Journal of Experimental Psychology:
Learning, Memory, and Cognition
2010, Vol. 36, No. 4, 906–919

© 2010 American Psychological Association
0278-7393/10/$12.00 DOI: 10.1037/a0019762

The Relevance Framework for Category-Based Induction: Evidence From Garden-Path Arguments

Aidan Feeney
Queen's University Belfast

John D. Coley
Northeastern University

Aimée Crisp
Durham University

Relevance theory (Sperber & Wilson, 1995) suggests that people expend cognitive effort when processing information in proportion to the cognitive effects to be gained from doing so. This theory has been used to explain how people apply their knowledge appropriately when evaluating category-based inductive arguments (Medin, Coley, Storms, & Hayes, 2003). In such arguments, people are told that a property is true of premise categories and are asked to evaluate the likelihood that it is also true of conclusion categories. According to the relevance framework, reasoners generate hypotheses about the relevant relation between the categories in the argument. We reasoned that premises inconsistent with early hypotheses about the relevant relation would have greater effects than consistent premises. We designed three premise garden-path arguments where the same 3rd premise was either consistent or inconsistent with likely hypotheses about the relevant relation. In Experiments 1 and 2, we showed that effort expended processing consistent premises (measured via reading times) was significantly less than effort expended on inconsistent premises. In Experiment 2 and 3, we demonstrated a direct relation between cognitive effect and cognitive effort. For garden-path arguments, belief change given inconsistent 3rd premises was significantly correlated with Premise 3 (Experiment 3) and conclusion (Experiments 2 and 3) reading times. For consistent arguments, the correlation between belief change and reading times did not approach significance. These results support the relevance framework for induction but are difficult to accommodate under other approaches.

Keywords: inductive reasoning, relevance, category-based induction, reading times, belief change

An important everyday cognitive ability involves evaluating the strength of inductive arguments. For instance, based on one's knowledge of certain relationships between political parties in the United Kingdom and the United States, an individual might be asked to infer that members of the Conservative party will tend to hold some policy on tax because he or she learns that members of the Republican party do. Of course such an argument is inductive because its conclusion does not necessarily follow from its premises; although they share many beliefs about the economy, Republicans and Conservatives may diverge on the issue of taxation. Often, as in the example we have just described, inductive arguments are made on the basis of category membership: Given that members of one category possess a property, individuals assess the likelihood that members of another category share that property. There is now a range of theoretical accounts of how people perform this kind of reasoning (e.g., Heit, 1998; McDonald, Samuels, & Rispoli, 1996; Osherson, Smith, Wilkie, Lopes, & Shafir, 1990; Tenenbaum, Kemp, & Shafto, 2007; for an overview see Heit, 2007). In this article, we describe an entirely new method we have used to test the relevance account (Medin et al., 2003) of how people evaluate such category-based arguments.

We are interested in the relevance account because, as we will show, it leads to processing-level predictions about category-based reasoning. Although existing accounts of induction make predictions about the types of processes involved in inductive reasoning (for a discussion, see Feeney, 2007), only the relevance account allows one to make predictions about processing time for premises of inductive arguments on the basis of their content. Our overall goal in this article is to derive and test such processing time predictions, and thus to demonstrate how the study of inductive reasoning may be informed by dependent variables other than argument strength, which up until now has been the primary focus of experimental and modeling studies.

Aidan Feeney, School of Psychology, Queen's University Belfast, Northern Ireland, United Kingdom; John D. Coley, Department of Psychology, Northeastern University; Aimée K. Crisp, Department of Psychology, Durham University, Durham, United Kingdom.

Experiments 1 and 2 were carried out while Aimée K. Crisp was in receipt of an undergraduate research bursary from the Nuffield Foundation. Aimée K. Crisp is currently funded by a postgraduate award from the Economic and Social Research Council. We thank Bob Metcalf and Darren Dunning for their assistance with programming and data collection in the United Kingdom; Amanda Civiletto for her work in programming and data collection in the United States; and Neal Pearlmutter, Jean-Baptiste Van der Henst, and Anna Vitkin for their useful input in the early stages of this research.

Correspondence concerning this article should be addressed to Aidan Feeney, School of Psychology, Queen's University Belfast, Northern Ireland BT7 1NN, United Kingdom. E-mail: a.feeney@qub.ac.uk

Category-Based Inductive Reasoning and Similarity

When category-based inductive reasoning is studied in the laboratory, arguments of the following kind, where sentences above the line are premises and the sentence below the line is a conclusion, are often used.

Argument 1:

Cows have Property P

Sheep have Property P

Argument 2:

Cows have Property P
Sheep have Property P

All animals have Property P

Argument 3:

Cows have Property P
Chipmunks have Property P

All animals have Property P

All three of these examples illustrate the importance of similarity to the evaluation of category-based inductive arguments. Most people judge Argument 1 to be strong based on the similarity between cows in the premise and sheep in the conclusion. However, most people also judge Argument 2 to be weaker than Argument 3. Because cows and sheep are more similar than cows and chipmunks, they cover the conclusion category animals less well than do cows and chipmunks. This effect is known as the diversity effect (see Heit & Feeney, 2005; Heit, Hahn, & Feeney, 2005). There are a variety of accounts for category-based induction, which explain a range of phenomena such as the diversity effect in terms of similarity relations (see Osherson et al., 1990; Sloman, 1993).

More recently, problems have begun to emerge for similarity-based accounts of how people evaluate category-based inductive arguments. One problem for similarity-based approaches is that there may be many different kinds of similarity relations between the categories in an argument (see Medin, Goldstone, & Gentner, 1993). Whales and fish, for example, are behaviorally similar but taxonomically dissimilar. Heit and Rubenstein (1994) have demonstrated that people think it more likely that taxonomically similar categories share biological properties than it is that behaviorally similar categories share them, and that it is more likely that behaviorally similar categories share behavioral features than it is that taxonomically similar categories do. Similar effects have been reported by Murphy and Ross (1999) in the domain of food categories, and property effects have been observed in young children (see Coley, Vitkin, Seaton, & Yopchick, 2005).

Apart from the question of which kind of similarity relation should be considered most relevant in a particular context, there have been several demonstrations that knowledge about relations other than similarity can affect the perceived strength of an argument. For example, cross-cultural work (López, Atran, Coley, Medin, & Smith, 1997) has shown that North American participants display sensitivity to the diversity effect because they use their knowledge of similarity relations when evaluating two-premise arguments. However, Central American participants (Itza' Mayans who live in Guatemala) use their knowledge of causal and ecological relations to evaluate such arguments and hence do not show sensitivity to diversity (or at least to diversity computed over taxonomic categories).

