# Same Words, Same Context, Different Meanings: People are unaware that their own concepts are not always shared

**Louis Martí (I_Am_A_Robot@berkeley.edu)**
Department of Psychology, 2121 Berkeley Way
Berkeley, CA 94704 USA

**Steven Piantadosi (stp@berkeley.edu)**
Department of Psychology, 2121 Berkeley Way
Berkeley, CA 94704 USA

**Celeste Kidd (celestekidd@berkeley.edu)**
Department of Psychology, 2121 Berkeley Way
Berkeley, CA 94704 USA

**Abstract**

A long-standing assumption in cognitive science has been that concepts are shared among individuals for common words. However, given that concepts are formed by the data we observe, and observations vary wildly across individual experiences, our concepts are not likely identical. Here, we present data in which 104 participants answer questions regarding their beliefs about the definitions of common everyday words, and the degree to which they think others agree. Our results suggest that even for common words, there exist many distinct extensions of ordinary and political concepts across individuals. There is also a pervasive bias which leads individuals to overestimate the degree to which others agree, which may explain why "talking past each other" is an anecdotally common experience when discussing important topics.

**Keywords:** Concepts; Metacognition; Individual Differences; Miscommunication

## Introduction

In 1964, the United States Supreme Court heard Jacobellis v. Ohio, a case in which theater owner Nico Jacobellis was fined for exhibiting a dramatic French film about adultery that contained material which the state considered obscene. Justice Potter Stewart, in explaining why he believed the film did not violate the state's obscenity laws, stated that he was unable to define pornography but also said "I know it when I see it". In this instance, the highest authority in the country was tasked with categorizing an edge-case which would affect the lives of millions and he admitted that his criteria for categorization was difficult to articulate.

Even words with seemingly precise meanings can be deceptively ambiguous. Especially if neither party anticipates a problem because of the word's commonality. These non-obvious misalignments can have serious consequences. For example, toxicologists and non-toxicologists likely have different concepts for the word "hazard". Toxicologists define the word as referring to anything that could potentially cause harm—even if unlikely. For example, a toxicologist would categorize water as a hazard because it is possible to overdose if excessive quantities are consumed. However, for non-toxicologists, the word "hazard" refers to things which are dangerous—*likely* to cause harm, not simply capable of causing harm under specific or unlikely circumstances. This misalignment caused problems when toxicologists with the World Cancer Association labelled coffee as a known hazard for developing cancer in mice and cell cultures. A California judge, who likely possessed the concept synonymous with "dangerous", interpreted the report as meaning coffee was dangerous for consumers and ruled that California had to warn consumers.

These examples illustrate that not only can words be hard to define, but we sometimes have very different ideas about what they mean. When two people use the same word, they may assume that they are each referring to the same (or at least a similar) concept. But how often is this assumption correct? Communication requires involved parties to understand each other correctly. A necessary component of this during language use is that words map onto the same meanings for all conversation partners, or, alternatively, that they are at least aware of the possibility for misalignment. If this is not in fact occurring, it could provide new insights into why and how people disagree and misunderstand one another. Understanding these dynamics could, likewise, be used to facilitate better communication in general. Thus, conceptual misalignment has important implications in a wide range of domains, including public policy, diplomacy, education, and politics.

All theories of concepts involve learning via interaction with and data accumulation from the world. These experiences vary (often wildly) across individuals. If individuals are using the same word to refer to two different concepts, confusion and miscommunication may occur. There is some empirical evidence that at least some of people's concepts do in fact vary across individuals (McCloskey & Glucksberg, 1978). Labov (1973) asked participants to categorize objects as either a "cup" or a "bowl" as he varied the heights and widths of the objects. For extreme values of either height or width, there was widespread agreement on the classification. However, as the values became more moderate, the classifications became more subjective. This demonstrates that people's concepts are fuzzy along the edges, even with everyday

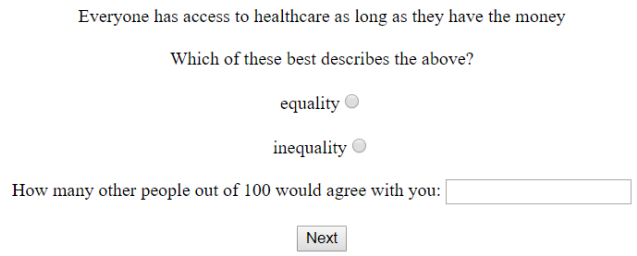


Figure 1: Participants saw 200 randomized trials as above.

objects.

