Teachers Know Best: The Impact of Taxonomic Distance and Teacher Competence on Evaluation of Negative Evidence

Chris A. Lawson (lawson2@uwm.edu)
Department of Educational Psychology; UW-Milwaukee; 2400 E. Hartford Ave
Milwaukee, WI 53211

Abstract

Inductive generalization involves extending knowledge from sparse samples of evidence to arrive at broad conclusions. Most of the research in this area has focused on generalization from sparse samples of positive evidence (cases known to share properties with known cases; e.g., birds have hollow bones). Much less is known about generalization from samples of negative evidence (cases known to lack the properties attributed to known cases; e.g., bats do not have hollow bones). This paper reports the results from three experiments that examined factors that were believed to influence adults' evaluation of negative evidence. Experiment 1 showed that when selecting among samples most useful for teaching about a particular category, participants (*N*=36) preferred samples with negative evidence rather than those with single, or additional, positive evidence. Experiment 2 revealed that participants (*N*=25) preferred samples with negative evidence that included a closer (rather than more distant) taxonomic match with the category in question. Finally, Experiment 3 revealed that adults (*N*=52) only preferred samples that provided a close match when evidence was provided by a competent informant. Overall these results emphasize the important role of pragmatic expectations when reasoning about samples that include negative evidence.

Keywords: Generalization; Inductive reasoning; Negative evidence; Pragmatics; Pedagogical sampling

Introduction

Inductive reasoning, the ability to use a single piece of evidence to support generalization, is central to acquiring and utilizing information. Learning that *sparrows* have hollow bones can serve as evidence that other, and perhaps all, birds have hollow bones. A considerable amount of research has focused on the processes that govern how we reason about positive evidence (i.e., new evidence about a category that shares the property in question). Much less is known about the influence of negative evidence (i.e., new evidence about a category that lacks the property in question). To what extent do we consider negative evidence in our inductive decisions?

A small set of studies have identified some conditions in which individuals will prefer to generalize, or endorse, evidence from samples that include negative evidence rather than those that only include positive evidence (Heussen, Voorspoels, Verheyen, Storms, & Hampton, 2011; Kalish & Lawson, 2007; Voorspoels, Navarro, Perfors, Ransom, & Storms, 2015). For example Huessen and colleagues (2011) found that when asked to decide which sample provided the strongest evidence to support a conclusion (e.g., *Swans have enzyme z*), adults judged arguments with premises that included a mixture of positive evidence and negative evidence (e.g., *Ducks have enzyme z* and *Sparrows do not have enzyme z*) as stronger than those with premises that included only a single piece of positive evidence (e.g., *Ducks have enzyme* z) (see also Voorspoels et al., 2015). In a property projection task Kalish and Lawson (2007) found that adults and 5-year-olds preferred to generalize a property to a target from a sample that included negative evidence (e.g., *A raven has enzyme x*, *and a swan does not have enzyme x*) than a sample that included positive evidence (e.g., *A raven has enzyme x*, *and a swan has enzyme x*).

These findings appear to be at odds with a well-known paradox in inductive logic. Formal logic dictates that, via contraposition, any piece of evidence that qualifies as not sharing both the category identity and the properties of a premise serves as support for a premise. In this formulation, the existence of a red racecar (a non-black, non-raven) serves as evidence to support the assertion that *All ravens are black* (Hempel, 1945). Further, the idea that negative evidence supports induction is inconsistent with models of inductive reasoning that emphasize that adding positive evidence strengthens induction (e.g., monotonicity; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). Thus, prior work on induction would seem to imply that the addition of negative evidence would weaken our inductive decisions (see e.g., Heussen et al. 2011).

Among other things, these views on negative evidence fail to consider the psychological influence of prior beliefs, and expectations, on our inductive decisions. Consider that most inductive reasoning tasks take place within a pedagogical context, in which a person (e.g., informant or teacher) provides evidence from which they are soliciting an inductive response. These conditions are subject to pragmatic rules of communication, such as those that specify that informants ought to provide relevant information (Grice, 1975; Sperber & Wilson, 1986). Findings from a range of studies indicate that because we expect informants have deliberately chosen the samples of evidence they present we will identify relations between properties and categories within the evidence that we believe are relevant to the task (e.g., Medin, Coley, Storms, & Hayes, 2003).

