Improving the Comprehension of Numbers in the News

Pablo J. Barrio
Columbia University
pjbarrio@cs.columbia.edu

Daniel G. Goldstein
Microsoft Research
dgg@microsoft.com

Jake M. Hofman
Microsoft Research
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 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 using digital platforms to improve numeracy among online readers.

1. INTRODUCTION

Consider a billion dollar budget cut or a million liter decrease in 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, and we shall show, people have difficulty remembering, estimating, and detecting errors in measurements sampled from everyday reading material.

Improving numerical literacy among the general population has been a long-standing challenge, with popular books [10] and programs [2] 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] Despite extensive literature on the topic [5] as well as a number of classroom-based studies on improving numeracy among students [7, 9] and journalists [11], there are few existing tools to help the common reader better understand unfamiliar measurements.

To date, most advances in numerical communication have fallen 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 miles" instead of as "miles per gallon" [6]. Likewise, creative ways to re-express the caloric content of foods (e.g., as the amount of exercise needed to burn them off) [3] and the energy consumption of appliances [8] 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 [4].

In this research, we go beyond the domains of risk and consumption to develop a more general system for improving numerical communication. We do so by taking advantage of digital platforms to both better understand how people consume quantitative information and to improve reading experiences. In particular, we show that simple sentences, termed *perspectives*, that employ percentages, ratios, rankings or other comparisons can be used to help people better understand arbitrary numerical measurements. 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 simple 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.

Take, for example, the quotes 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 our system (in italics), designed to make the measurement easier to understand. One of these quotes mentions the number of registered firearms in the United States. It can be difficult to estimate this number if one has never seen it before, and difficult to recall even if one has. Our experiments show that recall is substantially easier with the help of a perspective that rephrases the measurement as "about equal to 1 firearm for every person in the United States": while only 40% of people shown only the original quote were able to recall this number exactly, nearly 55% of participants who saw 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.

Where related work is concerned, a number of existing tools aim to improve online reading experiences. Popular sites such as Medium and NewsGenius allow readers to annotate articles with comments, but do not focus on quantitative information. Tools such as WolframAlpha and the Dictionary of Numbers focus on retrieving numerical information, but do not provide perspective clauses. In addition, we find no empirical research on the effects of these tools on numerical comprehension. Useful research has been conducted on simplifying the representation of numbers in text (e.g., writing "one half" instead of "50%") to improve reader understanding [1, 12], but not on perspective clauses. Accordingly, a test of the effect of perspective sentences on comprehension seems merited.

In the remainder of the paper we discuss how the quotes in Ta-

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

Quote and top-rated perspective
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.*
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.*

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.

ble 1 were generated and test the impact they have on numerical comprehension. First, we briefly describe the perspective 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 create and vote on perspectives that provide context to make the underlying measurements easier to understand. Based on user voting, the best perspectives are selected to appear within actual news articles as they are read.

Next, we test the effectiveness of perspectives through a series of large online experiments, which show that augmenting news articles with these perspectives improves people's ability to understand the magnitude of numerical measurements. In particular, we show that perspectives improve participants' ability to recall measurements they have read, to estimate unfamiliar amounts, and to detect errors in what they read. We begin by describing our framework for collecting helpful perspectives from crowd workers.

2. CROWDSOURCING PERSPECTIVES

We developed a crowdsourced system to collect perspectives for arbitrary measurements mentioned in online news articles from Amazon Mechanical Turk workers. Workers were shown a randomly selected quote from a New York Times front page article containing a highlighted measurement. Before and after the quote, they saw up to three adjacent sentences from the article in a smaller and lighter font. Below the quote, we displayed 10 templates that allowed workers to re-express the measurement in various formats (e.g., "x times larger than y", "about equal to x", "the largest x", "1 x for every y", or "in the top x%", ...). Each worker was allowed to add an unlimited number of perspectives for each quote 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 paid anywhere from \$0.05 to \$0.50 per perspective according to the perceived helpfulness of their contributions. In total, we collected 370 perspectives on 67 quotes from 80 different Mechanical Turk workers (an average of 4.6 perspectives per worker).

