987 treaty banning American and Russian missiles with a range of 300 to **3,400** miles. *To put this into perspective, 3,400 miles is about 1.27 times larger than the width of the continental United States of America.*
The Ohio National Guard brought **33,000** gallons of drinking water to the region, while volunteers handed out bottled water at distribution centers set up at local high schools. *To put this into perspective, 33,000 gallons of water is about equal to the amount of water it takes to fill 2 average swimming pools.*
Early on Saturday, municipal officials asked the **500,000** residents served by the city's water system to stop using tap water after the toxins were found at a city water treatment plant. *To put this into perspective, 500,000 residents is about 77% of the population of the whole Toledo metropolitan area.*
The storm killed thousands of people in Honduras, left **one million** homeless and destroyed what was left of a declining Banana industry, once the country's lifeblood, as well as other vital crops. *To put this into perspective, one million people is about 12% of the population of Honduras.*
The group says it has helped to preserve more than **120 million** acres around the world. *To put this into perspective, 120 million acres of protected land is about 1.15 times larger than the state of California.*
They also recommended safety programs for the nation's gun owners; Americans own almost **300 million** firearms. *To put this into perspective, 300 million firearms is about 1 firearm for every person in the United States.*
With its trove of knowledge about the likes, histories and social connections of its **1.3 billion** users worldwide, Facebook executives argue, it can help advertisers reach exactly the right audience and measure the impact of their ads – while also, like TV, conveying a broad brand message. *To put this into perspective, 1.3 billion users is about 4 times larger than the entire population of the United States.*
Facebook, which made $1.5 billion in profit on $**7.9 billion** in revenue last year, sees particular value in promoting its TV-like qualities, given that advertisers spend $200 billion a year on that medium. *To put this into perspective, 7.9 billion dollars annual revenue is about 25 dollars for every person in the U.S.*
Bob Baur, chief global economist for Principal Global Investors, made up a name for what he thinks has been holding back both consumers and businesses from the more exuberant kind of spending that would help close the gap of as much as $**1 trillion** between the economy's current level and its larger capacity for generating goods and services without setting off a significant rise in inflation. *To put this into perspective, 1 trillion dollars is about 3144 dollars for every American citizen.*

Table 1. Text and top-rated perspectives of selected quotes. The measurements of interest are shown in bold and the perspectives rephrasing them are shown in italics.

framework and the scalable, crowdsourced platform we created to collect perspectives from everyday workers. In the system, crowd workers are shown actual measurements taken from the news and asked to complete perspective templates that make the underlying measurements easier to understand. Based on worker voting, the best perspectives are selected to appear within actual news articles as they are read.

We then test the effectiveness of perspectives through a series of randomized, online experiments, which show that augmenting news articles with these sentences improves people's ability to understand the magnitude of numerical measurements. Our first experiment investigates perhaps the most basic aspect of comprehension, the ability to remember or at least to approximate numerical quantities one has read. To test this, we present people with quotes from the news and, after a forgetting period, ask them to recall the measurements contained in the quote. Our second experiment focuses on another level of comprehension, as reflected in the ability to make reasonable estimates of unknown quantities. In this experiment, we show participants quotes from news articles that are missing key measurements and ask them to make interval and point estimates as to what the missing values might be. In our third and final study we test yet another aspect of comprehension, where we present participants with quotes from news articles containing potentially erroneous measurements and ask them to identify possible errors. Participants in each experiment are randomly selected to either see the original article, or to see it augmented with a simple perspective sentence. Across these experiments we find that providing participants with perspective sentences improves their ability to recall measurements they have read, to estimate measurements they have not, and detect errors in manipulated measurements.

We begin by briefly discussing related work.

RELATED WORK

Despite much past work on the topics of numerical literacy and estimation [10, 17, 4, 14, 11] as well as a number of classroom-based studies on improving numeracy among students [15, 19, 2] and journalists [22], there are few existing tools to help the common reader better understand unfamiliar measurements. Popular sites such as Medium[2] and NewsGenius[3] allow readers to annotate articles with arbitrary information, and a recent tool by Liaw and colleagues [13] helps readers assess the trustworthiness of information, but none of these tools focus on quantitative information. Resources such as WolframAlpha[4] and Dictionary of Numbers[5] do focus on numbers, but do not consider the context in which these measurements are mentioned. Furthermore, we find no studies in the literature on their impact on comprehension.

Related research has been done in simplifying the representation of numbers themselves (e.g., writing "one half" instead of "50%") to improve reader understanding [1, 23], but not on actually re-expressing the numbers in other terms. To date, the largest advances in numerical communication lie within the policy domain. For instance, researchers have found that people make better decisions about automotive fuel consumption when information is re-expressed as "gallons per 100

[2] `http://medium.com`
[3] `http://news.genius.com`
[4] `http://wolframalpha.com`
[5] `http://dictionaryofnumbers.com`

miles" instead of as "miles per gallon" [12]. Likewise, creative ways to re-express the caloric content of foods (e.g., as the amount of exercise needed to burn them off) [7] and the energy consumption of appliances [18] have been proposed to help people understand their consumption. And decades of research in risk communication have uncovered ways to help people appreciate the medical, financial, and environmental risks around them [9].

In this paper, we aim to build upon these promising findings by broadening the scope to that of arbitrary measurements (not just measures of risk and consumption), and by providing a general-purpose method for conveying unfamiliar measurements to everyday readers. We measure the effectiveness of this method through three experiments, presented below, where we discuss cognitive mechanisms that explain why perspectives might aid numerical comprehension.

