Understanding the Effects of Different Comparison Sets of Information on the Comprehension of Generic Statements through a Bayesian Inference Model

(Working Paper)

Xiuyuan Zhang *

June 6, 2018

*Department of Psychology, The University of Chicago. xiuyuanzhang@chicago.edu. The funding for this project come from Communication and Learning Lab at The University of Chicago. This is a preliminary draft paper for completion of a grad-level course computational research at the University.

Abstract

The generic statement "Mosquitoes carry West Nile" is neither a claim about specific mosquitoes, nor all mosquitoes, nor even most mosquitoes. Very few mosquitoes carry West Nile; nonetheless, "Mosquitoes carry West Nile" is acceptable because mosquitoes are likely to carry West Nile relative to other insects. If generics are evaluated relative to a comparison set (e.g., the set of insects), the specific comparison set one uses then determines the inference made. We test this prediction in a series of studies on Mechanical Turk in which participants (n = 150) were asked to draw inferences about novel categories from generics. Participants read about a novel category and its comparison set (e.g., daxes are like trucks), then read a novel generic (e.g., daxes are heavy), and judged the prevalence of this feature (e.g., what percent of daxes do you think are heavy?). Across participants, we presented comparison sets (e.g., trucks, rocks, and bikes) with different prevalence rates for the feature of interest (e.g., heavy.) After reading the same generic, participants anchored their prevalence judgments for novel categories differently based on the particular comparison set. These results suggest that different comparison sets yield wildly different inferences about a novel generic. While we manipulated these comparison sets explicitly, in naturally occurring generics people bring their own background knowledge to bear. These differences may form the basis for transmission of stereotypes and misinformation–a hypothesis that we are exploring in ongoing work.

INTRODUCTION

Human beings live in a contextually-rich environment. Constantly many types of inputs are fed into various perceptual channels; living in a social group that has speech as one of its communicative tools for members of the group, linguistic inputs contribute to the systematic understanding a human being has toward their surroundings. During a speech event, for the addressee (a linguistic term for 'listener' used by Jakobson, addresser in this case is the speaker and the addressee is the listener) to have a basic understanding of the addresser's message, the addressee has to already have a systematic understanding not only of the possible meanings of each word in the message, but also of the social statuses of both the addresser and the addressee from experience and sociopolitical cultivation(Jakobson, 1960; Silverstein, 1976). To explore this linguistic question further in the field of psychology, we choose to study a representative syntactical structure: generic statement.

Generic statements are statements that inform the addressee about a category and its characteristic feature. What is interesting about generic statements is that, despite the statement about the category and its given feature being generally accepted as true, the percentage of the members in that category which has this given feature does not stay fixed. We call this percentage the prevalence rate that reflects a ratio of the amount of members of a category that possess this feature to the whole count of members of the said category. Generic statements do not have to have more than half of the members of the said category to possess a given feature for them to be accepted. If there is not a fix threshold that the prevalence rate has to surpass, what makes a generic statement acceptable to the addressee? Is it possible that this acceptance of the generic statement is not based solely on the generic statement itself but what addresses are made to think of when they hear this generic statement? We ask: how does the comprehension of an utterance from different addresses reflect the possible different sets of concepts they have?

THEORY & PAST WORKS

There are several past works in the field of social and developmental psychology that have greatly inform and motivated this paper. These projects collectively make an effort to connect the study of human behaviors with language. L. Ross and A. Ward's paper makes two major assertions about the communication between human beings as well as human beings' subjective recognition of themselves and others: (1) differences in subjective interpretation or construal matter, that they have a profound impact in the conduct of everyday social affairs. (2) social perceivers characteristically make insufficient allowance for such impact in the inferences and predictions they make about others. This paper includes a detailed account of multiple studies done by L. Ross and others on the topic of naive realism. Among the many studies that L. Ross and A.Ward referenced, one study revealed that the manipulation of labels and language can be used effectively to disengage normal mechanism of moral evaluation. It shows that, during the process of perspective taking, human beings frequently make mistakes on their estimation of what others are thinking. In their paper, L. Ross and A. Ward suggest that there are two direction in which the next step in psychology research should head to in studying this phenomena - studying the direct influence of manipulation of labels and language and the indirect (Ross & Ward, 1996).