If people do not agree even on category boundaries for concrete objects, how much misalignment might exist among more abstract concepts? And are people aware of the fact that concepts vary across individuals? To date, there has been no work exploring these questions.

Our first goal is to quantify individual differences in conceptual representations. By utilizing an approach which targeted edge-cases, we optimized our chances of detecting differences in definitional boundaries. Edge-cases are both theoretically and functionally important. They are crucial to conceptual definitions, and are arguably where the highest utility can be found due to the possible illusion of confidence people have about others' definitions. For example, common debates about abortion, gun-control, and welfare hinge on edge-cases.

We did this by collecting peoples assessments of whether particular scenarios applied to specific concepts. We then applied a clustering algorithm to group participants by similarity of their conceptual representations. We borrowed techniques from machine learning that have previously been applied in biology and ecology in order to estimate the total number of distinct representations on a population level. Our second goal is to quantify peoples metacognitive awareness of differences in each others conceptual structures. If people are unaware of the variability in others classifications of everyday concepts, it would make communication more difficult. In this paper, we will probe the variability of people's concepts and measure their awareness of any differences.

## Methods

We recruited 104 participants on Amazon Mechanical Turk and queried them regarding their beliefs about whether a particular word applied to a given phrase or sentence. For each of 200 trials (see Figure 1) a phrase or sentence was displayed. The participant was then presented with two opposites and asked to classify the phrase or sentence. For example, after reading the sentence "A murderer is killed", participants answered whether they thought it was *justice* or *injustice*. They also answered how many people out of 100 would agree with them.

| Word | Sentence | Reliability Pair |
|---|---|---|
| justice/injustice | A guilty man is executed | A man who is guilty is put to death |
| adult/child | A 17-year-old | An individual who is almost 18 |

Table 1: Sample reliability sentences

| Word | Phrase/Sentence |
|---|---|
| equality/inequality | Taking wealth from the rich and giving it to the poor |
| fairness/unfairness | Paying none of your workers because you don't have the money for everyone |
| justice/injustice | A thief's stolen property is stolen |
| peace/conflict | A field filled with corpses after a war is over |
| honesty/dishonesty | Making true but misleading statements |
| safety/danger | Preventing you from drinking soda |
| freedom/prohibition | Making murder illegal |
| transparent/secretive | Releasing your taxes behind a paywall |
| education/ignorance | Home schooling in the US |
| healthcare/illness | Insurance not paying for your medical bills despite you paying your premiums |
| day/night | Dusk |
| hot/cold | A temperate day |
| light/dark | Classical music |
| friend/enemy | A close acquaintance who insults you all the time |
| boy/girl | A transgender woman |
| love/hate | Spanking a child so that they will not become spoiled |
| adult/child | A 17-year-old |
| good/bad | An entire building full of murderers was destroyed |
| sun/moon | A star that orbits a planet |
| ceiling/floor | The top surface in an upside-down house |

Table 2: Sample stimuli presented to subjects in the experiment.

### Word Choices

Stimuli were divided into ten political words and ten frequently used nouns. Half of all participants answered questions regarding political words while the other half answered questions about the nouns. Political concepts were chosen by asking 130 mTurkers to list the top ten words they felt were most relevant to politics. The top ten most frequent words were then chosen as our political concepts. The ten nouns

were chosen by querying the MRC Psycholinguistic Database for the ten most frequent nouns and omitting words which were close semantic duplicates (e.g. boy vs. man). This was done in order to maximize semantic variability in our word pairs as much as possible.

### Sentence Construction

The specific sentences participants are responding to for each word are of crucial importance. One might imagine a set of sentences could be chosen for the word "boy" which would result in near universal agreement among participants. On the other hand, sentences could be constructed in such a way as to maximize disagreement (a 50/50 split for each binary response). If our goal is to discover whether people possess different concepts, the latter approach is appropriate. Specifically, since edge cases are often where the greatest variability lies, we will probe people's classifications of edge cases. This approach allows us to get a rough estimate of the maximum conceptual variability for each word (see Table 2 for a sample of phrases/sentences).