Analogous effects of expertise have been shown in studies of inductive reasoning. For instance, among the experts studied by Proffitt, Coley, and Medin (2000), only taxonomists used their knowledge of taxonomic relations among plants to evaluate a series of plant category-based inductive arguments. Landscape gardeners and maintenance workers relied on other knowledge. Other studies of experts demonstrate the influence of the property to be projected. Shafto and Coley (2003) found that expert fishermen were as likely as North American undergraduates to use knowledge of taxonomic relations when evaluating arguments about fish where the property to be evaluated was blank. When the property concerned transmission of a disease, however, the fishermen, but not the undergraduates, reasoned on the basis of ecological food–web relations. In sum, different kinds of similarity, as well as causal relations that fall outside the realm of similarity altogether, contribute to category-based induction.

The Relevance Framework

In response to these and other demonstrations (for a recent review, see Shafto, Coley, & Vitkin, 2007), Medin and colleagues (2003) have formulated a relevance framework for category-based induction. Because Medin et al. (2003) applied principles derived from relevance theory to category-based induction, they refer to their account as a framework rather than as a theory, and we will persist with their nomenclature here. The relevance theory of linguistic pragmatics (Sperber & Wilson, 1995) holds that relevance, as a property of inputs to the cognitive system, is determined by the cognitive effects derived from processing those inputs, and the cognitive effort required to achieve those effects. Cognitive effects include belief change and the combination of new information in the input with preexisting information so that contextual conclusions maybe drawn. Cognitive effort is inversely related to relevance. That is, inputs that require more effort to process are less relevant. Thus, information that is salient, or comes to mind easily, is often highly relevant.

In applying these principles to category-based induction Medin et al. (2003) argued that premises are assumed by participants to be relevant to the conclusion. In other words, participants assume the experimenter to be cooperative (Grice, 1975). Further, Medin et al. (2003) argued that because of the principle of cognitive effort, salient properties of the categories in the argument are highly relevant. For example, in processing the premise *Magpies have Property P*, salient properties of magpies are likely to come to mind, such as the color of their plumage. Medin et al. (2003) argued that because participants assume the experimenter to be cooperative, participants are likely to associate such properties with the blank property in the argument. Participants will also compare the properties of the premise and conclusion categories in order to test whether the salient property is plausible. So, the conclusion *Zebras have Property P* is further evidence that the blank property has to do with having black and white coloration.

This same comparison process is used when participants evaluate multiple-premise arguments. So, had the statement about zebras occurred as a second premise in an argument rather than as the conclusion, it would still have provided supporting evidence for the hypothesis that the blank property has to do with coloration.

To test their account, Medin et al. (2003) contrasted taxonomic (or similarity) relations with highly salient ecological (or other nontaxonomic) relations. That is, they created experimental materials where the cognitive effort required to process the ecological relation was less than the effort required to process the taxonomic relation. For instance, participants were asked to rate the strength of arguments such as 4 and 5 below.

Argument 4:

Penguins have Property X13
Eagles have Property X13

Camels have Property X13

Argument 5:

Penguins have Property X13
Polar bears have Property X13

Camels have Property X13

These arguments were constructed so that the most diverse pair of categories also shared a property that was not shared by the conclusion category. Accordingly, Medin et al. (2003) predicted, and observed, a reversal of the diversity effect. They also predicted that salient causal relations should affect judgments of argument strength. For example, they showed that adding a category to the conclusion (*Grass has some property therefore humans have that property* vs. *Grass has some property therefore cows and humans have that property*) increased the perceived strength of the argument. Medin et al. (2003) predicted this finding on the grounds that adding the extra conclusion category would make the causal relations between the categories in the argument more salient. Although some of these effects (in particular the nondiversity effect) have turned out to be less straightforward than originally appeared (see Heit & Feeney, 2005), others have been successfully replicated and extended (see Feeney, Shafto, & Dunning, 2007).

One feature of Medin et al.'s (2003) demonstrations is that they all work by manipulating cognitive effort. That is, a nontaxonomic relation or feature is more available to the reasoner (see Shafto et al., 2007) than is the taxonomic relation. In the experiments to be described here, we extend and test the relevance theory by considering sources of cognitive effect in the category-based induction task. Consideration of sources of cognitive effort and cognitive effects permit us to derive processing time predictions for category-based inductive reasoning.

Cognitive Effects and Garden-Path Arguments

One obvious and important cognitive effect in the category-based induction paradigm is a change to the participant's belief about the strength of the argument. Thus, if a premise is likely to substantially increase or decrease belief in the conclusion, that premise is highly relevant. However, another important cognitive effect of an input would be if it changed the participant's beliefs about the communicative intentions of the experimenter in choosing the premises in the argument. For example, a premise might make her less confident in a particular hypothesis that she had derived from earlier premises, about the property that the experimenter wished her to think of. When she reads the second premise about polar bears in Argument 5 above, the hypothesis derived from the first premise, that the experimenter intended her to think that the blank property had to do with membership of the category bird, is weakened. On the other hand, reading the second premise in Argument 4 (that eagles possess the property) strengthens the bird hypothesis.

In order to test the relevance framework by studying the cognitive effort expended in processing supportive and nonsupportive evidence, we developed a new category-based inductive reasoning task, based on garden-path arguments. Consider, for example, Argument 6.

Argument 6:

Brown bears have Property K
Panda bears have Property K
Zebra have Property K

All animals have Property K

We call Argument 6 a garden-path argument because upon reading the first two premises of the argument—according to the relevance account—the reasoner is likely to hypothesize that the experimenter intends to communicate that the blank property has something to do with membership of the near superordinate category *bear*. However, the reasoner has been led down a garden path, as this hypothesis is not supported by the third premise. Accordingly, the cognitive effects of reading the third premise should be high. Because according to relevance theory people will expend cognitive effort in processing an input in proportion to the cognitive effects to be derived from doing so, the theory predicts that participants should spend a relatively large amount of time in processing the third premise.

To emphasize this prediction, we consider Argument 7.

Argument 7:

Magpies have Property Z
Panda bears have Property Z
Zebras have Property Z

All animals have Property Z

Here, the first two premises suggest that the experimenter intends to convey that the blank property has to do with black-and-white coloration, and this hypothesis is supported by the third premise. Accordingly, Argument 7 is not a garden-path argument and its third premise does not achieve a high degree of relevance owing to its cognitive effects. Instead, it merely further supports the reasoner's preexisting hypothesis. Accordingly, we did not expect participants to spend a lot of time processing it.