GENERATING PERSPECTIVES

We developed a simple yet flexible framework to provide context around arbitrary measurements mentioned in online content. To do so, we designed a set of *perspective templates*, pictured in Figure 1, that allow a measurement to be re-expressed in variety of formats (e.g., "x times larger than y," "about equal to y," "the x-th largest y," or "in the top x% of all y,").

The templates were developed through an iterative process over the course of several months. We started with a seed set of templates that captured different contexts such as relative percentages and multiples. Each day we examined front page articles from the New York Times for numerical measurements and used the current set of templates to re-express these measurements in more familiar terms. We iteratively refined existing templates and added new templates until they were rich enough to capture all use cases we encountered, but simple enough to be understood by everyday readers.

Each of the 10 final templates decomposes a perspective into three factors: a scaling factor, an attribute, and a reference entity. For example, the first template in Figure 1 recasts the one million left homeless by a storm in Honduras as a percentage of a reference amount—e.g., as 12% of the population of Honduras, where 12% serves as the scaling factor, "population" is the attribute, and "Honduras" is the reference entity. Although our work does not rely on these exact templates being used in all contexts, the templates standardize the representation of contextual information and eliminate effects of chance wording in our experiments. Furthermore, templates have the advantage of generating structured data for future automatic generation of perspectives.

We used these templates to collect perspectives from workers on Amazon's Mechanical Turk online labor platform [16]. After a short training period that validated their ability to research and manipulate simple statistics, workers were presented with a randomly selected quote taken from an article that appeared on the front page New York Times[6] between March and September of 2014. As shown in Figure 1, up to three adjacent sentences from the article were displayed before and after the quote in a smaller and lighter font to provide context around it. Templates were presented in a randomized order to avoid a position bias favoring higher ranking options. Each worker was allowed to add an unlimited number of perspectives for each quote in the system, and was required to document each perspective by providing a URL for fact-checking any source information used. Finally, and to motivate users to submit high-quality perspectives, workers were told they would be paid anywhere from $0.05 to $0.50 per perspective according to the perceived helpfulness of their contributions.

[6] `http://www.nytimes.com/pages/todayspaper/`

In total we collected 370 perspectives on 64 quotes from 80 different Mechanical Turk workers, for an average of 4.6 perspectives per worker and 5.8 perspectives per quote. The overwhelming majority (76%) of the perspectives submitted by workers used a percentage or multiplier to provide context (i.e., "x% of y," "about equal to y," or "x times larger/smaller than y"). We left a "write your own" template option to check whether participants could not find a satisfactory template. This option was rarely used, consistent with the refined list of templates being relatively complete for this corpus.

To assess the quality of each contributed perspective, we asked workers to rate the helpfulness of perspectives on a scale from 1 (not helpful at all) to 5 (very helpful). Workers viewed randomly selected quotes along with one perspective collected for its corresponding measurement. Each worker rated 10 perspectives from quotes that they had not seen during the generation phase. This prevented malicious users from rating their own perspectives highly to increase their pay. We collected a total of 12,094 ratings from 1,862 unique workers, comprised of at least 25 ratings for each of the 370 perspectives.

Next, we evaluated the effectiveness of a dozen of the top-rated perspectives in a series of randomized experiments.

EVALUATING PERSPECTIVES

Our objective is to test whether perspectives help people appreciate and comprehend numerical measurements. As discussed above, we assume that comprehension will be reflected in three measures—recall, estimation, and error detection—which we assess in three separate experiments.

In the three controlled experiments, we use as stimuli 12 news quotes and the top rated perspective for each, shown in Table 1. These quotes were intentionally selected to cover a wide range of measurements in terms of both their amount and unit (e.g., ranging from 22.5 sacks in a football season to $1 trillion dollars in economic capacity). The treatment in each experiment is exposure to a perspective: participants were randomly selected to see (or not see) a perspective alongside each quote, and then asked to either recall its measurement, estimate a missing measurement, or detect whether a measurement has been manipulated. All experiments were run on Amazon's Mechanical Turk platform and restricted to workers with an approval rating of 95%.

To assess the quality and accuracy of responses in the experiments that follow, we compute the *relative log error* between the value submitted by each participant and the ac-

Figure 1. User interface for collecting perspectives.

tual measurement to which it is being compared. Relative log error is defined as the percent difference between the log of the actual value and the submitted one: $|\log(actual) - \log(submitted)|/\log(actual)$. This measure has two desirable properties. First, it accounts for the wide variation in submitted values, which span several orders of magnitude. Second, it allows us to assess responses to different questions, containing wildly different measurements, on a common scale. In this sense it can be useful to think of relative log error as absolute log error [20] adjusted to the scale of the actual value. That said, this measure may be unfamiliar to readers, and because we employ it in several analyses in this paper, an example may be in order. At the time of writing, the population of the United States is about 320 million. A relative log error of 10% would correspond to believing the US population to be as low as 45 million or as high as 2.3 billion. At 20% relative log error, these values would be 6 million and 16 billion (i.e., more than double the world's current population), while at 30%, they would be 900,000 and 114 billion. From this we can see that a relative log error of 30% or less captures an enormous range of estimates. Responses outside this range might include extreme misestimates as well as typographic errors or abbreviations, for instance responding with "2.3" when one means "2.3 billion," which we occasionally observe in our experiments. Accordingly, we limit all analyses and plots in this paper to responses within a 30% relative log error range.