In M. Rhodes et al. paper on the cultural transmission of social essentialism, the authors set out to study how generic languages facilitates the transmission of essentialist beliefs about social categories from parents to children. This study by M. Rhodes et al followed a prior project done by MG. Taylor along with M. Rhodes and S. Gelman on the existence of essentialist beliefs by participants from the age 7 - 10 years old who are in communities that respectively have politically (1)conservative and (2)liberal takes on the issue of race (Taylor et al., 2009). In their 2012 study, M. Rhodes et al. recruits both children and adults participants, introducing them to a novel creature called Zarpies. They studies different forms of expressing a generic

statement, from using bare plural, indefinite singular, specific singular, to using no-label as the referential subject. Participants are told stories about Zarpies framed in line with using the aforementioned four categories in the generic statement. The objective of study is to find out whether different forms of generic statements convey different messages to participants about the inherent quality of Zarpies, i.e. social essentialism. The result of this study indicates that generic languages using bare plural facilitates the idea of inheritance of qualities.

In this working paper, M.H Tessler and N.D. Goodman studies several aspects of generalization that are formulated through the use of language. Their work includes two major parts: first, they proposed their hypothesis on people's inference process based on a Bayesian inference process, counting the conditional probability based on some prior event as a crucial variable in their model. They went through the mathematical derivation for their model before went on the field, using Amazon Mechanical Turk to recruit their participants and run survey studies. For the generic statement study specifically, the authors collected data on participants' estimation for the baseline prevalence for certain category $C_{familiar}$ and its given feature F. They also provided different narratives to introduce participants to novel categories C_{novel} with the same set of features F, to see how the co-occurrence of $C_{familiar}$ and C_{novel} within the same narrative would influence the way in which participants make their estimation on feature prevalence rate P about the novel category C_{novel} (Tessler &Goodman, 2018).

L. Ross and A. Ward's work on naive realism, showcasing not only people's difference in their subjective construal of the world but also the potential negative consequence of such construal on a social level, has led us to ask: how did this difference in how people perceive themselves, the world and their interlocutors occur? Can we build a model that tries to pinpoint where it is more likely that this difference is generated? Although L. Ross and A. Ward brought up this question of naive

realism in the field of social psychology, it is an equally important subject of study for researchers in developmental psychology. Since the manifestation of adult naive realism can be traced back to the developmental stages for children, it is important to consider this question also understand the framework of an ontogenesis of subjective construals.

Following from L. Ross and A. Ward's direction as to study the direct effect of language manipulation, the studies by M. Rhodes et al. looks at the potentially negative effect of generic statement, transmitting cultural essentialism (Rhodes et al., 2012). One question that is worth asking is: what is the benefit to the wide use of generic statement if its presence is to the detriment of our society? One suspicion is that people are economical with their words, or that the most optimizing way of conveying information for them is through generic statement. Moreover, it is not directly harmful for people themselves when they use generic statement to describe others. The over-generalization doesn't work against their interest. It is then a problem of whether, when speaker's priority may be not presenting the most truthful information, what is the inferential process on the listener's part? This is one of the reasons that we are building our current speaker-listener interaction model.

Combining Rhodes et al. study with Tessler and Goodman's current project on the language of generalization, we set ourselves to study how certain essentialist's concepts are transferred from a speaker to a listener. At the moment, we are using a similar platform (Amazon Mechanical Turk) as Tessler and Goodman, running surveys that uses the introduction of a novel category C_{novel} and then ask about the prevalence rate of certain features F. Instead of setting up to directly provide participants with familiar comparison categories $C_{familiar}$, we are manipulating the demographics of participants(gender and age) and the question that we are asking to see if we could use their possibly already-existed stereotypes and biases to show this phenomenon. Another next step that we are pursuing is that we also are in commu-

nication with M. Rhodes to discuss methods that we can extend our own research on young children offline in our lab, so that we can get a more complete picture of how this understanding of generic statement or estimation for novel categories using prior background information changes or remains the same across developmental stages.