Using arguments of this type, we also tested the similarity-based theories by examining participants' ratings of argument strength. In this case, the premises in Argument 7 are more taxonomically

diverse (one bird, two mammals), and hence provide better coverage of the conclusion category *animals* than the premises of Argument 6 (three mammals); thus, similarity-based approaches predict that Argument 7 should be seen as stronger than Argument 6. In contrast, the relevance framework suggests that salient relations among premise but not conclusion categories may weaken or even reverse standard diversity effects. For example, despite the diversity of its premises, Argument 7 might not be seen as stronger than Argument 6 because the single salient property present in all premises of Argument 7 (black-and-white coloration) is not present in the conclusion; this creates a mismatch between what is seen as relevant for premises and for conclusion, and thus weakens the argument. In all three of the experiments that follow, we used a variety of garden-path and non-garden-path arguments to test predictions about relations between cognitive effort—as measured by reading–response time—and cognitive effect—as measured by perceived argument strength. In general, the relevance framework predicts that premises with larger cognitive effect should be processed more deeply (i.e., require more effort).

Experiment 1

In Experiment 1, we sought to test our basic prediction that premises with larger cognitive effect will be processed more deeply—that is, deemed more relevant—than premises with little cognitive effect. Specifically, we proposed that while reading premises of a multipremise argument, participants would form hypotheses about the nature of the property consistent with salient relations among premise categories (see also McDonald et al., 1996); premises with greater cognitive effect (i.e., those that suggest a new or different hypothesis) should be processed more deeply and thus yield higher reading times than those with less cognitive effect (i.e., those that confirm an existing hypothesis).

To test this idea, we developed sets of related three-premise arguments (see Table 1 for an example). Each set consisted of three types of arguments. In *consistent* arguments, all premise categories shared a salient relation that could be seen as relevant to the nature of the to-be-projected property. In Table 1, *doves*, *polar bears*, and *snow tigers,* are all *white animals*. For consistent arguments, the first two premises establish this relation, and the third premise supports it. Because of its confirmatory nature, the third premise should have few cognitive effects, and we therefore expected relatively fast reading times for the third premise in consistent arguments. In *garden-path* arguments, the first two categories shared a salient relation that is different from that shared by the second and third premises. Our intention was to lead participants down a garden path by setting up an expectation with the first two categories that was subsequently not met by the third. For example, in Table 1, *grizzly bears* and *polar bears* are both *bears*, whereas *polar bears* and *snow tigers* are both *white animals*. Thus, the likely garden-path hypothesis suggested by the first two premises—that the property has to do with membership in the category *bear*—is weakened by the third premise. Because of its inconsistency with the first two premises, the third premise should yield substantial cognitive effects, so we predicted relatively slow reading times for the third premise in garden-path arguments. Finally, in *neutral* arguments, the first two premise categories shared no salient property and were therefore unlikely to suggest any hypothesis; the third premise was the first likely source of a hypothesis. In Table 1, *donkeys* and *polar bears* share no salient relation (at least for us), whereas *snow tigers* and *polar bears* are *white animals*. In neutral arguments, the third premise presents the first coherent hypothesis about premise relations, and should therefore yield substantial cognitive effects. As such, we predicted relatively slow reading times for the third premise in neutral arguments.

It is important to note here that we made predictions about differential reading times for exactly the same sentence. In Table 1, the third premise in each argument is *Snow tigers have property X19.* Note also that in each case the third premise is preceded by exactly the same second premise; thus we controlled for local priming effects due to variations in strength of association between the categories in Premises 2 and 3. The only overt difference among the three argument types is the category in the first premise, through which we manipulated the nature of the hypothesis (if any) participants were likely to entertain upon reaching the third premise.

In all of the experiments described in this article, we used entirely blank properties (e.g., Property X12). Property effects have been widely reported in the literature (for a review and account based on the relevance framework, see Coley & Vasilyeva, in press), and seem to work via the property setting up a context in which participants retrieve or select relevant relations. Because we were concerned that consideration of context might affect our manipulations of cognitive effects, we used blank properties that we hoped would create a neutral context against which participants might consider relevant relations between the categories in the arguments.

In addition to examining cognitive effects and reading time, we had a secondary goal of testing similarity-based predictions about

Table 1
Examples of Premise and Conclusion Categories Used to Construct Arguments Used in Experiments 1–3

	Argument type		
Category	Consistent	Garden path	Neutral
First premise	Doves	Grizzly bears	Donkeys
Second premise	Polar bears	Polar bears	Polar bears
Third premise	Snow tigers	Snow tigers	Snow tigers
Specific conclusion (Experiments 1, 2, 3)	White animals	White animals	White animals
General conclusion (Experiments 1 and 2)	Animals	Animals	Animals
Garden-path conclusion (Experiments 2 and 3)	Bears	Bears	

argument strength. The argument types described above were constructed so that they also varied in the taxonomic diversity of the premise categories. Specifically, the premise categories for consistent arguments were always drawn from multiple superordinate categories (e.g., in Table 1, *birds* and *mammals*), whereas the premise categories for garden-path and neutral arguments were always drawn from a single superordinate category (in Table 1, *mammals*). As such, according to similarity-based approaches to category-based induction (Osherson et al., 1990; Sloman, 1993), consistent arguments with general conclusions should be judged stronger than garden-path and neutral arguments with general conclusions because their diverse premises provide greater coverage of a general conclusion category. In contrast, according to the relevance framework, when the categories in an argument make a hypothesis about the nature of the blank property readily available, and that hypothesis is inconsistent with the given conclusion category, argument strength should be weakened. Accordingly, the relevance framework predicts that consistent arguments with general conclusions should be no stronger, and perhaps even weaker, than garden-path and neutral arguments with general conclusions because of a mismatch between the relevant relation among premise categories and the conclusion category (see also McDonald et al., 1996). We also presented arguments with specific conclusions corresponding to the target relation for the consistent arguments (see Table 1). We had no predictions about these specific arguments, except that if participants are actually noticing our target relations, consistent-specific arguments should be rated as relatively strong.

Method

Participants. Forty-one students (mean age = 23.5 years) from Durham University (Durham, United Kingdom) were recruited by e-mail to take part in the study. Volunteers were paid £3 for their participation.