Experiment 1: Recall

In this experiment, we test whether perspective sentences help people remember what they have read. Why would perspectives aid memory? For example, why would knowing that one million people is about 12% of the population of Honduras help people remember that one million people were mentioned in the seventh quote in Table 1? Several mechanisms may jointly play a role. The first is mere repetition, which influences the probability of remembering [24]. The second mechanism is elaboration. As readers think about one million being 12% of the population, they spend more time simply processing the number one million in working memory, which makes it more likely to be retrieved later [6]. The third mechanism is that the information in the perspective can be used to reconstruct the forgotten target value. If, as in the previous example, the reader estimates Honduras' population at around 8 million, then if one million is forgotten but 12% and 8 million are retained, the reader can approximate one million by taking 12% of 8 million. Fourth and finally, the additional information in the perspective can serve as a retrieval cue for the target value [25]. We therefore predict that perspectives will aid people in their efforts to remember what they have read. We expect that some of the benefit of perspectives will be due to mere repetition, but also expect gains beyond this because of the multiple mechanisms at play.

At a high level, in this experiment, participants read six news quotes, in plain text, containing numbers. After a forgetting period, they were asked to recall or estimate the measurement of interest from each quote. In all formats, the focal quotes were surrounded by a few sentences of text from the actual news article from which they were taken. Quotes could appear in one of three presentation formats. In the "original" format, quotes were as they appeared in the news. In the "repeated quote" format, the quote containing the measurement was repeated in the margin in the style of a "call out box." In the "perspective" format, the quotes containing the measurements were followed by inline perspective sentences. After reading the quotes, participants played Tetris, followed by a surprise quiz in which they were shown the quote with the measurement missing and asked to fill in the blank and guess what its value might be. These guesses are the dependent variable in this experiment.

On the first page of the experiment, participants were told the experiment would consist of three phases: "first, reading quotes from several news articles; next, playing a brief game of Tetris; and third, answering some questions." For each participant, six quotes were randomly drawn from the set of 12 in Table 1. Each worker was randomly assigned to the repeated quote condition or the perspective condition. In the repeated quote condition, participants saw three quotes in the original format and three in the repeated quote format, in a random order. The perspective condition was identical, except with the three modified quotes in the perspective (as opposed to repeated) format.

Next, to provide forgetting time, participants were presented with a Javascript version of the game Tetris and were instructed to play for 120 seconds. Afterwards, they were redirected to the final phase in which they were told they would be shown the six quotes, one at a time, and asked to fill in the missing value in each quote before a 30 second countdown timer runs out. The countdown timer was used to prevent people from searching for the answers online. In addition, and also to reduce cheating, participants were also told they would be paid whether or not they answered correctly. Participants were told that if they did not wish to input a guess, they could simply let the timer run out. After one practice item—a new quote for which the correct answer was provided—participants made their guesses for the missing values in the six quotes. Javascript enforced that participants submitted valid numbers, which we accepted as numerals, words, or some combination (e.g., "1 million"), and both the raw string and the parsed floating point value were saved.

The experiment took place online and participants were 819 workers from the Amazon Mechanical Turk online labor market, who were paid \$1.50 for participation. As a result of the random assignment, 405 participants saw the repeated quote condition while 414 saw perspectives. After dropping results from those who did not fully complete the experiment, we were left with 379 and 381 in each condition, with completion rates of 94% and 92% (a non-significant difference, $p = .47$, χ^2 test). Therefore we collected 2,280 responses for quotes that were shown in the original presentation (3 per person in both conditions), 1,137 in the repeated quote condition (3 per person), and 1,143 with an inline perspective (3 per person). Participants did not submit a guess (timed out) in 11.0%, 10.2%, and 11.2% of items in the original, repeated

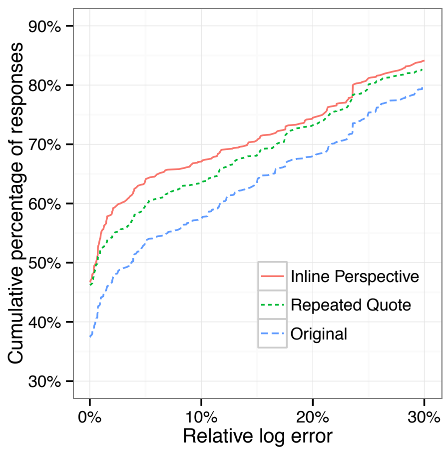

Figure 2. Accuracy of recalled values as measured by relative log error, for original quotes, quotes with repetition, and quotes with inline perspectives.

quote, and perspective conditions, respectively, also a non-significant difference ($p = .72$, χ^2-test).

Figure 2 shows relative log error by condition for all non-timed out responses, averaged across all 12 quotes. For each level of relative log error on the horizontal axis, the vertical axis displays the percentage of responses with at most this amount of error. For example, in the perspective format, approximately 67% of responses have a log-error of 10% or less, while in the original format only 57% do. In terms of relative log error in recall, perspectives provide a clear improvement over the original quotes alone. The repeated quote condition falls between these two, suggesting that part, but not all, of the benefit of perspectives may be due to repetition. Specifically, we assessed the difference in percentage of responses at each 1% relative log error value shown in Figure 2 and found a significant improvement for the perspective condition over the original quote for *every* such value (all p-values < 0.01, χ^2 test).

To assess differences in the accuracy of responses between conditions, we regressed relative log error against condition and quote. We find that the perspective condition provides a significant 3.2 percentage point improvement in relative log error over the original format ($p < 0.001$). To put this in perspective, a relative log error of 3.2 percentage points in estimating the U.S. population corresponds to guessing as low as 171 million or as high as 599 million.