### Reliability

Each trial had an associated reliability trial which presented the same phrase or sentence except for a minor modification which did not change the meaning. These were added in order to assess subject attention and reliability. (see Table 1 for a sample of phrases/sentences)

### Analysis

Ascertaining whether participants possess different conceptual representations for a given word is a non-trivial problem. We first run into the problem of how to quantify differences between conceptual representations. What does it mean for person A's concept of "justice" to be twice as far from person B's as person C's is? We address this by representing each person's concept using a binary response vector. Next, we run into the issue of measurement noise and participant reliability. If person A answers that "A clear night with a full moon" is "light" but also answers that "A cloudless night with a full moon" is "dark", it would be reasonable to label these responses as unreliable noise. The "reliability" sentences provide semantic duplicates for each sentence, allowing us to quantify the reliability of participants in the task.

Once we have quantified participants' reliability and concepts, our last major issue is deciding how much of a difference between two concepts is sufficient to call them distinct. For the concept blue, one individual might be centered on the 470 nanometer wavelength while another might be centered on 480 nanometers. What would not be clear, however, is whether that disparity is sufficiently different so as to reasonably characterize the individuals as having separate concepts for blue. We approach this challenge by clustering our participants such that people with similar concepts will be grouped in the same cluster. We do this by adopting Bayesian approaches that find the optimal partition of participant responses using a trade-off between data-fit and simplicity. If the responses of one participant are very similar to those of another participant, they will likely be placed in the same cluster. On the other hand, if two participants have very different responses, they will likely be placed in different clusters, despite the process's overall conservative preference for fewer total clusters (Anderson, 1991). More specifically, we will use a Chinese Restaurant Process prior. If $[x_1, x_2, \ldots, x_k]$ is a vector denoting how many of the $n$ subjects have each concept (for a given word), then the CRP prior is

$$P([x_1, x_2, \ldots, x_k]) = \frac{1}{n!} \prod_i (x_i - 1) \tag{1}$$

Within each "table" of the CRP, we use a Beta-Bernoulli likelihood, meaning that subjects assigned the same cluster are assumed to generate the same latent vector of binary answers. This vector is then measured with noise ($\alpha$), and the latent probabilities are integrated out. Thus, if $y_j$ and $n_j$ are the number of "yes" and "no" responses in a given cluster assignment to the $j$'th item of a given concept, then the likelihood is,

$$\prod_j \frac{\Gamma(2\alpha) \cdot \Gamma(y_j + \alpha) \cdot \Gamma(n_j + \alpha)}{\Gamma(y_j + n_j + 2\alpha) \cdot \Gamma(\alpha)^2 \cdot y_j! \cdot n_j!} \tag{2}$$

With this setup, we used a Gibbs sampler to sample from the posterior on clusters given the responses for each concept. This analysis provides us with the number of distinct conceptual representations possessed by our participants. While this is useful information, what we are actually interested in is the total number of conceptual representations which exist on the planet. Ecologists have faced a very similar problem in estimating the number of species which exist. Often, they possess observed counts of individuals and of species in a given area (the Amazon rainforest for example) and would like to estimate the true number of species for that area (Bunge & Fitzpatrick, 1993). Here, we use the number of sampled concepts across the number of sampled individuals to estimate the total number of concepts which exist across the population of Earth for each of our words. Given that our clustering algorithm has a conservative preference for fewer clusters, this preference will extend to our global estimate.

## Results

We excluded participants who did not have a reliability greater than 70% (9 out of 104 participants). We also excluded participants who gave the same answer to all agreement-prediction questions (2 out of 95 participants). We did not require participants to perform flawlessly, however, as this would be an unrealistically high bar for humans completing so many trials. Of the remaining participants, their probability of giving the same answer to both questions in the reliability pair was a respectable 86%.