Materials. The experimental stimuli consisted of 16 argument sets, each comprising six related three-premise arguments. The second and third premises were the same for each argument within a set, were drawn from the same superordinate category, and shared a salient relation (e.g., speed, habitat, appearance, behavior); the first premise determined whether the argument was consistent, garden path, or neutral. For consistent arguments, the first premise category was drawn from a different superordinate than Premise Categories 2 and 3 but shared the same salient relation evident in the later premise categories (e.g., *magpies, panda bears, zebras*). For garden-path arguments, the first premise was drawn from the same superordinate as Premise Categories 2 and 3 and shared a salient relation with the second, but not the third, premise category creating a situation where the first and second premise categories shared one relation and the second and third premise categories shared a different one (e.g., *brown bears, panda bears, zebras*). For neutral arguments, the first premise category was also drawn from the same taxonomic superordinate as Premise Categories 2 and 3, but there was no other salient relation shared between the first two premise categories (e.g., *otters, panda bears, zebras*). Each argument could be paired with either a general conclusion (e.g., *magpies, panda bears, zebras/ALL ANIMALS*) or specific conclusion (e.g., *magpies, panda bears, zebras/ALL BLACK-AND-WHITE ANIMALS*). Specific conclusions corresponded to the salient property or relation shared by Premise Categories 2 and 3 (or all three in the case of consistent arguments).

Design. The experiment had a 3 (argument type: consistent, garden path, or neutral) × 2 (conclusion: specific or general) within-participant design. Dependent variables were reading time for the third premise and the proportion of trials in each condition where participants judged the conclusion to be strongly supported by the premises. Participants were presented with the general version of eight of the argument sets, and the specific version of the other eight, for a total of 48 arguments of the form of Arguments 6 and 7 above. Pairing of argument set with conclusion was counterbalanced across participants. The properties attributed to the objects or animals in the premises were blank letter–number combinations, such as Property X6. The specific letter–number combination was different for each problem. Arguments were presented in a different random order for each participant.

Procedure. Participants were individually tested using a computer running custom-written software. Before the main experiment, participants were given oral and written instructions and completed three example problems. Participants initiated the experimental session by pressing the spacebar to start the presentation of the premises. Premises were presented one at a time, and when the participant had read the first premise, she pressed the spacebar and the next premise appeared. The previous premise was not retained on the screen. Once the participant had read all three premises, a conclusion appeared and the participant had to decide whether the argument in the conclusion was strong or not strong by pressing one of two keys on the keyboard. The response keys corresponding to a strong or a weak argument endorsement were counterbalanced across participants. As well as recording judgments about each argument, for each premise the computer recorded the reading times from when the premise was displayed on the screen until the next press of the space bar.

Results

Data treatment. For our analysis of reading times, we computed the mean reading time for each premise in each condition. Where participants had recorded a premise reading time less than 100 ms or greater than 10 s, we replaced the outlier with the participant's mean premise reading time for that premise (first, second, or third) in that condition. We replaced 0.3% of all reading times in this way. For our analysis of judgments, for each participant we calculated the proportion of arguments in each condition that were judged strong.

Reading time and cognitive effort. Means for Premise 2 and 3 reading times, broken down by argument type, are to be found in Table 2. We argued that the relevance framework predicts that Premise 3 should require more cognitive effort when it appears in

Table 2
Mean Second- and Third-Premise Reading Times, Collapsed Across Conclusion, From Experiment 1

	Second premise		Third premise	
Argument type	*M*	*SD*	*M*	*SD*
Consistent	1,587	600	1,447	440
Garden path	1,502	412	1,634	469
Neutral	1,668	467	1,674	520

a garden-path argument than when it appears in a consistent argument. If so, Premise 3 should take longer to read in a garden-path arguments than in a consistent argument, and should result in a reading slowdown relative to Premise 2 for garden-path arguments. Note that because participants read the premises before they saw the conclusions in this experiment, for the purposes of this analysis we were able to collapse across the conclusion variable. Examination of the mean reading times in Table 2 suggests that our predictions have been borne out. This was confirmed by a 3 (argument type: consistent, garden-path, neutral) $\times$ 2 (premise: second, third) repeated measures analysis of variance (ANOVA) showing a significant interaction between argument type and premise, $F(2, 80) = 11.0$, $p < .001$, $\eta_p^2 = .22$. (Here and throughout the article we report analyses by participants. In every case we also carried out analyses by items. Unless otherwise noted, the results of the analyses by items and by participants were consistent.) Planned comparisons revealed that—as predicted by the relevance framework—participants took longer to read the third premise when it appeared in a garden-path argument than when it appeared in a consistent argument, $t(40) = 3.45$, $p = .001$, Cohen's $d = 0.42$. The third premise also took longer to read for neutral arguments than for consistent arguments, $t(40) = 4.12$, $p < .001$, $d = 0.48$. Moreover, for consistent arguments, participants spent less time reading the third premise than they did the second premise, $t(40) = 2.63$, $p < .02$, $d = 0.27$, whereas for garden-path arguments, participants spent more time reading the third premise than the second premise, $t(40) = 2.97$, $p < .006$, $d = 0.30$. Second and third premise reading times did not differ for neutral arguments, $t(40) = .12$, $p = .91$, $d = 0.01$.

Argument strength judgments. The proportion of arguments judged to be strong, broken down by premise set and conclusion are to be found in Table 3. A 2 $\times$ 3 within-participant ANOVA on judgment proportions revealed a significant main effect of conclusion, $F(1, 40) = 18.6$, $p < .001$, $\eta_p^2 = .32$, and a significant main effect of argument type, $F(2, 80) = 21.24$, $p < .001$, $\eta_p^2 = .35$. Both of these main effects were subsumed by the significant interaction, $F(2, 80) = 24.65$, $p < .001$, $\eta_p^2 = .38$. As expected, the proportion of arguments judged strong in the consistent specific condition was much greater than the proportion of strong judgments in any other condition. We used planned comparisons to test predictions, derived via coverage, that consistent general (i.e., more diverse) arguments would be judged stronger than neutral general arguments and garden-path general (i.e., less diverse) arguments. Proportion of arguments judged strong did not differ for garden-path versus consistent arguments, although consistent arguments were judged marginally weaker than neutral arguments, $t(40) = 1.85$, $p = .07$, $d = 0.20$.

Table 3
Mean Proportion of Arguments Endorsed as Strong in Experiment 1

	General conclusion		Specific conclusion	
Argument type	*M*	*SD*	*M*	*SD*
Consistent	.33	.29	.71	.33
Garden path	.34	.31	.40	.27
Neutral	.39	.32	.36	.27

Discussion

The analysis of reading times in this experiment confirm our prediction, derived from the relevance framework, that participants will expend more effort processing premises likely to have greater cognitive effects. In consistent arguments, the third premise confirmed likely hypotheses about what the experimenter intended to communicate about the nature of the blank property and therefore had little cognitive effect. In contrast, in garden-path arguments the very same premise did not support the likely hypothesis, whereas in neutral arguments it supported an initial hypothesis; in both cases, the third premise had much greater cognitive effect. We predicted that reading times would increase with cognitive effect, and as predicted, participants spent longer reading the third premise in garden-path and neutral arguments than they did in consistent arguments. For garden-path arguments only, third premise reading times were longer than second premise reading times. For consistent arguments, third premises were read more quickly than second premises, and there was almost no difference for neutral arguments. Moreover, in all three cases, the third premises were identical, and the immediately preceding premises were identical as well. Nevertheless, reading times varied as predicted. No current theory of induction, other than the relevance framework, appears to predict these results.