To provide further insight into the accuracy of responses, Figure 3 shows relative log error for each quote individually. This reveals substantial variation in the improvement provided by perspectives. Compare, for instance, the 300 million firearms quote to the 7.9 billion dollars one. Perspectives provide great benefit for the former (stated as 1 firearm per person in the U.S.) but not the latter (when phrased as $25 per person in the U.S.). This could be due to the ease or difficulty of recalling the numbers themselves, or it could be due to the quality of the accompanying perspective in each case. As discussed below, this highlights one avenue for future research on the design and impact of perspectives. That said, perspectives appear to help substantially in the vast majority of quotes.

To conclude, we see improvements from perspectives over the original quotes both for exact recall (a relative log error of zero) and for cases in which the value cannot be recalled exactly (a relative log error greater than zero). With the aid of perspectives people remember roughly half of the numbers they see, compared to a third of numbers without them. Our experiments also demonstrate that the benefits of perspectives exceed that of mere repetition. These results are encouraging, but recall demonstrates only one aspect of comprehension. In the following sections we test two more—estimation and error detection.

Experiment 2: Estimation

The previous experiment demonstrated that perspectives help people retain and make estimates about information they have recently read. While knowledge and recall of important quantities is certainly one aspect of numeracy, there are many others, such as ability to estimate unknown quantities. In this experiment we tested workers' accuracy in estimating the values of quantities they had *not* previously been exposed to, both with and without the aid of perspectives.

Why might perspectives improve estimation? Take our running example of the individuals left homeless by the storm in Honduras. When asked to estimate the number of such people without any further information one might entertain unrealistic values, such as those larger than the country's population. Now imagine that when participants entertain an estimate, they see it put into perspective as a percentage of Honduras' population. This gives participants a choice. They can either estimate the number of people directly, or they can estimate the percentage, which should be an easier task. For example, participants might infer that a percentage less than .001% could not be correct because such a low figure would not have made the pages of the New York Times. At the other extreme, participants may infer that values above 75% could not be correct because if the devastation were so vast, they would have heard of it before. This latter kind of reasoning is called a "lack of knowledge inference" [8]. With perspectives, people can make use of two routes (reasoning about the original units or those in the perspective sentence) to arrive at estimates, similar to how the perceptual system can substitute one kind of information for another in what is known as vicarious functioning [5].

We recruited online workers who were paid $0.80 to provide estimates for six randomly selected quotes. Workers were shown the example quotes with a missing measurement and first asked to provide a plausible range, followed by a best estimate for its value based on this range. Each participant was randomly assigned to see either the original quote (the control condition) or the quote with an inline perspective (the treatment condition) for all six quotes that they saw. In addition to the quote, workers in the treatment condition were also shown a highlighted, inline perspective that rephrased candidate values as they were entered. For example, if the

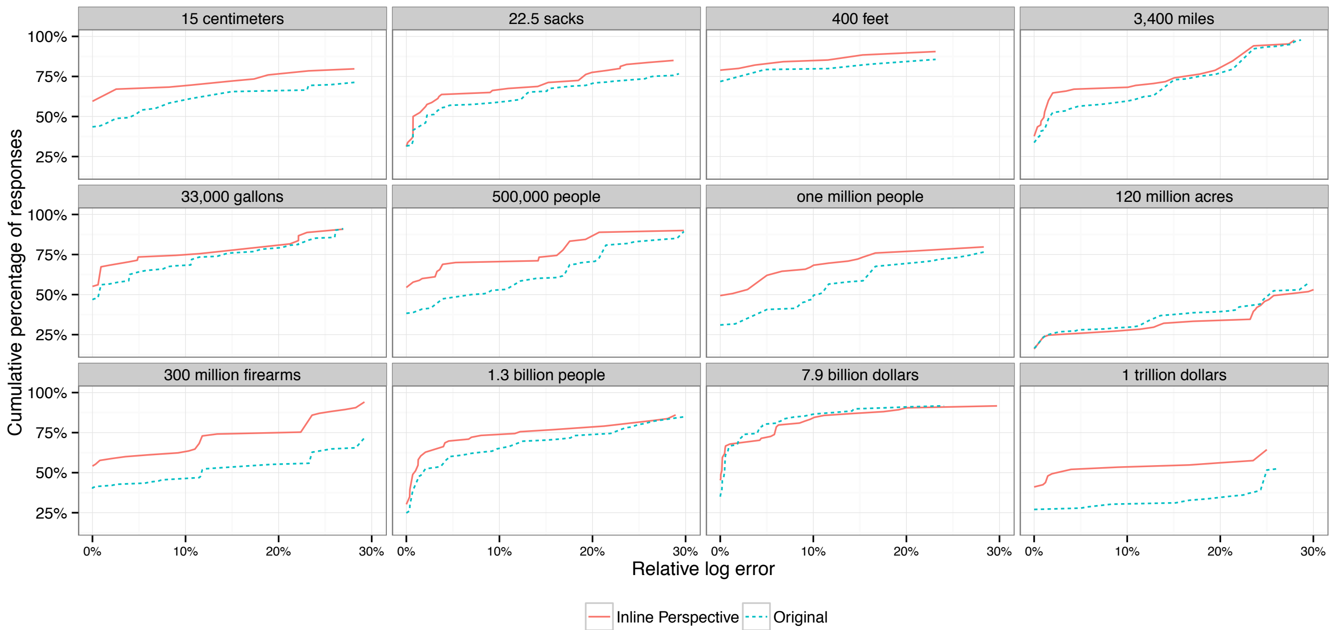

Figure 3. Relative log error for each quote in the recall experiment, by presentation format.

Honduran quote was shown with a candidate value of 8 million people, the perspective expressed this as 97.5% of the population of Honduras.