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.

3. EVALUATING PERSPECTIVES

Our objective is to test whether perspectives help people appreciate and comprehend numerical measurements. We assume that comprehension will be reflected in three measures—memory, estimation, and error detection—which we assess in three separate experiments. In each experiment, we use as stimuli 12 news quotes and the top rated crowdsourced perspective for each, a sample of which is shown in Table 1. 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 experiment that follow, we compute the relative log error between the value submitted by each participant and the actual 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: $|\log(actual) - \log(submitted)| / \log(actual)$, which allows us to account for (relatively common) large errors and to compare measurements of different magnitudes on a common scale.

3.1 Recall

In our first experiment, we test whether perspective sentences help people estimate or remember what they have read. Participants in this experiment read six news quotes, in plain text, containing numbers. Participants were randomly assigned to either see the original quote or to see the quote along with an additional perspective sentence. After a forgetting period, they were asked to recall the measurement of interest from each quote. Our hypothesis is that exposure to the perspective results in higher recall rates for the measurements of interest.

In all formats, the focal quotes were surrounded by a few sentences of text from the actual news article from which they were taken. 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. 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 the corresponding inline perspective sentence generated by our crowdsourced system. After reading the quotes, participants played Tetris (to provide forgetting time), followed by a surprise quiz in which they were shown the quote with the measurement missing and asked to fill in the

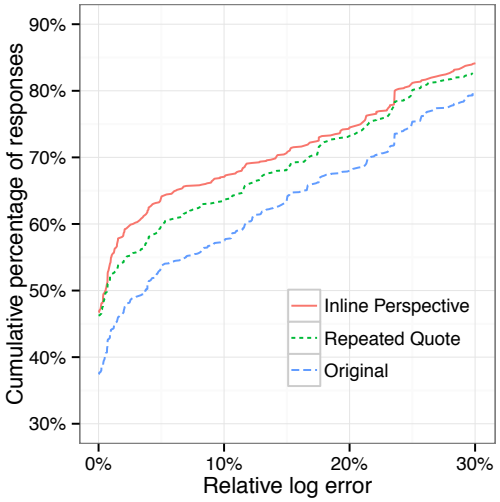

Figure 1: Accuracy of estimates as measured by relative log error, for original quotes, quotes with repetition, and quotes with inline perspectives.

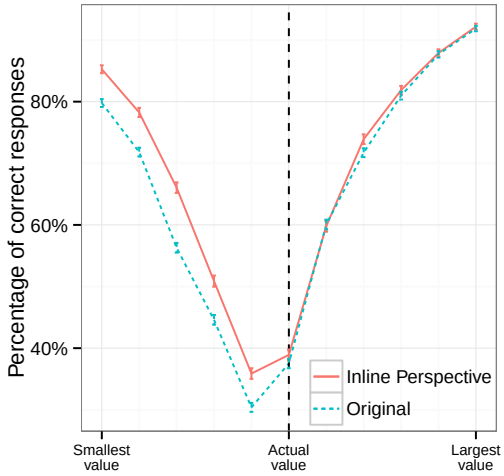

Figure 2: The percentage of responses correctly classified as "too low", "plausible", or "too high" in the estimation experiment for original quotes compared to those with inline perspectives. Error bars shown one unit of standard error above and below the mean.

blank and estimate what its value might be. These estimates are the dependent variable in this experiment.

Each worker saw six randomly selected quotes and 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. Participants gave answers before a 30 second countdown timer ran out, to prevent searching for answers online.

Figure 1 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. We compared the difference in percentage of responses at each 1% relative log error value shown in Figure 1 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).

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. Similarly, repeating the quote gives a 1.9 percentage point improvement over the original condition ($p = .0137$).

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). Our experiments demonstrate that the benefits of perspectives exceed that of mere repetition, a strategy that quickly grows tiresome and fail to teach readers anything new. These results are encouraging, but recall demonstrates only one aspect of comprehension. In the following sections we test two more—estimation and error detection.