HYPOTHESIS

On a theoretical level, we are interested in learning about how people learn new knowledge about categories that they were just introduced to. If a listener hears something that is unknown to her, what is the next step of cognitive process that would happen for her? We hypothesize that the listener will not completely generate a new understanding of this new category out of nothing, which is against both literature in developmental psychology and linguistics. Since the language structure human beings have is one that is systematic and have mappings with cognitive concepts and our experiences, we think that a listener, upon hearing a novel category and a familiar feature, will infer a prevalence rate of that novel category given the feature based on her prior experience of the world knowledge that contains sets of entries. This inference is thus conditional on one's prior knowledge, which is why we think that building a model using Bayes Theorem would be appropriate. A listener will search within her repertoire the set of objects that have a prevalence rate that is of certain threshold θ given the feature. Based on Grice's Maxim of Quality, a listener would assume that if a speaker utters the generic containing a novel generic, it would be a piece of information that is informative and valuable. Thus, as our initial results show, listeners would infer the prevalence rate of the novel category and the given feature based on some prior understanding of the comparison set.

While we manipulated the background knowledge in our first experiment, we are aiming at using participants' demographic differences to get at an estimation of prevalence for certain feature and categories that better describe certain probability

terms from the equations below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

We are interested in learning about $P(Feature)$ and $P(Category)$. $P(Feature)$,the probability of certain feature in the world knowledge can be expressed in the following way, where X is a set of all categories that has this $Feature$ as a characteristic feature.

$$P(Feature) = \sum_{x_i \in X} P(Feature|X = x_i)P(X = x_i) \tag{2}$$

$$\sim \sum_{x_j \in X'} P(Feature|X' = x_j)P(X' = x_j) \tag{3}$$

We first make a qualitative prediction that for a listener who hears a generic statement about a novel category and its feature, she would infer that the prevalence rate of this $Feature$ given that novel category as

$$P(Feature|Novel\,Category) \gg P(Feature) \tag{4}$$

$$\gtrapprox P(Feature|X = x_i) \tag{5}$$

where different x_i, considered as comparison sets (background knowledge), are explicitly provided to participants by us during our experiment. Following M.H. Tessler and M.C. Frank, we also hypothesize that a listener would infer that the prevalence rate of this feature given novel category is related and higher than the comparison set they have to compare against this novel generic with.

$$\frac{P(Feature|Category\,x_i)}{\mathbf{E}(\sum_{x_i \in X} P(Feature|X = x_i))} \gg \theta \tag{6}$$

$$\frac{P(Feature|Novel\,Category)}{\mathbf{E}(\sum_{x_i \in X} P(Feature|X = x_i))} \gg \theta \tag{7}$$

where θ is an unknown threshold which the listener thinks that the speaker's surpass for the speaker to utter the generic statement 'Categories are Feature'. At the moment, we are getting $P(Heavy|X = x_i)$ from survey data.

Thus, using the above conditional probability conditions, we derive a speaker and listener model. In the model below, I have chosen one example from our survey as events for the probability calculation so that one can get more contextual intuition on how this model works. In both models, we apply a softmax since the output of the softmax function can be used to represent a categorical distribution, which is what we have here, a probability distribution over different possible feature and category sets.

For a speaker,

$$\begin{aligned} P(utter\,utterance) &= f(P(Friendly|Feps), P(Friendly)) && (8)\\ &= \alpha \frac{P(Friendly|Feps)}{P(Friendly|Feps) + P(Friendly)} && (9)\\ &= \frac{e^{\alpha P(Friendly|Feps)}}{e^{\alpha P(Friendly|Feps)} + e^{\alpha P(Friendly)}} && (10) \end{aligned}$$

For a listener,

$$P(utterance|speaker\,uttered\,''Feps\,are\,friendly'') \tag{11}$$

$$= \frac{P(speaker\,uttered\,utterance|Feps\,are\,friendly)P(Feps\,are\,friendly)}{P(speaker\,uttered\,utterance)} \tag{12}$$

$$= \mathbf{E}(P(speaker\,uttered\,utterance|Feps\,are\,friendly)P(Feps\,are\,friendly)) \tag{13}$$

$$= \mathbf{E}(P(speaker\,uttered\,utterance|Feps\,are friendly)) \cdot 0.5 \tag{14}$$

$$= \mathbf{E}(\int_{P(Friendly|Feps)} \frac{e^{\alpha P(Friendly|Feps)}}{e^{\alpha P(Friendly|Feps)} + e^{\alpha P(Friendly)}} \delta P(Friendly|Feps)) \cdot 0.5 \tag{15}$$

where $P(speakeruttered''Fepsarefriendly'') = 1$ is assumed since, by the time the participant(the listener) does the survey, the event of speaker uttering an utterance already has happened. Further, $P(Fepsarefriendly)$ is drawn from uniform distribution, and thus have its expected value to be 0.5.