Expectations about the intentions of informants also guide reasoners to consider the processes by which evidence was

selected (Navarro, Dry, & Lee, 2012; Tenenbaum & Griffiths, 2001). Reasoning patterns vary depending on whether evidence was described as having been selected randomly without consideration of the property or category in question (i.e., weak sampling) rather than if the evidence was selected deliberately with respect to a particular category or property (i.e., strong sampling) (see e.g., Lawson & Kalish, 2009; Navarro et al., 2012). In the former case our expectations about the unbiased intentions of the informant signal to us that any regularities within the sample are suspiciously coincidental and therefore warrant a stronger set of inferences than the latter case for which sampling would be viewed as intentionally selective (e.g., Xu & Tenenbaum, 2007). Drawing from this work Voorspoels and colleagues (2015) found that participants preferred to generalize from mixed samples of evidence (positive and negative evidence) rather than positive samples of evidence, when a cover story described that the evidence was selected deliberately.

The studies reported in this paper were designed in consideration of these issues. One of the goals was to explore whether subtle cues about the deliberate selection of evidence would be sufficient to solicit greater attention to negative evidence. Rather than describing the methods by which evidence was selected, in these studies the evidence was merely described as having been selected by a "teacher". Labeling someone as a teacher should emphasize their role as an individual who provides relevant information, and therefore under these circumstances participants were expected to prefer samples with negative evidence rather than samples with positive evidence.

Another goal of the present studies was to explore the extent to which different types of negative evidence would support inductive decisions. Kalish and Lawson (2007) showed that negative cases that implicated a contrast at a close level of abstraction provided stronger support for induction than those that highlighted a more superordinate contrast (for similar results see Lee, Lovibond, Hayes, & Navarro, 2019; Shafto, Goodman, & Griffiths, 2014; Voorspoels et al., 2015). When considering the hypothesis about black ravens, evidence about red racecar might not be useful, yet evidence about white swans might be. What are some factors that make some types of negative more useful? The current studies tested sensitivity to different types of negative cases by manipulating the taxonomic distance between the concept in question and the negative evidence.

Finally, it is important to note that while prior research on negative evidence on inductive generalization has evaluated judgments of argument strength or property induction, the present studies used a slightly different method. In the present studies participants were presented with scenarios in which teachers were described as trying to help their students learn new concepts by using different samples (i.e., evidence). The samples were manipulated to include different types of evidence (e.g., negative vs. positive). Participants were asked to evaluate the evidence provided by the teachers to determine which had done the best job helping their students learn about the concept in question.

Experiment 1

In Experiment 1 participants were provided two sets of evidence each of which was said to have been supplied by different teachers. The items were designed such that the different teachers presented conflicting sets of evidence. A teacher either provided a set of evidence with a positive single case (e.g., dogs have omat bones), positive evidence about two cases (e.g., dogs have omat bones and cats have omat bones), or evidence with a positive case and a negative case (e.g., dogs have omat bones and cats do not have omat bones). Thus, the structure of the task was similar to Kalish and Lawson (2007). However, there were some differences. First, this study included a range of domains and solicited judgments about categories (e.g., "birds"), rather than individuals (e.g, "this bird"). Second, rather than being asked to make an inductive projection, in this task participants were asked to determine which sample was the most helpful for learning about a concept. The prediction was that participants would favor the samples that included negative evidence and therefore would select those samples over the other two.

Method

Participants. Thirty-six adults participated in this experiment. Participants were recruited from introductory Psychology courses and received course credit for their participation. Participants were sampled from a large eastern US city. There were approximately equal numbers of males and females.

Design. Participants responded to 15 items. Each item included two samples both of which were introduced in the context of two teachers trying to decide which clues were better for helping their students learn about a topic. Overall there were five examples from each of three sample pairings: *single* vs. *positive*, *single* vs. *negative*, and *positive* vs. *negative*. Each sample included evidence about different categories represented by items drawn from the basic level of abstraction (e.g., Rosch et al. 1975). The specific categories presented in the evidence varied across the three samples. All samples included evidence about the category that was the focus of the teaching lesson (e.g., bears). The negative and positive samples always included a category represented by an item that came from the same superordinate category. For example, for the item in which the teachers were trying to help their students learn about bears, in the *negative* vs. *positive* pairing a participant may have heard about Teacher A, who told her students, "Bears eat flaxum; Birds do not eat flaxum" and Teacher B who told her students, "Bears eat flaxum; Birds eat flaxum." For the *single item* from this set the Teacher told her students that, "Bears eat flaxum".