Participants completed two steps for each quote, first selecting a plausible range and then a best estimate. In the first step they were shown 11 candidate values for the missing measurement and were asked to classify whether each was "too low," "plausible," or "too high" by clicking one of three buttons. We used the results of the previous experiment to select candidate values so that the examined range was large enough to contain the majority of reasonable estimates, but small enough to exclude obviously wrong values. Specifically, we constructed a candidate range for each question that was centered around the true value, equal in size to the range of inner 80% of responses in the recall experiment. We then took candidate values from this range at 5 logarithmically-spaced values above and below the true value for each question. For example, this produced a range from 2,000 to 490 million people for the displaced Hondurans, with the correct answer of one million people in the middle. This corresponded to a range of 0.2% to 6,000% of the Honduran population in the perspective that was shown to participants in the treatment condition.

To guard against anchoring effects, participants were also randomly assigned into one of two conditions where these values were shown in either ascending or descending order. Each click moved the participant to the next value until they had made judgments on all 11 candidates. This determined a "plausible range" for the measurement, defined by the largest value they judged to be "too low" and the smallest value marked as "too high."

The second step presented participants with a slider that allowed them to select a fine-grained estimate for the missing value from this plausible range. To prevent defaults from biasing responses, the slider was initialized without a selected value. The scale on the slider was also randomly assigned at the participant level to be either linear or logarithmically spaced. The missing value updated as the participants hovered their mouse over the slider, clicking to select a final estimate. In addition to the changing measurement, participants in the treatment condition were shown a dynamic perspective that continuously updated as they moved their mouse. Once a best estimate was selected the participant was asked to double check their guess before clicking submit to move to the next quote.

As a result of the random assignment, 1,071 participants were assigned to see the original quote, while 1,024 were assigned to see the perspective. After ineligible participants (who had completed any of our previous experiments) were turned away and after eliminating participants who did not complete the experiment, this left 657 and 511 in each group. This corresponds to completion rates of 87% and 77% for eligible workers in the control and perspective conditions. The difference in completion rates, which is significant ($p < .001$, χ^2 test), is likely due to user interface issues (for example, longer page lengths) as we did not observe any significant differences in either of our other experiments, which made similar use of perspectives. Future experiments will be better instrumented to detect such user interface problems.

Figure 4 shows the percentage of correct responses for each condition in the first stage of the experiment, computed from more than 77,000 clicks. Each value on the horizontal axis corresponds to one of the 11 candidate values shown in the

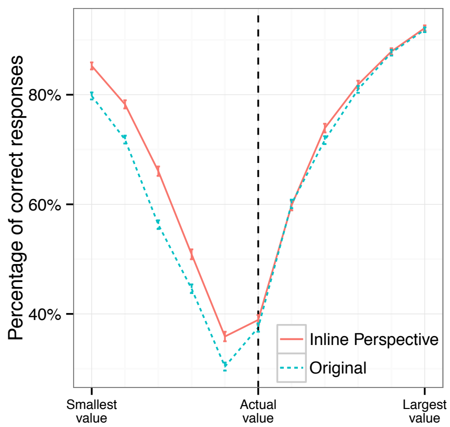

Figure 4. The percentage of responses correctly classified as "too low," "plausible," or "too high" in the estimation experiment for original quotes compared to those with inline perspectives. Error bars show one unit of standard error above and below the mean.

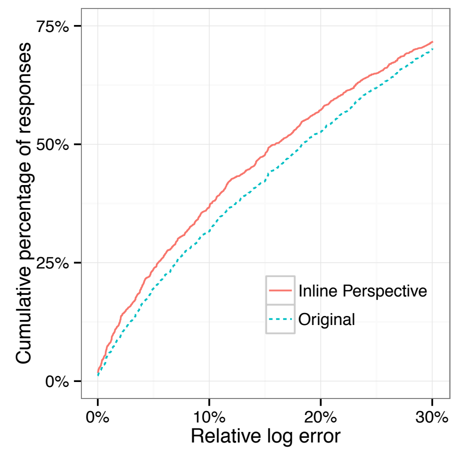

Figure 5. Accuracy of estimated values as measured by relative log error for original quotes and quotes with inline perspectives.

first stage. A correct response corresponds to the user clicking "too low" when the candidate value is below the actual value, "too high" when the candidate value is above it, and "plausible" when the actual value is presented. The u-shaped trend in this figure shows that participants found the extreme candidate values highly implausible—with over 80% of responses correctly rejecting these values—but had substantially more difficulty in correctly identifying the actual value. Furthermore, perspectives aided participants in rejecting incorrect intermediate values, particularly those below the actual value, where we observed improvements of 5 to 9 percentage points over the control condition.

To quantify the improvement that perspectives bring, we fit a logistic regression to predict the percentage of correct responses shown in Figure 4. Specifically, we regressed success rate against an indicator for whether a perspective was shown, an indicator for each candidate level, and the interaction of the two, as well as an indicator for value order. This shows substantial improvements in accuracy from the presence of perspectives and a small but significant benefit to presenting values in ascending (rather than descending) order (all $p < .001$).

Figure 6 shows the results of the second stage of the experiment, in which participants provided their best estimate for the missing value. The red and blue curves show the distribution of these estimates across quotes for the perspective and control groups, respectively, while the dashed line shows the actual value. In many but not all of the quotes, perspectives appear to improve the quality of estimates by reducing the variance of responses (the red curves are more concentrated about their peaks) and shifting them towards the actual value (the peaks are closer to this value).