### There are about five distinct concepts per word

Figure 2 shows the estimated true number of concepts (y-axis) across 4,000 iterations of our clustering algorithm (us-

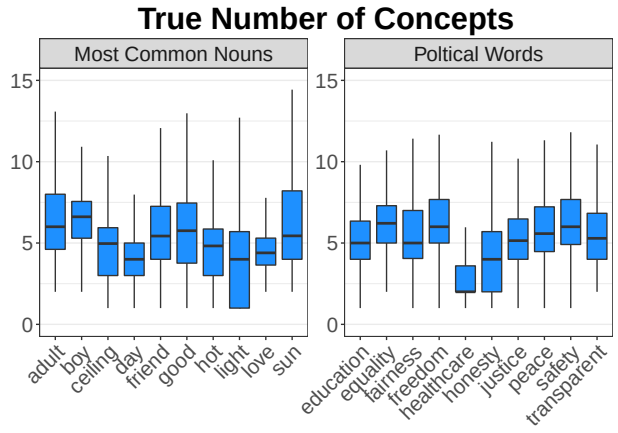


Figure 2: Estimated true number of concepts for 4,000 iterations of our clustering algorithm using a simplicity prior after a 1,000 iteration burn-in. Median estimates are roughly the same regardless of the type of word (about 5).

ing a simplicity prior) after 1,000 iterations of burn-in for each word (x-axis). Across the 4,000 samples from each word, the median estimates are always about the same regardless of word type: roughly five concepts. We also ran our clustering algorithm using a uniform prior which resulted in the same pattern of estimates except increased by two. In comparison to a simplicity prior, a uniform prior will prefer a larger estimate of concepts. That a uniform prior resulted in seven concepts, not 7,000, strongly suggests that the true number of concepts for our chosen words is close to our estimates.

## Additional participants are unlikely to significantly increase our estimates

Additionally, we can run our algorithm with varying amounts of data in order to confirm our results. As the number of participants we sample increases, we should expect the distance between our sample estimate and global estimate to narrow and eventually converge. Figure 3 shows the number of concepts (y-axis) by the number of people sampled (x-axis). For most words, as the number of people sampled increases, the true number of concepts (in blue) also increases. This suggests that our estimates for these concepts are relatively conservative, as it is unlikely we have sampled enough to cause this process to plateau. This, in addition to the inherent conservatism in our clustering algorithm (the simplicity prior), suggests these are lower bound estimates for our tested concepts. However, given the slow rate of increase between sample sizes, it is unlikely our estimates would ever grow significantly, even with many more participants.

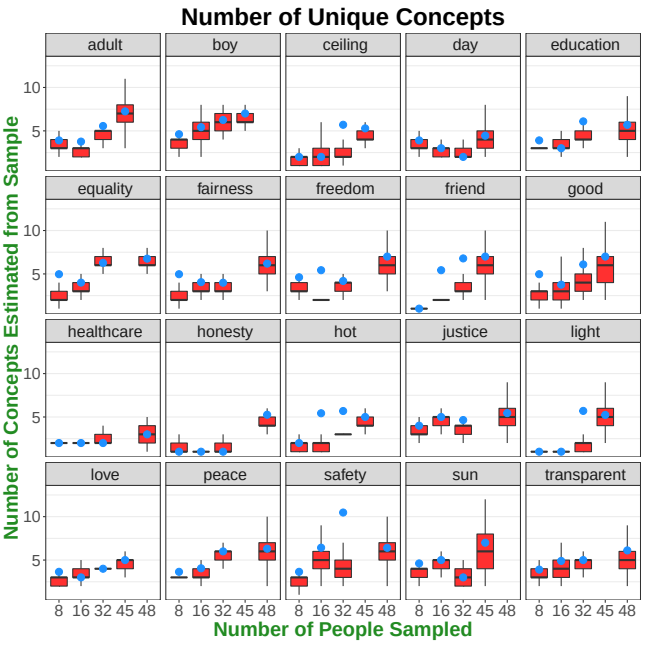


Figure 3: Number of concepts (y-axis) depending on the number of people sampled (x-axis) using a simplicity prior. Box-plots represent 25% to 75% quantiles of the number of concepts in our sample. Blue dots represent the estimated number of unique concepts on Earth based on our sample estimates. As the number of people sampled increases, the number of estimated concepts tends to increase by a slowing amount. This suggests that although our current estimates are conservative, the true number is not much higher.