After presentation of the two samples participants were asked to judge which teacher had provided the best examples to help their students learn about a specific category. The specified goal of both teachers was to teach about the first premise that was introduced. For example, continuing from the above example, participants were asked, "Which teacher provided better clues to help her students learn about bears?"

Procedure. Participants were interviewed in a quiet location in a laboratory on their campus. All materials were presented on a laptop computer. Participants were told that they would read some hypothetical scenarios in which some teachers were trying to find the best way to teach their students about different things. The instructions were as follows:

> You are going to read scenarios in which different teachers are trying to help their students learn about the same concept, but each has provided different clues, or facts, to help them learn. Your task is to decide which teacher has given their students the best clues to help them learn.

For each item participants read that both teachers were interested in teaching their students about a certain topic (e.g., bears). The sample provided by each teacher was then randomly presented. For example, in the *single vs. negative* evidence case, participants might be told,

> Two teachers are interested in teaching their students about BEARS. Teacher A tells her students that bears have funti blood. Teacher B tells her students that bears have funti blood and that lizards do not have funti blood. Which teacher do you think has given her students better clues to help them learn about bears?

After each response a new item was then presented. The task lasted approximately 10 minutes.

Results and Discussion

The primary analysis considered whether participants favored one evidence type more in each of the evidence pairs. Separate comparisons revealed that participants consistently selected samples of negative evidence whether the alternative choice was a single exemplar (73% vs. 27%), $t(24)=3.94$, $p=.001$ (two-tailed), $d=1.35$, or a sample that included positive evidence (74% vs. 26%), $t(24)=3.84$, $p=.001$, $d=1.26$. When single evidence was pitted against positive evidence, there was no difference in the proportion of choices for each evidence type (49% vs. 51%), $t<1$, *ns.*

These results support the prediction that participants would prefer the sample with negative evidence over the other two samples. Thus these results are consistent with findings from the inductive reasoning literature showing that people prefer to generalize from samples that included negative evidence (Heussen et al., 2015; Kalish & Lawson, 2007; Lee et al., 2019). The most interesting comparison is between the positive evidence and the negative evidence, in which the teachers both presented evidence about the same categories. One interpretation is that participants understood that the negative evidence was intentionally selected as a contrast, to highlight a property that was exclusive to the concept in question. However, it is also possible that participants merely responded to the technique used by the teacher: participants might expect that teachers are prone to use negative evidence as a way to highlight meaningful information. Thus, when asked to judge who has chosen better clues, participants may have simply chosen the teacher that used negative evidence. One of the goals of Experiment 2 was to examine this possibility.

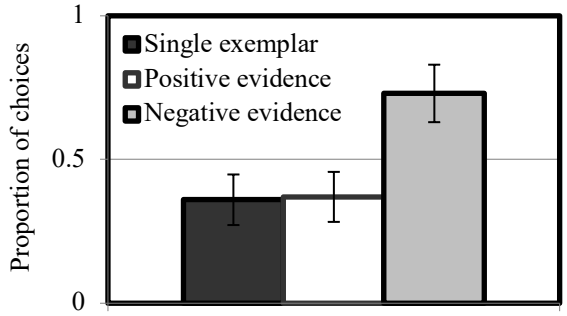

Figure 1. Proportion of choices in which participants selected each evidence type when it was among the evidence pairs. Note that bars represent 1 +/- SE from the mean.

Experiment 2

Experiment 2 was designed with several goals in mind. One of the goals was to explore the effect of category "distance" on the evaluation of negative samples. Negative evidence is useful to the extent that it can be viewed as providing a useful contrast to the concept in question (Kalish & Lawson, 2007). In hoping to teach students about Bears, evidence about other mammals that lack the critical properties provides the type of contrast that highlights the property generalizes to (only) bears. Without providing an explicit contrast, positive evidence is more likely to broaden rather than narrow generalization (though see Gentner & Namy, 2006). Learning about another mammal that shares a property with a bear might suggest the property generalizes broadly, to other mammals. The goal of this study was to determine the extent to which the relative taxonomic distance between evidence and target concepts influences how adults reason about positive and negative samples of evidence.

Another goal of this experiment was to determine whether participants merely prefer an informant who provides negative evidence, regardless of the contents of the sample. This issue was examined by asking participants to select between samples that included negative evidence with exemplars that were taxonomically close to the category in question with samples that were taxonomically far to the category in question.

Method

Participants. Twenty-five adults participated in this experiment. Participants were recruited from introductory Psychology courses and received course credit for their participation. Participants were sampled from a large eastern US city. There were roughly equal numbers of males and females.