METHOD

The table below includes our feature selection, its corresponding comparison categories (chosen based on our estimate of their low, medium, and high prevalence rate) and novel categories.

Feature	Alternative Comparison Categories	Novel Category
friendly	Puppies (H), Goats (M), Squirrels (L)	Feps
tasty	Pizzas (H), Fruits (M), Vegetables (L)	Kobas
heavy	Trucks (H), Stones (M), Bikes (L)	Dands

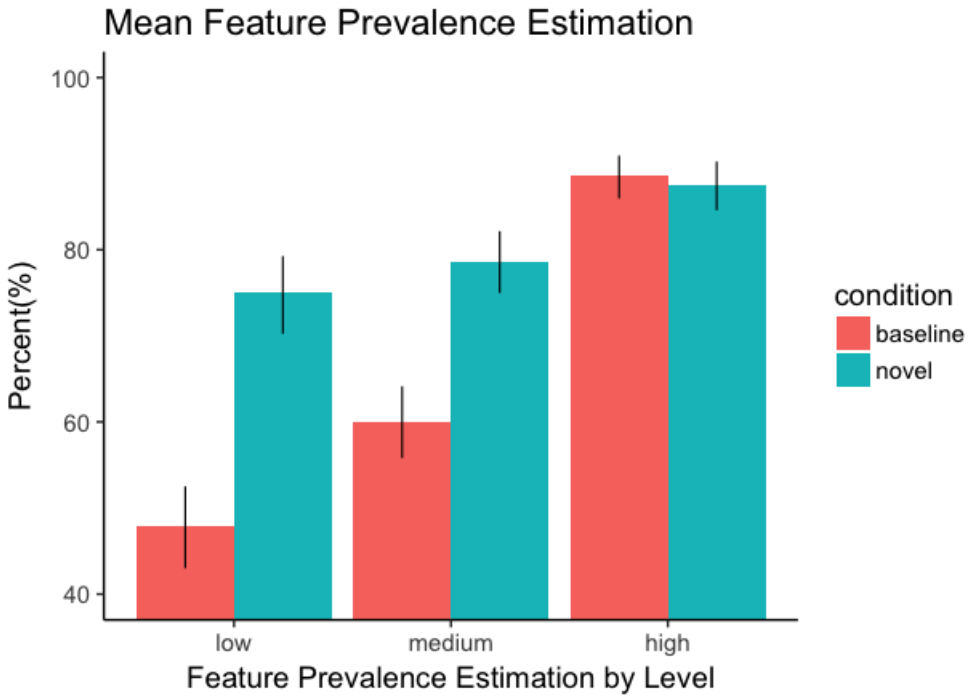
(H = high prevalence, M = medium prevalence, L = low prevalence)

Table 1: Conditions for Survey Questions

We run three separate survey studies on Amazon Mechanical Turk. All three surveys provide participants a narrative that introduces them to an imaginary country, Akar. Sample questions from all three surveys are provided in the section below. The first survey recorded participants's evaluation of a given generic statement as True or False as well as their prevalence rate estimates for all abovementioned 9 categories (3 per feature.) The second survey introduced a novel category C_{novel} along with a familiar comparison category $C_{familiar}$, then asked participants to estimate the prevalence rate of the feature F in novel category C_{novel}. The third survey is similar to the second, with the difference that we only stated 'C_{novel} are like $C_{familiar}$.' before asking participants to estimate the prevalance rate of feature F for C_{novel}. This survey serves as a sanity check, checking whether participants treat C_{novel} as equivalent to $C_{familiar}$ and provide a smiliar rather than higher estimate for C_{novel} comparing to estimates for $C_{familiar}$.

RESULT

The graph below shows a comparison between our baseline prevalence estimation for participants by different features and novel category prevalence estimation for participants. There is a general trend that follows our hypothesis, which is that, provided with a comparison set (our operationalized construct for a listener's background knowledge,) the listener's estimation of a novel category is influenced by this comparison set.