As in the previous experiment, we assessed the accuracy of these estimates by computing the cumulative percentage of responses at each relative log error value up to 30%, shown in Figure 5. We found a significant improvement for the perspective condition over the original quote for every such error value between 1% and 25% (all p-values $< .001$, χ^2 test). For example, in the perspective format, approximately 39% of responses have a relative log error of 10% or less, while in the original format only 33% do. We see such improvements across many of the individual quotes as well, most strikingly in the 120 million acres quote. Conversely, several quotes show relatively little benefit from perspectives, such as the record 22.5 sacks in a season, where Figure 6 shows that participants have a reasonably accurate estimate even without the aid of perspectives.

To model these effects, we regressed relative log error for participants' best estimates on indicators for the perspective format, scale type (log vs. linear), and each quote. We also included an interaction term between the format and quote to capture differences in the impact of each perspective. We observe a slight benefit to using a linear scale in the slider, corresponding to a 1 percentage point improvement in relative log error ($p < .01$). More importantly, this reveals that, holding all else equal, perspectives reduce relative log error by 7.1 percentage points ($p < .001$), with some variation by quote as noted above. These results would be of marginal importance if most of the benefits of perspectives come from choosing a reasonable plausible range, in which case this regression merely recapitulates the results of the first stage. To test this we repeated this analysis limited to the set of reasonably well-informed participants whose plausible range included the actual value. Among this subset we find an even larger benefit from perspectives, corresponding to a 11.8 percentage point reduction in relative log error ($p < .001$).

Thus far we have seen that perspectives improve memory for what one has read as well as the ability to estimate unknown quantities. We turn now to our third and final measure of numerical comprehension, error detection.

Experiment 3: Error detection

In our final experiment we look at people's ability to detect errors in quotes from news articles, both with and without the

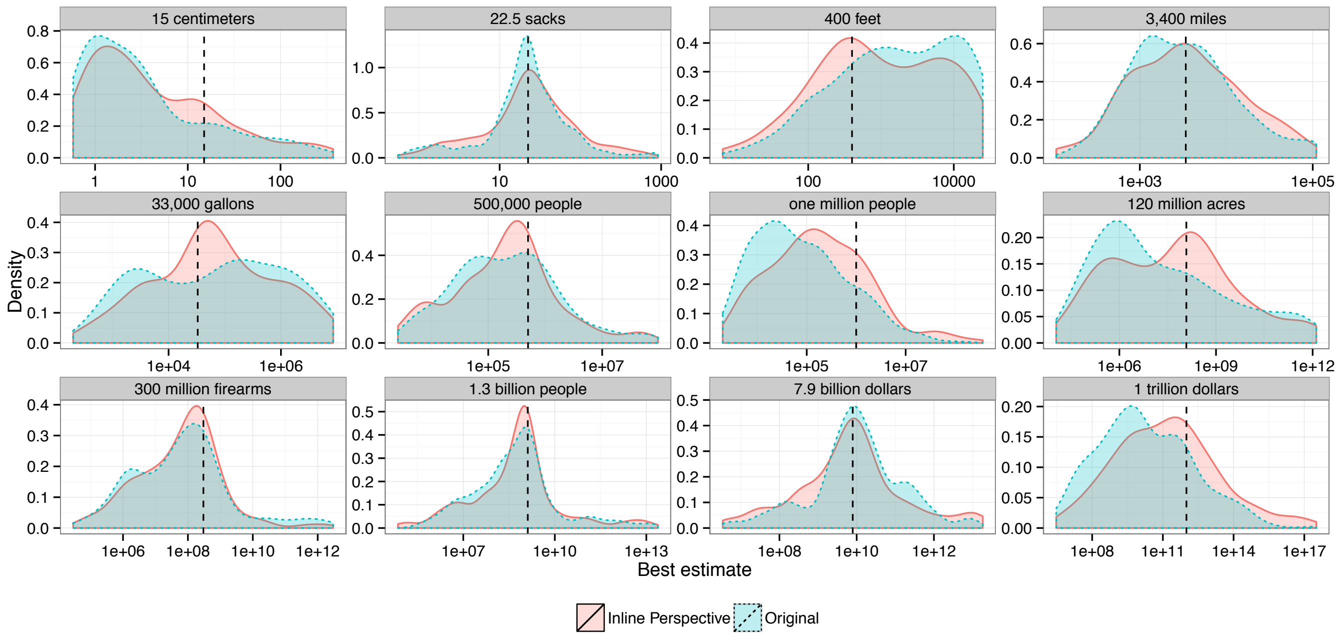

Figure 6. The distribution of participants' best estimates for missing measurements by condition.

aid of perspectives. Why should perspective sentences aid in error detection? Consider the example stating that 120 million acres of land worldwide were preserved by a nature conservancy group. Were this accidentally printed as one million acres, many readers might miss the mistake because acres are unfamiliar units. The addition of a perspective that rephrases one million acres as 1/100th the area of California might flag the measurement as too small to be newsworthy. In addition to putting things in more familiar units, perspective sentences may aid in error detection because they reiterate the key measurement, giving the participant a second chance to notice that something may be amiss. That said, because numbers that make the news tend to be exceptional, perspective sentences could cause correct values to be perceived as implausible. For instance, it may be even harder to believe that there are 300 million guns in the United States when this statistic is phrased as one gun per citizen, the highest ratio in the world by a large margin. Accordingly, it is unclear whether perspectives will help in the task of error detection, which is what we test in this experiment.

Online workers were once again recruited from Mechanical Turk and paid $1.00 to look for errors in all 12 quotes. Each quote was shown as plain text, with its corresponding measurement highlighted. Participants were told that this measurement "may or may not be modified from the original value that appeared in the actual article" and asked a simple question with a binary outcome: "Do you think the number highlighted in blue is the one that was actually printed in the original article?" Each participant was randomly assigned to either see a perspective (treatment) or not (control) across all 12 quotes presented to them. Those in the treatment condition received two extra instructions. The first explained that the perspective was always accurate with respect to the displayed number, regardless of whether the number itself had been modified. The second was to use the perspective sentence as an aid when reasoning about the highlighted number.