Materials, Procedures, & Design

Participants responded to sixteen items. The method was similar to Experiment 1 in that participants were presented with a scenario in which two teachers were described as having presented competing examples and participants were asked to judge which teacher provided the most compelling examples to help their students learn. The primary exception was that in Experiment 2 the examples provided by the teachers varied in their taxonomic distance relative to the category in question; one of the teachers presented information about a Close taxonomic match and the other teacher presented information about a Far taxonomic match. The Close match was always a member from the same subordinate as the category in question, whereas the Far match was from a different basic level. For example, for one item participants were told that teachers using examples to teach their students about *grizzly bears*, and that one teacher presented additional information about *polar bears* (Close) and that the other teacher presented additional informational about *deer* (Far). The items were the same as those used in Experiment 1, with exceptions to the aforementioned modifications.

Evidence type (negative, positive) was manipulated within subjects in such a way to create four distinct pairs of evidence pair contrasts: Close-Negative (CN) vs. Far-Negative (FN), CN vs. Far-Positive (FP), Close-Positive (CP) vs. FN, and CP vs. FP. In all other respects the study design was the same as in Experiment 1.

Results and Discussion

A repeated measures ANOVA, with evidence type (negative, positive) and taxonomic distance (Close, Far) serving as within-subjects variables, revealed there was significant effect of taxonomic distance, $F(1,24)=26.30$, $p<.001$, $\eta^2=.52$. As suggested by Figure 2, participants made a greater proportion of choices of the Close matches than they did Far matches, all ps<.01 (Tukey's HSD). No other main effects or interactions were significant.

Further analyses revealed that participants exhibited a significantly greater preference for Close-Negative (CN) samples over both types of Far samples: CN vs. FN, 74% vs. 26%. $t(24)=4.66$, $p<.001$, and CN vs. FP, 76% vs. 24%, $t(24)=3.78$, $p<.001$ (two-tailed), both ds>1.25. In both cases, the number of participants who preferred CN evidence over both types of Far samples (17 out of 25 cases for both options) was greater than would be expected by chance, $p=.03$, binomial theorem. In contrast, there was not a significantly greater preference for CP samples over either Far samples, both ts<1.50, ps>.23.

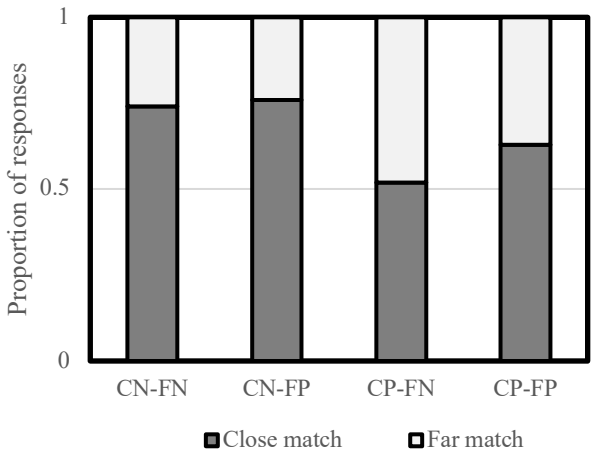

Figure 2. Proportion of choices in which participants selected either the teacher who presented the Close match or the teacher who presented the Far match for each of the four evidence pairs; Close-Negative (CN) vs. Far-Negative (FN), CN vs. Far-Positive (FP), Close-Positive (CP) vs. FN, and CP vs. FP.

The results support the prediction that negative evidence that represents a target in close taxonomic proximity to the category in question would be favored over negative evidence that represents a target that is less close in proximity. The former provides a better contrast for the category in question than that latter. Additionally, results from evidence pairs in which the close cases were represented by positive evidence suggest that the taxonomic relationship between target and evidence alone cannot account for these effects. Additionally, the overall response pattern indicate that participants do not simply prefer negative samples over positive samples, but rather prefer samples that establish a contrast that highlights the relevant category in question.

Experiment 3

To this point the results are consistent with the idea that negative evidence is useful when we expect it has been chosen to help us learn about a particular concept. Such an expectation hinges on the belief that an informant is a capable and willing partner. Experiments 1 and 2 suggest that mentioning an informant was a teacher is sufficient to draw attention to the relevant contrast implied by the negative evidence. The final experiment considered the degree to which information about the effectiveness of a teacher impacts how we reason about the evidence they provide. If participants rely on their prior beliefs about the competence (not just the intentions) of teachers, they should prefer negative evidence that presents a contrast that highlights the category in question when the teacher is described as effective, rather than when the teacher is described as ineffective.