The results of evaluations of specific arguments support our contention that participants tended to form a hypothesis online about relations among premise categories; specific consistent arguments, in which all three premise categories shared the relation made explicit in the conclusion, were rated much higher than specific neutral or garden-path arguments.

Finally, strength ratings for general arguments provide no evidence of a diversity effect. Similarity-based accounts must predict that arguments with taxonomically diverse premises—that is, premises that provide better coverage of a conclusion category—will be perceived to be stronger than arguments with less diverse premises. In contrast, the relevance framework suggests that salient relations among premise but not conclusion categories may weaken or even reverse standard diversity effects. Results revealed no support for diversity; consistent-general arguments received the lowest proportion of strong ratings. Indeed, results revealed a nondiversity effect in that (less diverse) neutral-general arguments were perceived as marginally stronger than (more diverse) consistent-general arguments. This clearly fails to support similarity-based predictions.

It is possible that the dichotomous strong–weak response format may have reduced our ability to detect differences in perceived argument strength, or that sequential premise presentation may have interfered with argument evaluation and thereby masked diversity effects. To examine these possibilities, we presented an additional 29 participants with the same arguments in standard form (simultaneous presentation of premises and conclusions); participants rated argument strength using a 9-point Likert scale from 1 (*not strong*) to 9 (*strong*). Results replicated the argument strength patterns from Experiment 1 in detail, suggesting that the lack of a diversity effect was not an artifact of our methodology.

Experiment 2

The results of Experiments 1 clearly demonstrated that participants engage in significantly greater cognitive effort when pro-

cessing premises that are presumably inconsistent with their hypotheses about inference-relevant relations, or that lead to formation of new hypotheses, than when processing premises that are presumably consistent with an existing hypothesis. Thus far, we have made assumptions about what online hypotheses our participants are entertaining, and consequently about the magnitude of the cognitive effects of the third premise. In this experiment, we aimed to test these assumptions by examining the degree of belief change following receipt of a confirming or inconsistent premise.

A second aim of this experiment was to directly test the claim made in relevance theory that additional cognitive effort processing information will be most likely to be observed in cases where processing that information has greatest cognitive effects. In the case of category-based induction, the primary source of cognitive effect is people's beliefs about the strength of the argument. Thus, in Experiment 2 we attempted to examine the link between cognitive effort, as measured by the reading time for the third premise, and cognitive effects, as measured by the change in participants' beliefs about the argument once they had processed the third premise. Our hypothesis was that for garden-path arguments we should find a correlation between the time spent processing the third premise and the amount of change in participants' beliefs upon processing that premise. As an alternative test of the same hypothesis, we also examined the relation between conclusion reading times and belief change.

In this experiment, we focused on a comparison of consistent and garden-path premise sets, which provide the clearest contrast in potential for cognitive effect. For consistent premise sets (e.g., *magpies, panda bears, zebras*), the third premise should have little cognitive effect, because it simply reinforces a relation that is presumably already salient. For garden-path premises (e.g., *grizzly bears, panda bears, zebras*), the third premise should have a large cognitive effect because it is inconsistent with a likely hypothesis that the relevant relation has to do with membership of the category *bears* and also lends support to a previously unlikely hypothesis (*black and white animals*). In order to quantify the cognitive effects of the third premise in both cases, we examined the degree to which its presentation led to changes in argument strength ratings. To do so, we needed participants to evaluate arguments before and after the presentation of the third premise. For consistent premises, this was straightforward. Participants rated the strength of two-premise (e.g., *magpies, panda bears/black-and-white animals*) and three-premise (e.g., *magpies, panda bears, zebras/black-and-white animals*) versions of arguments made up of consistent premises and a conclusion consistent with the hypothesis suggested by those premises; we computed cognitive effect by taking the difference of these ratings.

For garden-path arguments, because the cognitive effect of the third premise should stem from it being inconsistent with one likely hypothesis as well as strengthening a previously unlikely hypothesis, measuring effect was less straightforward. To assess the extent to which the third premise ruled out the garden-path hypothesis, we compared the strength of two- and three-premise garden-path arguments presented with the garden-path conclusion (i.e., the conclusion we thought the first two premises were likely to bring to mind; see Table 1 for examples and the Appendix for a complete list). For example, we compared (*grizzly bears, panda bears/bears*) to (*grizzly bears, panda bears, zebras/bears*). Likewise, to assess the extent to which the third premise lent support to a previously unlikely hypothesis, we compared the strength of two- and three-premise garden-path arguments presented with the specific consistent conclusions presented in Experiment 1. For example, we compared (*grizzly bears, panda bears/black-and-white animals*) to (*grizzly bears, panda bears, zebras/black-and-white animals*). For completeness, we also assessed the strength of consistent two- and three-premise arguments with garden-path conclusions (e.g., *magpies, panda bears/bears* and *magpies, panda bears, zebras/bears*). Please note that the "consistent conclusions" used in Experiment 2 are identical to the "specific conclusions" used in Experiment 1. We introduce the change in nomenclature because unlike in the preceding experiment, Experiment 2 included two different types of specific conclusions.

Our design allowed us to test our assumptions about the relative cognitive effects of the third premise in garden-path versus consistent arguments. For garden-path arguments, we expected the third premise to markedly increase the perceived strength of arguments with consistent conclusions, and markedly decrease the perceived strength of arguments with garden-path conclusions. For consistent arguments, we expected the third premise to moderately increase the perceived strength of arguments with consistent conclusions, and have little effect on the perceived strength of arguments with garden-path conclusions, which should be relatively weak in any case. Greater differences between strength ratings of three-premise and two-premise garden-path arguments than between comparable consistent arguments would indicate that the third premise had a larger overall impact on belief change for garden-path arguments, and therefore had more cognitive effect. Finally, the design also allows us to look at specific relations between magnitude of cognitive effect and degree of cognitive effort; after all, the third premise should only take relatively long to read if it is causing a reassessment of the merits of the argument. On the other hand, if there is little change in perception of argument strength, there should be little effort devoted to processing the third premise.

Method

Participants. Forty-eight members of staff and students at Durham University were recruited by e-mail and were paid £3 to participate in this experiment.