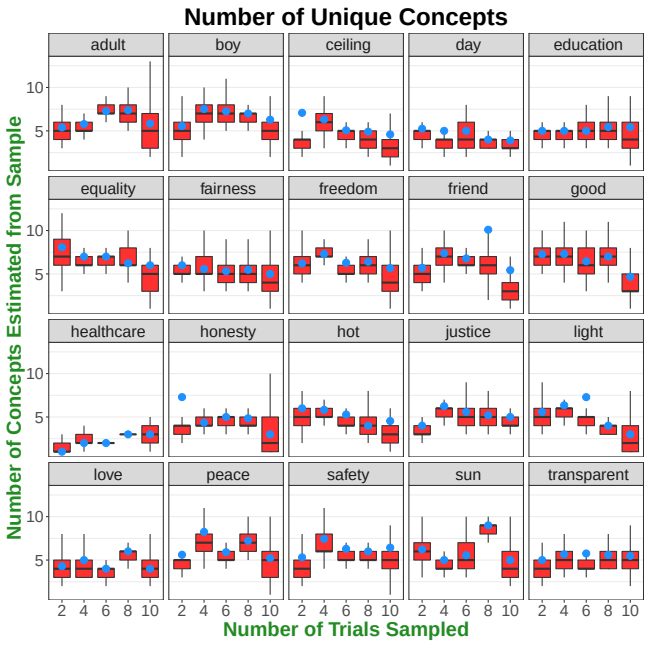


Figure 4: Number of concepts (y-axis) depending on the number of sentences sampled (x-axis) using a simplicity prior. Box-plots represent 25% to 75% quantiles of the number of concepts in our sample. Blue dots represent the estimated number of unique concepts on Earth based on our sample estimates. As the number of sentences sampled increases, the number of estimated concepts tends to stay the same. This suggests that our sentence choices are sufficiently varied enough to capture concept diversity.

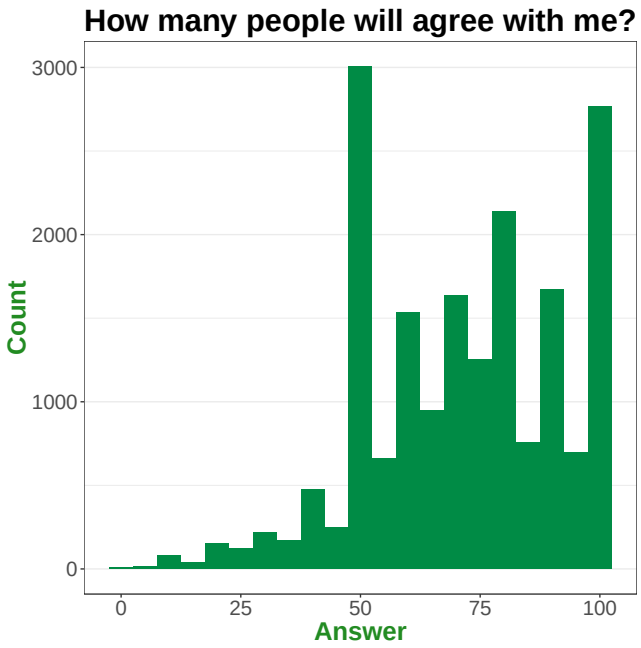


Figure 5: Raw counts (y-axis) of participant answers to the question "How many other people out of 100 would agree with you?" (x-axis). People overwhelmingly think the majority of other people will agree with their assessment.

## Additional sentences are unlikely to significantly increase our estimates

Figure 4 shows the number of concepts (y-axis) by the number of sentences sampled (x-axis). If concepts were unique for each individual, one would expect the number of estimated concepts to steadily increase as the number of sentences increased. Instead, we see our estimate as relatively stable, even sometimes *decreasing* as more sentences are sampled. This also suggests that our sentence choices were sufficiently varied to capture concept diversity.

## Most individuals underestimate conceptual variability

We then examined people's guesses about how often others' agreed with them. Figure 5 shows raw counts (y-axis) of participant responses (x-axis). The figure illustrates a very strong "like me" bias where the overwhelming number of responses indicate a belief that most others' will agree with their categorization. The second most common response was that *all* others will agree with their assessment.