Method

Participants. Fifty-two adults participated in this experiment. Participants were recruited from introductory Psychology

courses and received course credit for their participation. Participants were sampled from a moderately large eastern US city. There were roughly equal numbers of males and females.

Materials, Procedures, & Design

The method employed here was similar to the methods used in the experiments reported above with a few exceptions. First, in this task participants were told about a single teacher who was trying to teach their students about a concept. Participants were then asked to determine which additional piece of information the teacher would provide as an example to help their students learn the concept. Second, evidence type (negative, positive) was manipulated between subjects. Finally, two additional between-subjects variables were included. The Teacher competency variable was manipulated with a cover story about the effectiveness of the teacher.

> *Effective teacher* condition (*N*=25): "Mrs. Johnson is a really effective teacher. She always presents material in a way that makes the content clear. Her students learn quite a bit from her and she has won many awards for her teaching."
>
> *Ineffective teacher condition (N=27): "*Mrs. Johnson is a really ineffective teacher. She always presents material in a way that makes the content unclear. Her students learn very little for her and she is often criticized for her teaching"

Participants were then presented 15 items. For each item the teacher was described as giving a lesson on a particular topic in which she provided a fact about a concept (e.g., "Mrs. Johnson was teaching a lesson on Trout and she told her students that Trout have the neurotransmitter glibon"). Participants were asked to judge which other facts she might choose to help them learn about the concept. The additional facts were represented by three different targets all of which were described as lacking the property that was attributed to the category in question (e.g., x's DO NOT have the neurotransmitter glibon). The targets included two items from the same basic-level as the category in question (e.g., other fish), one of which was a typical member (e.g., bass) and another which was an atypical member (e.g., flounder). The third item represented an item from a different category within the superordinate (e.g., a frog is an animal, but not a fish). These items were adapted from the first two experiments.

Results

The analyses considered the proportion of responses for each of the three targets in each of the Teacher conditions (Figure 3). A mixed ANOVA (Target Type X Teacher reputation) revealed a main effect of Target, which was conditioned by an interaction with Teacher effectiveness, $F(2,49)=9.76$, $p=.002$, $\eta^2=.16$. This interaction was due to differences in responses to the Typical and Superordinate targets. Participants were significantly more likely to select the Typical targets in the Effective teacher condition than the Ineffective teacher condition, $F(1,50)=10.48$, $p=.002$, whereas the opposite pattern (Ineffective > Effective) emerged for the Superordinate targets, $F(1,50)=9.80$, $p=.003$.

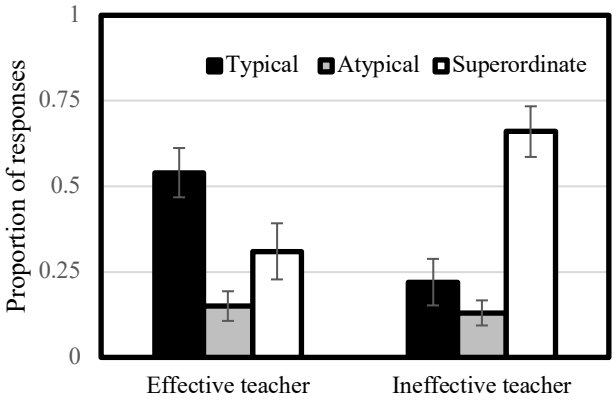

Figure 3. Proportion of responses for each of the three targets in each of the evidence conditions. Note that bars represent 1 +/- SE from the mean.

Additional analyses of individual patterns confirmed these group patterns. In the Ineffective teacher condition a significant number of participants (18 out of 27) preferred the superordinate targets, $p<.001$ (binomial theorem, assuming an equal likelihood of choosing one of the three targets). In contrast, in the Effective teacher a significant number of participants (16 out of 25) preferred the Typical targets, $p=.001$ (binomial theorem).

These results indicate that information regarding competency and past effectiveness of a teacher influenced the type of negative evidence they were believed to have selected. Those described as effective teachers were expected to have chosen the negative evidence that provided the closest taxonomic match, and therefore the more relevant contrast, to the category in question. In contrast, those described as ineffective teachers were expected to have provided the most distant taxonomic match, and therefore the least relevant contrast to the category in question. Overall, these results are consistent with the other findings here in showing that participants rely on information about an informant when evaluating the generalizability of samples that include negative evidence.