Materials. The experimental stimuli consisted of the consistent and garden-path arguments used in Experiment 1 (see Appendix). Because the design called for only 15 argument sets, we dropped one of the sets used in the earlier experiments. Arguments were presented with either three premises (identical to those in earlier experiments) or two premises, in which case the first two premises were presented and the third omitted. Orthogonally, arguments were presented with either consistent or general conclusions (as in previous experiments) or with garden-path conclusions, which corresponded to the relation presumably rendered salient by the first two premises of garden-path arguments. We included items with general conclusions so that we could test ratings of argument strength for diversity and nondiversity effects, as we did in Experiments 1.

Design. The experiment had a 2 (argument type: consistent vs. garden path) × 2 (premises: two vs. three) × 3 (conclusion:

consistent, garden path, general) within-participant design. Dependent measures were time to read the second and third premises, time to evaluate the conclusions, and argument strength rating, measured on a 9-point scale.

Each participant was presented with a total of 60 arguments, including four arguments from each argument set (two- and three-premise versions of the consistent and garden-path versions of the argument). All arguments from a given set were presented to an individual participant with the same conclusion (consistent, garden path, or general). In total, each participant evaluated five argument sets presented with consistent conclusions, five with garden-path conclusions and five with general conclusions. Pairing of argument set with conclusion was counterbalanced across participants so that each form of each argument was evaluated an equal number of times.

Procedure. The procedure was identical to that in Experiment 1 (i.e., each premise was presented sequentially, contingent on a button press) except that argument strength was rated on a 9-point scale. In order to measure belief change following presentation of the third premise, we had all participants initially evaluate the 30 two-premise arguments. Next they completed an unrelated judgment task, after which they evaluated the 30 three-premise arguments.

Results

Data treatment. We performed the same data-trimming procedures on our reading time data as we did in Experiment 1. This resulted in the replacement of 0.3% of the data.

Item validation. Mean argument ratings for two- and three-premise arguments are presented in Table 4. Planned comparisons revealed that, as expected, two-premise garden-path arguments with garden-path conclusions were rated as stronger than two-premise garden-path arguments with consistent conclusions, $t(47) = 10.78$, $p < .001$, $d = 1.17$. In contrast, two-premise consistent arguments with consistent conclusions were rated as stronger than two-premise consistent arguments with garden-path conclusions, $t(47) = 6.79$, $p < .001$, $d = 0.94$. These results confirm that we did indeed lead participants down the garden path; after participants read the second premise, garden-path arguments rendered the garden-path conclusion more likely than the consistent conclusion, whereas consistent arguments rendered the consistent conclusion more likely than the garden-path conclusion.

Table 4
Mean Strength Ratings for Two-Premise and Three-Premise Arguments From Experiments 2 and 3

Argument type and conclusion	Two-premise arguments		Three-premise arguments	
	M	*SD*	*M*	*SD*
Experiment 2				
Consistent arguments				
Consistent	4.45	1.87	5.65	2.19
Garden path	2.88	1.47	3.36	1.67
Garden path arguments				
Consistent	2.95	1.41	4.20	1.88
Garden path	4.92	1.94	3.76	1.71
Experiment 3				
Consistent arguments				
Consistent	3.84	2.25	4.56	2.30
Garden path	2.52	1.38	2.69	1.35
Garden path arguments				
Consistent	2.42	1.34	3.23	1.50
Garden path	3.90	2.08	3.04	1.50

Argument strength and cognitive effect. Planned comparisons on strength ratings for two- and three-premise versions of the arguments confirmed our assumptions about the different cognitive effects of the third premise for consistent and garden-path arguments. As is evident in Table 4, for consistent arguments the third premise increased strength ratings, but this effect was bigger when the conclusion was consistent, $t(47) = 5.81$, $p < .001$, $d = 0.60$, than when it was garden path, $t(47) = 2.73$, $p < .01$, $d = 0.31$. For garden-path arguments, the third premise also increased strength ratings for the consistent conclusion, $t(47) = 6.16$, $p < .001$, $d = 0.76$, but decreased strength ratings for the garden-path conclusion, $t(47) = 4.56$, $p < .001$, $d = 0.64$. These results confirmed our assumptions about the effects of the third premise; for consistent arguments, the third premise reinforced an existing hypothesis (corresponding to the consistent conclusion), whereas for garden-path arguments, the third premise both weakened an existing hypothesis (corresponding to the garden-path conclusion) and strengthened an alternative hypothesis (corresponding to the consistent conclusion). Note that for garden-path arguments, the attenuating effect of the third premise on the hypothesis corresponding to the garden-path conclusion is an example of nonmonotonicity (see Osherson et al., 1990).

In order to quantify the overall cognitive effect of the third premise, we calculated the average absolute change in argument strength from two-premise to three-premise versions of consistent and garden-path arguments presented with either consistent or garden-path conclusions. If the third premise has a larger cognitive effect for garden-path arguments, we would expect a reliably higher average absolute change score for garden-path arguments. This is exactly what we observed; the average absolute change due to the third premise was greater for garden-path arguments ($M = 1.90$, $SD = 0.97$) than for consistent arguments ($M = 1.55$, $SD = 0.84$), $t(47) = 3.62$, $p = .001$, $d = 0.39$.

For both two- and three-premise arguments, we compared mean strength ratings for the taxonomically more diverse consistent premise sets to ratings of the garden-path premise sets, when each were presented with general conclusions. Once again, contrary to the predictions of similarity-based approaches, planned comparisons uncovered no evidence of a diversity effect in either the two-premise (consistent: $M = 3.02$, $SD = 1.60$; garden path: $M = 2.89$, $SD = 1.58$), $t(47) = .85$, $p = .4$, $d = 0.08$, or the three-premise (consistent: $M = 3.88$, $SD = 1.69$; garden path: $M = 3.91$, $SD = 1.80$) case, $t(47) = .32$, $p = .75$, $d = 0.02$.