From the table and the graph below, one can see that there is a consistency that confrims our hypothesis that the prevalence rate estimation of a feature and a given novel category is higher than its corresponding baseline familiar category and the same feature. Yet, as one can also see, when level = high, the prevalence estimations are very close to each other. This is not unexpected, since level = high is a condition where people reach more agreement on the high prevalence rate for the given feature and category. In the table below, one can see that we have two major conditions: baseline, novel, and each has three levels: low, medium, and high. This

correspond to the first table shown during the Method sections, where, instead of grouping by different features, we are focusing on the effects between different levels for the baseline (background knowledge) and the novel (newly introduced category.) This table provides some summary statistics on the data.

No.	Condition	Level	Mean	Std. Error	Sample Size
1	baseline	low	47.69655	2.399619	145
2	baseline	medium	59.95172	2.163931	145
3	baseline	high	88.64138	1.283504	145
4	novel	low	74.93407	2.310289	91
5	novel	medium	78.62637	1.883457	91
6	novel	high	87.50549	1.395035	91

Table 2: Data Analysis by Conditions and Levels

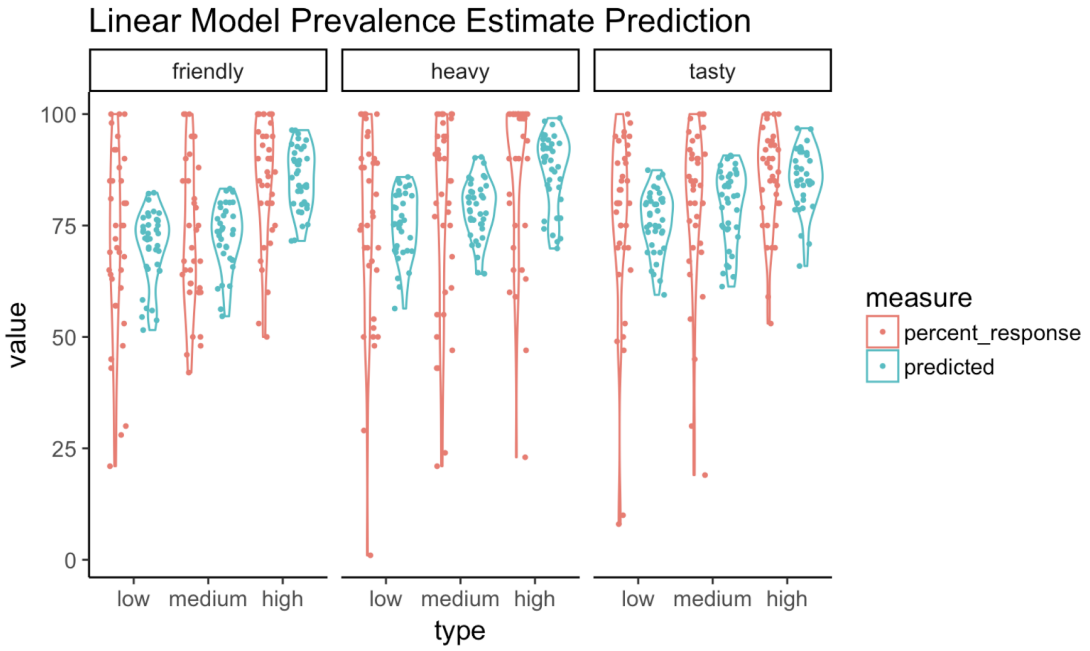
	Estimate	Std. Error	t value
(Intercept)	70.02	2.88	24.31
level - medium	3.11	4.07	0.76
level - high	13.40	4.07	3.29
feature - heavy	4.91	4.07	1.21
feature - tasty	6.41	3.99	1.60
baseline category - fruits	0.24	5.99	0.04
baseline category - goats	0.08	6.03	0.01
baseline category - pizzas	-5.35	6.03	-0.89
baseline category - puppies	1.98	5.99	0.33

Table 3: Summary of Linear Fixed Effect Regression Results

After taking a look at the initial results, we ran a linear fixed effect regression model and the results are shown in Table.3. One can see from the regression result, the level high has a relatively large positive coefficient estimate compared to other variables. The result for our regression does not fully explain the correlation between different levels and the percentage estimation. Although, one could see that level-high does have a higher estimate than level-medium. What is strange here is the baseline category - pizzas, since this category is marked as a level-high category but has a negative estimate of -5.35. Looking at the initial data the graph below, one can

see that there some rather low estimates for the high level categories in participants' response, this may cause the negative estimate.