General Discussion

How people use evidence to arrive at decisions is a central question in the study of human reasoning. Most of the research on this topic has focused on reasoning about positive evidence. Much less is known about the influence of negative evidence. To what extent, if any, does learning about cases that lack the property known to be true of the category in question impact our decisions?

Using a novel method in which participants were asked to choose among different samples provided by teachers, the results from three studies support the conclusion that negative evidence has a significant impact on inductive decisions.

Experiment 1 showed that participants preferred samples with negative evidence over those with either single evidence or positive evidence to support learning about a category in question. Experiments 2 and 3 revealed that participants consider the contrast established by the negative evidence and the reputation of the informant when determining which sample of negative evidence supports generalization. Overall, these results support the view that adults rely on pragmatic considerations when evaluating negative evidence.

These results are consistent with prior work that has shown that negative evidence facilitates inductive decisions. For example, similar to Voorspoels et al. (2015), these results indicate that the impact of negative evidence depends on participants' assessment of pragmatic task features. While their study showed that the impact of negative evidence was influenced by information about sampling procedures, the present studies indicate that evidence about the status and competency of an informant guides the determination about samples of negative evidence. Overall, both sets of findings confirm that pragmatic factors, such as the expectation that informants are deliberate and intentional in the selection of evidence, render negative evidence useful for generalization.

Drawing from this perspective, the results are also consistent with the idea that negative evidence that establishes a contrast that highlights the category in question is viewed as especially relevant for generalization (see also Lee et al., 2019; Kalish & Lawson 2007). These findings indicate that under conditions in which the evidence under consideration is meant to apply to a particular category (e.g., bears) a reasoner will favor negative evidence about a category member at the level of abstraction closest to the category in question (e.g., deer). In such cases negative evidence can be viewed as a meaningful contrast; presumably chosen to underscore that a property is to-be-generalized to the category in question (e.g., Clark, 1990; Nordmeyer & Frank, *ms*.).

These findings appear to be at odds with other work that indicates that negative evidence is likely to either weaken (e.g., Osherson et al., 1990) or have no meaningful impact (e.g., Hempel, 1945) on inductive judgements. Both of these accounts are accurate under conditions in which the available evidence is sampled randomly from an infinite pool of cases. However, in pedagogical contexts evidence is often chosen by an informant whose intentions and competence become important matters to consider when evaluating the evidence they have provided. The present studies demonstrate that in such cases negative evidence, especially in cases in which it presents a relevant contrast with the category in questions, is a strong cue from which to generalize.

References

Clark, E.V. (1990). On the pragmatics of contrast. *Journal of Child Language, 17*, 417-431.

Gentner, D., & Namy, L.L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science, 15*, 297-301.

Hempel, C. (1945). Studies in the Logic of Confirmation. *Mind, 5*, 1-26.

Heussen, D., Voorspoels, W., Verheyen, S., Storms, G., & Hampton, J.A. (2011). Raising argument strength using negative evidence: A constraint on models of induction. *Memory & Cognition, 39*, 1496-1507.

Kalish, C. W., & Lawson, C. A. (2007). Negative evidence and inductive generalization. *Thinking and Reasoning*, 13, 394-425.

Lawson, C.A., & Kalish, C.W. (2009). Sample selection and inductive generalization. *Memory & Cognition, 37,* 596-607.

Lee, J.C., Lovibond, P.F., Hayes, B.K., & Navarro, D.J. (2019). Negative evidence and inductive reasoning in generalization of associative learning. *Journal of Experimental Psychology: General, 148*, 289-303.

Medin, D.L., Coley, J.D., Storms, G., & Hayes, B.L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10*, 517-532.

Navarro, D.J., Dry, M.J., & Lee, M.D., (2012). Sampling assumptions in inductive generalization. *Cognitive Science, 36,* 187-223.

Nordmeyer, A.E., & Frank, M.C. (unpublished ms.). *Negation is only hard to process when it is pragmatically infelicitous.*

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,*185–200.

Rosch, E., Mervis, C.B., Gray, W., Johnson, D., Boyes-Braehm, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

Shafto, P., Goodman, N.D., & Griffiths, T.L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*, 55-89.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.

Tenenbaum, J.B., & Griffiths, T.L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629-641.

Voorspoels, W., Navarro, D.J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology, 81*, 1-25.

Xu, F., & Tenenbaum, J.B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.