Reading time and cognitive effort. Mean second- and third-premise reading times for consistent and garden-path arguments presented with consistent and garden path conclusions are displayed in Table 5. To test whether we had replicated the reading time effects observed in Experiment 1, we subjected these means to a 2 (argument type) × 2 (conclusion) × 2 (premise number) within-participant ANOVA. The analysis revealed a significant interaction between argument type and premise, $F(1, 47) = 20.22$, $p < .001$, $\eta_p^2 = .30$. Planned comparisons on the means involved in this interaction revealed that we have clearly replicated the

Table 5
Mean Reading Times for Second and Third Premises From Experiments 2 and 3

Argument type and conclusion	Second premise		Third premise	
	M	*SD*	*M*	*SD*
Experiment 2				
Consistent arguments				
Consistent	1,291	556	1,401	513
Garden path	1,261	521	1,364	626
Garden path arguments				
Consistent	1,181	481	1,517	605
Garden path	1,168	393	1,613	682
Experiment 3				
Consistent arguments				
Consistent	1,032	471	1,346	553
Garden path	1,235	602	1,421	511
Garden path arguments				
Consistent	981	627	1,360	657
Garden path	928	491	1,309	519

reading time findings from Experiment 1. Specifically, the third premises of garden-path arguments took significantly longer to read than the third premises of consistent arguments $t(47) = 3.33$, $p < .005$, $d = 0.35$.

Reading times for Premise 3 were slower than for Premise 2, both for garden-path arguments, $t(47) = 6.62$, $p < .003$, $d = 0.81$, and for consistent arguments, $t(47) = 2.26$, $p < .03$, $d = 0.22$. This last finding contrasts with the results of Experiment 1. Nevertheless, consistent with the proposal that Premise 3 requires more cognitive effort in garden-path arguments, the slowdown from Premise 2 to Premise 3 was significantly greater for garden-path arguments ($M = 391$ ms, $SD = 409$ ms) than for consistent arguments ($M = 107$ ms, $SD = 327$ ms), $t(47) = 4.5$, $p < .001$, $d = 0.78$.

Relations between reading time and cognitive effect. The relevance framework predicts that cognitive effort and cognitive effect should be related. In our paradigm, this translates into the prediction that the magnitude of change in argument strength from two-premise to three-premise arguments should be related to the time participants spend reading Premise 3—and possibly to the time they spend evaluating the conclusion—of three-premise arguments. Moreover, this relation should be most evident in garden-path arguments, where we have shown the third premise to have the most pronounced cognitive effect. To test this prediction, for each participant in each condition we calculated their average third-premise reading time, their average conclusion evaluation time, and their average belief change from two-premise to three-premise versions of an argument. The first and second are measures of the cognitive effort associated with each experimental condition, while the third is a measure of the cognitive effects associated with processing the third premise of arguments in that condition. For garden-path arguments with garden-path conclusions, we predicted a negative association between the measure of effect and the measure of effort because increased effort expended to process Premise 3 should result in a decrease in perceived argument strength. Conversely, for garden-path arguments with consistent conclusions we predicted a positive association between effect and effort because increased effort expended to process Premise 3 should result in an increase in perceived argument strength. Although we did not directly compare conclusion reading times because, unlike third-premise reading times, they involved different sentences, we did examine relations between conclusion reading–response times and changes in perceived argument strength. Because cognitive effort may be manifest in the time it takes to evaluate and respond to the conclusion as well as the time it takes to read and process the third premise, for garden-path arguments we made the same predictions for relations between conclusion reading–response time and change in perceived argument strength as we had for third-premise reading times. For consistent arguments, we expected little or no relation between effort and effect.

Correlations are presented in Table 6, where it may be seen that only for garden-path arguments presented with garden path conclusions did we find evidence of a significant association. Specifically, conclusion evaluation times for that condition were negatively associated with belief change scores. In other words, participants who spend longer evaluating the garden-path conclusions when presented with garden-path arguments were more likely to adjust their strength ratings downward having processed the third premise of the argument.

Discussion

The results of this experiment replicate and extend the results of Experiment 1 in several respects. First, we have clearly replicated the reading time findings from Experiment 1. Participants take longer to read the third premise of garden-path arguments than to read the third premise of consistent arguments. In addition, the difference between Premise 2 and 3 reading times is greater for garden-path arguments than for consistent arguments. The results of this experiment have also provided support for our assumptions about what hypotheses participants may or may not entertain prior

Table 6
Correlations Between Belief Change Scores and Reading Times From Experiments 2 and 3

Argument type and conclusion	Correlations between change in argument strength and measures of effort	
	Premise 3 reading time	Conclusion reading–response time
Experiment 2		
Consistent arguments		
Consistent	−.15	−.01
Garden path	−.13	.07
Garden path arguments		
Consistent	.09	.12
Garden path	−.21	−.33*
Experiment 3		
Consistent arguments		
Consistent	−.09	.14
Garden path	−.03	.08
Garden path arguments		
Consistent	.13	.28†
Garden path	−.36*	−.40*

† $p < .10$. * $p < .05$.

to encountering the third premise in our arguments; the first two premises of garden-path arguments rendered the garden-path conclusion likely, whereas the first two premises of consistent arguments rendered the consistent conclusion likely. Third, we have provided support for our assumption that the very same (third) premise leads to greater belief change (i.e., has greater cognitive effects) in a garden-path argument than in a consistent argument. Moreover, we have provided evidence for the prediction made by relevance theory (Sperber & Wilson, 1995) that cognitive effects are related to cognitive effort: Participants who spent more time reading and evaluating the conclusions of three-premise garden-path arguments presented with garden-path conclusions made larger downward adjustments in their evaluation of the strength of those arguments. One reason that we did not find a significant correlation between third-premise reading times and belief change for garden-path arguments might be that participants did not rate two and three versions of each argument in sequence. Accordingly, when it came to rating the strength of the three-premise version of an argument, participants might not have been able to remember their rating for the equivalent two-premise version, thus increasing the amount of error variance associated with our measure of cognitive effect. In Experiment 3, we presented both versions of each argument in sequence.

Experiment 3

The results of Experiments 1 and 2 have clearly demonstrated that participants engage in significantly greater cognitive effort when processing premises that fail to confirm their hypotheses about inference-relevant relations, or that lead to formation of new hypotheses, than when processing premises that are consistent with an existing hypothesis. In addition, the results of Experiment 2 have confirmed our assumptions about what online hypotheses our participants are entertaining, and consequently about the magnitude of the cognitive effects of the third premise. Although Experiment 2 provided some evidence of the associations between measures of cognitive effect and cognitive effort that are predicted by relevance theory, we did not find an association between belief change and third-premise reading times. In this experiment, we changed an aspect of our procedure that we hypothesized might have been interfering with our ability to detect a significant association between effect and effort. Instead of presenting participants with all two-premise arguments followed by all three-premise arguments, in Experiment 3 we had participants judge the two- and three-premise argument from each set in sequence. By doing so we hoped to draw particular attention to the addition of the third premise, and thereby obtain a more sensitive measure of the correspondence—if any—between the effort involved in processing the third premise and the cognitive effects thereof.

Method

Participants. Thirty-nine undergraduate students enrolled in an introductory psychology course at Northeastern University in Boston participated for partial course credit.