Each quote was presented in one of two conditions: either with the value that appeared in the original quote (the "actual" condition) or a predetermined plausible, but incorrect value (the "modified" condition). The modified value for each quote was chosen from the results of the estimation experiment above, using modal incorrect responses from the control group. This roughly corresponds to the most common incorrect value chosen when people were asked to estimate the measurement without any additional information, and results in a much more difficult test than the glaring typographic error discussed above. For instance, in the case of the Honduran storm, the modified value is 30,000 people—a number which is not entirely unreasonable, but is still substantially lower than the actual value of one million. The actual or modified condition was randomly assigned without replacement at the quote level for each participant, so that each person saw six quotes containing actual values and six with modified values in a randomly selected order.

As a result of the random assignment, 1,065 participants were assigned to see the original format, while 1,147 were assigned to see the perspective format. After ineligible participants (who had completed any of our previous experiments) were turned away and after eliminating participants who did not complete the experiment there were 660 and 644 in each group. This corresponds to completion rates of 98% and 97% for eligible workers in the control and perspective conditions, a non-significant difference ($p = .18$, χ^2 test).

Figure 7 shows participants' accuracy in error detection across quotes for both the control and perspective conditions,

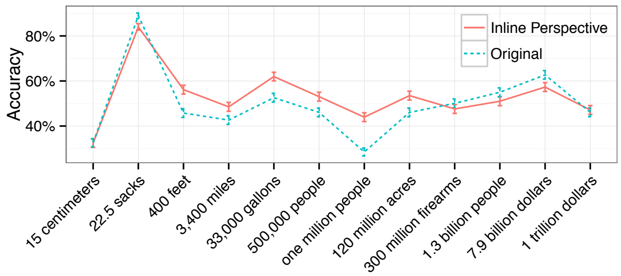

Figure 7. Classification accuracy for each quote, by condition, in the error detection task. Error bars show one standard error above and below the mean.

where a correct response corresponds to the user clicking "unlikely" when presented with a modified value or "plausible" for an actual one. Accuracy is rather low, varying from 30 to 60 percent for all but one quote, perhaps due to two likely causes. First, as mentioned above, the modified values we selected were not far from participants' estimates in the previous experiment—that is, these values were chosen to appear plausible. Second, regardless of condition, participants were overly liberal in accepting values—they selected "plausible" approximately two thirds of the time when only half of the presented values were correct.

We observe an average improvement of 3.2 percentage points in the presence of perspectives. To quantify this we regressed accuracy on indicators for the perspective format, manipulation condition (modified or not), and each quote. We also included an interaction term between format and manipulation as well as format and quote. This regression shows the expected interaction between format and manipulation, that is, perspectives helped in detecting erroneous quotes ($p < .05$). As shown in Figure 7, the impact of perspectives varied by quote. Gains from perspectives ranged as high as 15 percentage points, as in the Honduran quote. However, in select quotes we observe reversals, the largest of which is a 5 percentage point decrease in accuracy for the 7.9 billion dollar quote. We note that some of the reversals and weak patterns seem to roughly correspond to the cases in which people's uninformed estimates in Figure 6 (the blue densities) were rather accurate and low in variance. As we discuss below, whether perspectives should be selectively applied in such settings is a compelling hypothesis for future research.

DISCUSSION

In this paper we developed a framework that improves numerical communication. It is flexible enough to apply to wide range of settings, but simple enough to be understood and used by everyday readers. We examined how crowdsourced perspectives affect readers' comprehension and found that perspectives substantially improve people's ability to recall measurements they have read, estimate ones they have not, and detect errors in manipulated measurements.

We see this as the first of many steps in leveraging digital platforms to improve numeracy among online readers. As demonstrated here, perspectives are helpful in a variety of settings, but their utility depends on the underlying task, the considered measurement, and details of the perspective. This raises a series of questions around when perspectives should (and shouldn't) be employed, and what makes some perspectives useful but others less effective: How does one construct an effective perspective for a given statistic? Are certain types of perspectives (e.g., comparables or percentages) more useful than others (e.g., ranks and percentiles)? How does the saliency of the scaling factor affect comprehension? What is the tradeoff between the accuracy of a perspective and its helpfulness? How important is the use of a familiar reference entity, and to what extent should this be personalized to the individual reader? Can the discovery of these details be automated via information retrieval and machine learning algorithms?

Detailed answers to many of these questions fall outside of the scope of this work and require their own systematic studies. To see why, consider the example that rephrases 120 million acres as 1.15 times the area of California. It is possible—and perhaps even likely—that it would be just as effective to state this as "about equal to the area of California." It might even be the case that this simpler statement outperforms the more accurate, but likely more difficult to remember, perspective used in our study. Likewise, we could phrase 120 million acres as twice the area of Michigan, as this is factually more accurate than equating it to California's area while still employing a relatively simple multiplier. That said, some readers may be unfamiliar with Michigan's area as a reference quantity, which could have a negative impact on comprehension. Isolating these effects requires a carefully designed study that exogenously explores these different choices to uncover why some perspectives are more effective than others.

Another direction for future work is further exploration of how perspectives impact comprehension, learning, and generalization. Does repeated exposure to perspectives change the way people think when they encounter a new measurement, even in the absence of seeing a perspective for it? This could be tested by showing participants perspectives for one quantity and later asking them to estimate another. For example, once people know there is approximately one firearm per person in the United States, does this improve their ability to estimate the number of firearms in another country?