3.2 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 reason about unknown quantities. In this experiment we test workers' accuracy in estimating the values of quantities they have *not* previously been exposed to, both with and without the aid of perspectives.

We recruited a new set of 1168 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 a highlighted inline perspective that rephrased candidate values (the treatment condition) for all six quotes that they saw. For example, if a worker entered a candidate value of 8 million people for the Honduran quote, the perspective expressed this as 97.5% of the population of Honduras, which might prompt the user to rethink their estimate.

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. 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 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

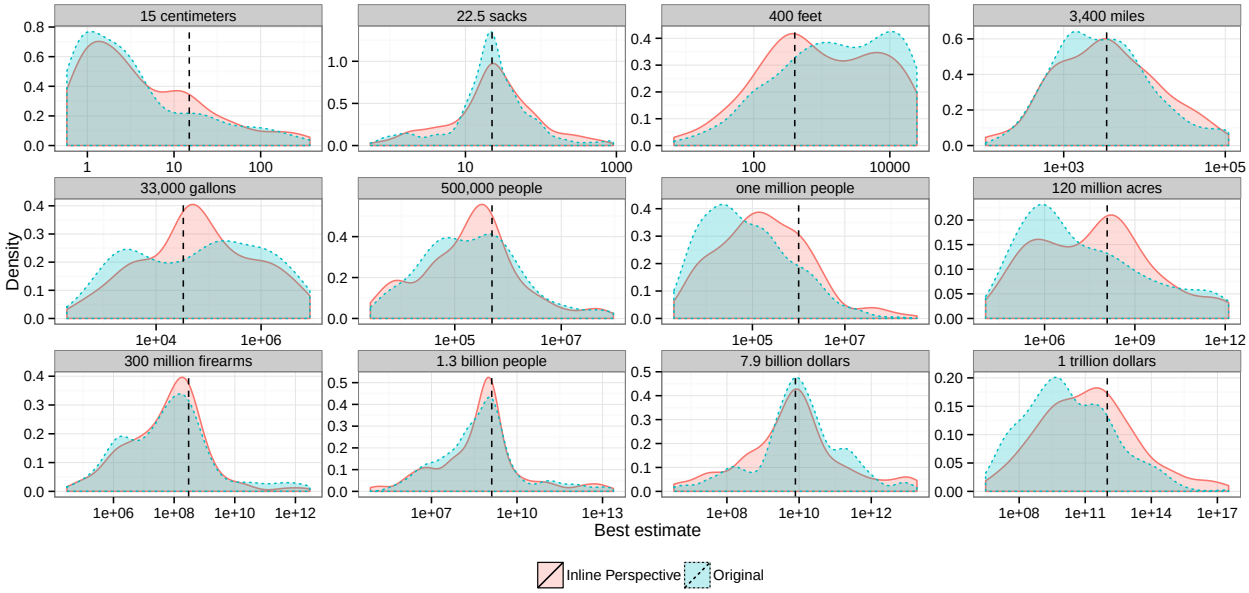

Figure 3: The distribution of participants' best estimates for missing measurements by condition.

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 moving to the next quote.

Figure 2 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 stage 1. 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 identifying the actual value. Furthermore, perspectives aided participants in rejecting incorrect values, particularly those below the actual value, where we observed improvements of 5 to 9 percentage points over the control condition.

Figure 3 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%. 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 3 shows that participants have a reasonably accurate estimate even without the aid of perspectives.

3.3 Error detection

In our final experiment we looked at people's ability to detect errors in quotes from news articles, both with and without the aid of perspectives. If perspectives improve understanding, we would expect to increase the odds that people can identify correctly printed measurements as well as values that are implausibly small or large. As an extreme example, consider a simple typographic error in the quote about the Honduran storm where "1 million people" is accidentally written as "10 million people". While this is likely to be picked up by careful readers, it might be missed by many others. The addition of a perspective that rephrases this as 20% larger than the entire Honduran population should make it much more likely that someone would catch this mistake. It is less clear if perspectives help—and if so, how much—for more subtle errors, which is precisely 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 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 dis-