We then assessed the relationship between people's categorizations to their guesses about the answers of others. Figure 6 presents people's predicted answers (y-axis) and their actual answers (x-axis). As the figure shows, a sizeable number of trials are not well predicted by participants. A perfect prediction rate would result in all trials landing on the y = x line. Although there are many trials which fall on or near this line, there also seems to be a consistent trend of partic-


Figure 6: Participants' actual responses vs. the responses they expected others' to give. Each data point represents the mean response for a trial/choice pair. Most data points are above the y = x line, illustrating that people largely overestimate to which others' agree with their assessments.

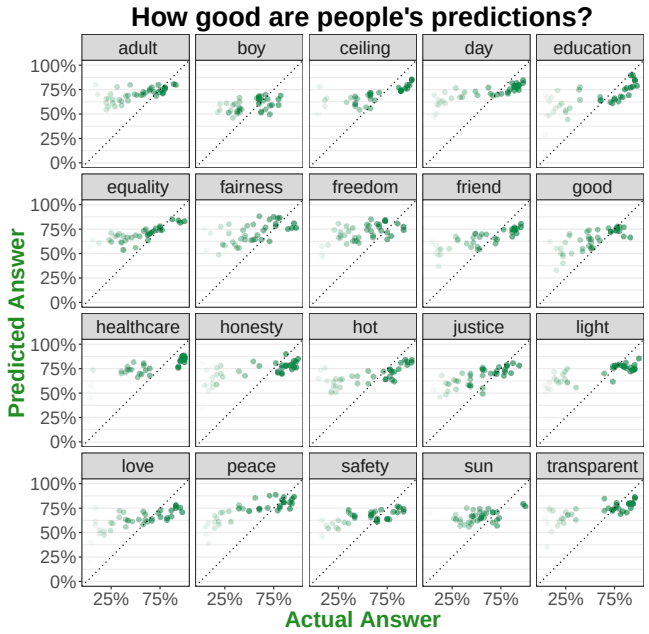


Figure 7: Participants' actual responses vs. the responses they expected others' to give, binned by word. Each data point represents the mean response for a trial/choice pair. People largely overestimate the degree to which others' agree with their assessments, regardless of the concept being assessed.

ipants overestimating the number of people who agree with them as the number of points above the y = x line shows. In fact, very few trials are underestimated and those which are, are only barely underestimated. In contrast, many trial predictions wildly overestimate people's actual responses. Examining the data by word (see Figure 7) shows these trends are not confined to a small subset of words but rather, are widespread.

## Conclusions and Discussion

The degree to which conceptual representations are shared and the degree to which people are aware of any differences are also fundamentally important aspects of any theory of conceptual structure, but both have been largely neglected.

These results, along with prior literature, provide strong evidence that the diversity in conceptual representations has been underestimated. As Figure 2 shows, concepts have roughly five to seven different representations, even for basic words such as "day" or "night". This is a surprising finding from multiple points of view. If you believe everyone holds the same concept for the same word, anything greater than one will be unexpected. On the other hand, if you believe concepts are infinitely distinct across individuals and across time, our estimate will also be unexpected.

Furthermore, individuals seem to be unaware of these differences. Figure 6 illustrates the poor relationship between people's actual answers and people's guesses about the answers of others. Taken together, these findings have strong implications for the way humans communicate. Misunderstandings are likely to occur if two individuals are operating with different representations of the same word.

### Limitations

It is possible that participants may have interpreted some sentences differently. If two participants have completely different interpretations of the same sentence, they may in reality possess the same concept, but appear to possess different concepts. We do not believe that this possibility could have driven our reported effects, however, because sentences were constructed in order to reduce ambiguity (though, of course, eliminating all ambiguity is impossible).

### Summary

There is measurable variability in the conceptual representations attached to particular words (greater than zero but less than infinity); importantly, this variability applies to both concrete words (e.g., "sun") and abstract ones (e.g., "freedom"). More importantly, our data shows that individuals are poorly calibrated to this variability and generally underestimate it. This is important, because communication requires that interlocutors understand one another. These results could help explain a previously unappreciated source of miscommunication and misunderstanding between people.

## Acknowledgements

We would like to thank members of the Kidd Lab, and the Computation and Language Lab for providing valuable feedback.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, *98*(3), 409.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, *88*(421), 364–373.

Labov, W. (1973). The boundaries of words and their meanings. *New ways of analyzing variation in English*.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.