Finally, how should perspectives be deployed in practice, and what impact do they have on opinion formation and decision making? For instance, a typical voter in the United States may be unaware of how many registered firearms there are in the country. Mere exposure to the fact that there are 300 million such firearms might not affect their stance on gun control, as voters may have difficulty contextualizing this information. Stating this fact as one gun per person citizen, however, is likely to be more impactful, both because it is an easily understandable measurement and because it highlights the extremely high rate of gun ownership in the United States compared to the rest of the world. Conducting field experiments that measure such effects—especially through a live site, browser plug-in, or live editing tool—would give further insights into the real-world feasibility and impact of perspectives on numeracy and decision making.

REFERENCES

1. 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`

2. Daniel B Berch. 2005. Making sense of number sense: implications for children with mathematical disabilities. *Journal of Learning Disabilities* 38, 4 (2005), 333–339. `http://dx.doi.org/10.1177/00222194050380040901`

3. Michael Blastland and Andrew W Dilnot. 2009. *The Numbers Game: The Commonsense Guide to Understanding Numbers in the News, in Politics, and in Life*. Gotham Books, New York, NY, USA.

4. Norman R Brown and Robert S Siegler. 1993. Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review* 100, 3 (July 1993), 511–534. `http://dx.doi.org/10.1037/0033-295X.100.3.511`

5. Egon Brunswik. 1955. Representative design and probabilistic theory in a functional psychology. *Psychological Review* 62, 3 (1955), 193–217. `http://dx.doi.org/10.1037/h0047470`

6. Fergus IM Craik and Endel Tulving. 1975. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology* 104, 3 (1975), 268–294. `http://dx.doi.org/10.1037/0096-3445.104.3.268`

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, 1 (2013), 173–181. `http://dx.doi.org/10.1016/j.appet.2012.11.013`

8. Dedre Gentner and Allan Collins. 1981. Studies of inference from lack of knowledge. *Memory & Cognition* 9, 4 (1981), 434–443. `http://dx.doi.org/10.3758/BF03197569`

9. Gerd Gigerenzer. 2014. *Risk Savvy: How to Make Good Decisions*. Viking Books, New York, NY, USA.

10. J Greeno. 1991. Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education* 22, 3 (Jan 1991), 170–218. `http://www.jstor.org/stable/749074`

11. David Landy, Noah Silbert, and Aleah Goldin. 2013. Estimating Large Numbers. *Cognitive Science* 37, 5 (July 2013), 775–799. `http://dx.doi.org/10.1111/cogs.12028`

12. Richard P Larrick and Jack B Soll. 2008. The MPG Illusion. *Science* 320, 5883 (2008), 1593–1594.

13. Raymond Liaw, Ari Zilnik, Mark Baldwin, and Stephanie Butler. 2013. Maater: Crowdsourcing to Improve Online Journalism. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, New York, NY, USA, 2549–2554. `http://doi.acm.org/10.1145/2468356.2468828`

14. Isaac M Lipkus, Greg Samsa, and Barbara K Rimer. 2001. General Performance on a Numeracy Scale among Highly Educated Samples. *Medical Decision Making* 21, 1 (Feb. 2001), 37–44. `http://dx.doi.org/10.1177/0272989X0102100105`

15. Z Markovits and J Sowder. 1994. Developing number sense: An intervention study in grade 7. *Journal for Research in Mathematics Education* 25, 1 (Jan 1994), 4–29. `http://www.jstor.org/stable/749290`

16. W Mason and S Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (Jan 2012), 1–23. `http://dx.doi.org/10.3758/s13428-011-0124-6`

17. Alistair McIntosh, Barbara J Reys, and Robert E Reys. 1992. A proposed framework for examining basic number sense. *For the Learning of Mathematics* 12, 3 (1992), 2–44. `http://www.jstor.org/stable/40248053`

18. Dennis L McNeill and William L Wilkie. 1979. Public policy and consumer information: Impact of the new energy labels. *Journal of Consumer Research* 6, 1 (1979), 1–11. `http://www.jstor.org/stable/2488721`

19. Edward L Munnich, Michael A Ranney, and Daniel M Appel. 2004. Numerically-Driven Inferencing in instruction: The relatively broad transfer of estimation skills. In *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society*. 987–992.

20. Raymond S Nickerson. 1980. Motivated retrieval from archival memory. In *Nebraska Symposium on Motivation*. University of Nebraska Press, Lincoln, Nebraska, USA.

21. John Allen Paulos. 1988. *Innumeracy: Mathematical Illiteracy and Its Consequences*. Hill and Wang, New York, NY, USA.

22. Michael Andrew Ranney, Luke F Rinne, Louise Yarnall, Edward Munnich, Luke Miratrix, and Patricia Schank. 2008. Designing and Assessing Numeracy Training for Journalists: Toward Improving Quantitative Reasoning Among Media Consumers. *Proceedings of the Eighth International Conference for the Learning Sciences* (2008), 246–253.

23. Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013. One Half or 50%? An Eye-Tracking Study of Number Representation Readability. In *INTERACT 2013*. Vol. 8120. Springer Berlin Heidelberg, 229–245. `http://dx.doi.org/10.1007/978-3-642-40498-6_17`

24. Endel Tulving. 1962. Subjective organization in free recall of "unrelated" words. *Psychological Review* 69, 4 (1962), 344–354. `http://dx.doi.org/10.1037/h0043150`

25. Endel Tulving and Shirley Osler. 1968. Effectiveness of retrieval cues in memory for words. *Journal of Experimental Psychology* 77, 4 (1968), 593–601. `http://dx.doi.org/10.1037/h0026069`