The graph below shows our prediction using the linear fixed effect model against our data gathered from Amazon Mechanical Turk:



From looking at this graph, one can see that the predictions capture some characteristics of the initial results from MTurk. The initial results have long tails for almost all categories, irregardless of whether the prevalence estimate level it belongs to. The prediction also exhibits this characteristic. Moreover, the prediction captures the increasing of prevalence estimate trend moving from different levels for each feature. However, not all of these results are significant due to the variations in our initial data and possible confounding variables that we did not capture.

DISCUSSION

Results from above revealed a preliminary difference of participants' prevalence estimations on novel categories given different familiar category. This conclusion agrees with our hypothesis and model, suggesting first that the mentioning of a familiar set together with a novel category allows participants to draw associations

between the familiar category and the novel category. Moreover, this association is later used to aid participants to make prevalence estimation on the novel category and a given feature. Participants connect the given feature to the familiar category as well, despite the absence of explicit mentioning of the familiar category and the given feature. Further, one can see that participants, as addressees in speech events, may have assumed Gricean Maxim of Quantity, thinking that the addresser's utterance is informative and thus the participants not only make an estimation that is similar to that estimation for the familiar category, but also overestimates the prevalence rate. The possible speculation behind this overestimation could be that participants believe that if the statement about the novel category is not informative if it has not been a higher prevalence estimation for its paired comparative familiar category.

The rationale behind this project is to explore the possible reasons that people misunderstood each other and where during the transmission of information do messages get construed with certain implication. It is possible that this difference in people's understanding of the same message based on their own background experience has contributed to the generation and transmission of discrimination and stereotyping.

CHALLENGES

There were and still are many challenges to our current project. We suspect there are various interaction variables that might have affected our results. Firstly, while Amazon Mechanical Turk is a useful tool to conduct online surveys, there exist greatly many uncertainties since participants who take the survey are not in one controlled environment, as they would have if they come into the lab to fill the survey. Despite having attention checks and filtered out participants who did not give the correct attention check, there is no direct method to evaluate whether they have answered the questions we asked or have just slid the bar for estimation percentage casually. It is also possible that the orders of questions might have affect on participants answers

even though we have randomized the question orders.

As mentioned in the result section, since people seem to have a tendency to overestimate the prevalence percentage, for the high prevalence rate categories, the difference between the familiar categories and the novel categories are not that significant. This is because there is a cap at a hundred percent. The unique challenge here is that we want to provide generic statements that, if given by an addresser to an addressee, would be generally accepted as a true statement. This requirement is in line with our higher-level assumption that this process of using Bayesian inference occur most saliently when generic statements are accepted as true statements. This limits what we could choose as our categories for given features.

FUTURE DIRECTIONS

In our ongoing and future projects, we aim to study the possible effects of people's background knowledge on their prevalence estimates for novel generic statements using their demographic differences, such as gender, age, political affiliation, and etc. Using these background information, we would be able to avoid explicitly inserting a comparison set to participants and present a result that has more of an correspondence with our society. Furthermore, we would like to design a kid-friendly version of this experiment to test at what age during children language acquisition do they develop similar capacity as adults in making this prevalence estimation judgment. Whether there is a critical period in which children start to develop this inference skill and how do they acquire novel categories to expand their background knowledge.

REFERENCE

1. Roman, Jakobson. "Closing statement: Linguistics and poetics." *Style in language.* (1960): 350-77.

2. Silverstein, Michael. "Shifters, linguistic categories, and cultural description." *Meaning in anthropology.* (1976): 11-55.

3. Ward, Andrew, L. Ross, E. Reed, E. Turiel, and T. Brown. "Naive realism in everyday life: Implications for social conflict and misunderstanding." *Values and knowledge.* (1997): 103-135.

4. Rhodes, Marjorie, Sarah-Jane Leslie, and Christina M. Tworek. "Cultural transmission of social essentialism." *Proceedings of the National Academy of Sciences109, no. 34* (2012): 13526-13531.

5. Taylor, Marianne G., Marjorie Rhodes, and Susan A. Gelman. "Boys will be boys; cows will be cows: Childrens essentialist reasoning about gender categories and animal species." *Child development 80, no. 2* (2009): 461-481.

6. Tessler, Michael Henry, and Noah D. Goodman (Working Paper). "The language of generalization: Probability, vagueness, and interaction." *arXiv preprint arXiv:1608.02926* (2018).