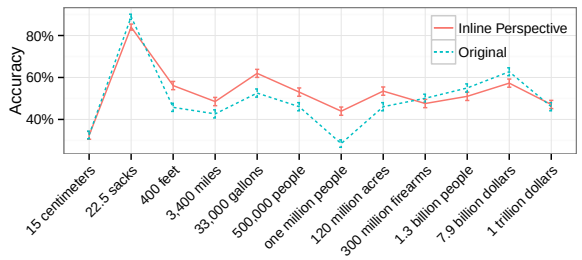

Figure 4: Classification accuracy for each quote, by condition, in the error detection task. Error bars show one standard error above and below the mean.

cussed 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 with their actual values and six with modified values in a randomly selected order.

As a result of the random assignment, 1065 participants were assigned to see the original format, while 1147 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, this left 660 and 644 in each group. This corresponds to completion rates of 98% and 97% for eligible workers in the control and perspective conditions, an insignificant difference ($p = .18$, χ^2 test).

Figure 4 shows participants' accuracy in error detection across quotes for both the control and perspective conditions, 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 interactions 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 4, 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 3 (the blue densities) were rather accurate and low in variance. Whether perspectives should be selectively applied to quotes for which people's uninformed guesses are poor is a compelling hypothesis for future research.

4. CONCLUSION

In this work 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 tested whether crowdsourced perspectives improve readers' comprehension and found that participants randomly assigned to see perspectives were substantially more accurate at estimating or recalling measurements and better at detecting errors in measurements they have read.

We see this as the first of many steps in leveraging digital platforms to improve numeracy among online readers. As shown here, perspectives are helpful in a variety of settings, but their utility depends on the underlying task and varies with the considered measurement. 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. 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 around it? Finally, how should perspectives be deployed in practice, and what impact do they have in more realistic settings? Conducting fields experiments though a live site, browser plug-in, or live editing tool would would give further insights into the real-world feasibility and impact of perspectives.

5. REFERENCES

[1] S. Bautista, R. Hervás, P. Gervás, R. R Power, and S. Williams. How to make numerical information accessible: Experimental identification of simplification strategies. In *INTERACT 2011*, volume 6946, pages 57–64. Springer Berlin Heidelberg, 2011.

[2] M. Blastland and A. W. Dilnot. *The numbers game: The commonsense guide to understanding numbers in the news, in politics, and in life*. 2009.

[3] S. Dowray, J. J. Swartz, D. Braxton, and A. J. Viera. Potential effect of physical activity based menu labels on the calorie content of selected fast food meals. *Appetite*, 62(0):173–181, 2013.

[4] G. Gigerenzer. *Risk savvy: how to make good decisions*. Viking Books, New York, NY, USA, April 2014.

[5] J. Greeno. Number sense as situated knowing in a conceptual domain. *Journal for research in mathematics education*, Jan 1991.

[6] R. P. Larrick and J. B. Soll. The mpg illusion. *Science*, 320(5883):1593–1594, 2008.

[7] Z. Markovits and J. Sowder. Developing number sense: An intervention study in grade 7. *Journal for research in mathematics education*, Jan 1994.

[8] D. L. McNeill and W. L. Wilkie. Public policy and consumer information: Impact of the new energy labels. *Journal of Consumer Research*, pages 1–11, 1979.

[9] E. L. Munnich, M. A. Ranney, and D. M. Appel. Numerically-driven inferencing in instruction: The relatively broad transfer of estimation skills. *CogSci*, 2004.

[10] J. A. Paulos. *Innumeracy: Mathematical illiteracy and its consequences*. 1988.

[11] M. A. Ranney, L. F. Rinne, L. Yarnall, E. Munnich, L. Miratrix, and P. Schank. Designing and assessing numeracy training for journalists: Toward improving quantitative reasoning among media consumers. pages 246–253, 2008.

[12] L. Rello, S. Bautista, R. Baeza-Yates, P. Gervás, R. Hervás, and H. Saggion. One half or 50readability. In *INTERACT 2013*, volume 8120, pages 229–245. Springer Berlin Heidelberg, 2013.