Materials. We used the consistent and garden-path arguments from the 16 argument sets used in Experiment 1 (see the Appendix), with a few minor changes to render the stimuli appropriate for participants in the northeastern United States rather than in the United Kingdom (e.g., *parakeet* replaced *budgie*; *skunk* replaced *magpie*). Two- and three-premise versions of each argument were presented. Arguments were presented with either consistent conclusions (as in previous experiments) or with garden-path conclusions, which corresponded to the relation presumably rendered salient by the first two premises of garden-path arguments.

Design. The experiment had a 2 (argument type: consistent vs. garden path) × 2 (premises: two vs. three) × 2 (conclusion: consistent, garden path) within-participant design. Dependent measures were time to read each premise, and argument strength rating, measured on a 9-point scale as in Experiment 2.

Each participant was presented with 32 pairs of arguments. Each pair consisted of the two- and three-premise versions of the same argument (i.e., the same premise–conclusion combination). Each participant was given two argument pairs from each argument set that differed in both premise and conclusion (i.e., for a given argument set, a participant was given either the consistent–consistent and garden path–garden path pairs, or the consistent–garden path and the garden path–consistent pairs). In total, each participant evaluated eight argument pairs for each possible premise–conclusion combination. Pairing of argument set with premise–conclusion was counterbalanced across participants so that each form of each argument was evaluated an equal number of times.

Procedure. The experiment was conducted using Superlab 4.0 experiment presentation software. At the level of presentation of arguments, the procedure was identical to that in Experiment 2 (i.e., each premise was presented sequentially, contingent on a button press). In order to highlight the third premise, we presented each argument pair consecutively with the three-premise argument always presented immediately after the two-premise version.

Results

Data treatment. Reading times over 10 s or under 100 ms were replaced with the participant's mean reading time for that premise (first, second, or third) in that condition. This resulted in the replacement of 3.2% of the data.

Item validation. Mean argument ratings for two- and three-premise arguments are presented in Table 4. Planned comparisons revealed that, as expected, two-premise garden-path arguments with garden-path conclusions were rated as stronger than two-premise garden-path arguments with consistent conclusions, $t(38) = 5.64$, $p < .0001$, $d = 0.85$. In contrast, two-premise consistent arguments with consistent conclusions were rated as stronger than two-premise consistent arguments with garden-path conclusions, $t(38) = 4.42$, $p < .0001$. These results confirm that we again lead participants down the garden path; after participants read the second premise, garden-path arguments rendered the garden-path conclusion more likely than the consistent conclusion, whereas consistent arguments rendered the consistent conclusion more likely than the garden-path conclusion.

Argument strength and cognitive effect. Planned comparisons of strength ratings for two- and three-premise versions of the arguments again confirmed our assumptions about the different cognitive effects of the third premise for consistent and garden-path arguments. As is evident in Table 4, for consistent arguments the third premise increased strength ratings for the consistent conclusion, $t(38) = 5.85$, $p < .0001$, $d = 0.32$, but had no effect on strength

ratings for the garden-path conclusion, $t(38) = 1.34$, $p = .190$, $d = 0.12$. For garden-path arguments, the third premise also increased strength ratings for the consistent conclusion, $t(38) = 5.92$, $p < .0001$, $d = 0.57$, but decreased strength ratings for the garden-path conclusion, $t(38) = 3.94$, $p = .0003$, $d = 0.47$. These results again demonstrate that for consistent arguments the third premise reinforced an existing hypothesis (corresponding to the consistent conclusion), whereas for garden-path arguments the third premise both weakened an existing hypothesis (corresponding to the garden-path conclusion) and strengthened an alternative hypothesis (corresponding to the consistent conclusion). These results suggest that although Premise 3 had substantial cognitive effects for three out of four types of arguments, those effects were especially pronounced for garden-path arguments. Indeed, as in Experiment 2, absolute change was greater for garden-path arguments ($M = 1.88$, $SD = 1.60$) than for consistent arguments ($M = 1.38$, $SD = 0.91$), $t(38) = 2.05$, $p < .05$, $d = 0.38$ (although the test by items was nonsignificant, $t[15] = 1.78$, $p = .095$, $d = 0.60$).

Reading time and cognitive effort. To examine reading time effects, we again carried out 2 (argument type) × 2 (conclusion) × 2 (premise number) within-participant ANOVAs by participants and by items on second- and third-premise reading times. Mean reading times are presented in Table 5. As in Experiments 1 and 2, this analysis yielded an interaction between argument type and premise number, $F(1, 38) = 5.08$, $p = .030$, $\eta_p^2 = .12$. Planned comparisons on the means involved in this interaction revealed that unlike in Experiments 1 and 2, mean reading times for Premise 3 did not differ for consistent versus garden-path arguments, $t(38) = 1.04$, $p = .307$, $d = 0.09$. As in Experiment 2, reading times for Premise 3 were slower than for Premise 2, both for garden-path arguments, $t(38) = 6.71$, $p < .0001$, $d = 0.71$, and for consistent arguments, $t(38) = 4.46$, $p < .0001$, $d = 0.50$. Nevertheless, as in Experiment 2, the slowdown from Premise 2 to Premise 3 was significantly greater for garden-path arguments ($M = 380$ ms, $SD = 353$ ms) than for consistent arguments ($M = 250$ ms, $SD = 350$ ms), $t(38) = 2.25$, $p = .030$, $d = 0.37$. These results suggest that although sequential presentation may have eliminated absolute differences in reading times for Premise 3, the increase in cognitive effort to process Premise 3 relative to Premise 2 was still greater for garden-path arguments than for consistent arguments.

Relations between cognitive effort and cognitive effect. As in Experiment 2, we examined relations between mean third-premise reading time and mean conclusion evaluation time (as measures of cognitive effort) and mean belief change from two- to three-premise versions of an argument (as a measure of cognitive effect). If increased cognitive effort is associated with increased cognitive effects, changes in strength ratings should be associated with one or both reading time measures. Correlations are presented in Table 6 (one case was dropped from this correlation analysis because of a change score that was over three standard deviations from the mean).

As predicted, for garden-path arguments there was a reliable relation between cognitive effort and cognitive effect. For garden-path arguments with garden-path conclusions, as we saw above, the addition of Premise 3 reduced perceived argument strength. Correspondingly, as the negative correlations in Table 6 suggest, the magnitude of this reduction was predicted by the length of time taken to process both Premise 3 and the conclusion. For garden-path arguments with consistent conclusions, the addition of Premise 3 increased perceived argument strength; the positive correlation in Table 6 suggests the magnitude of this increase was predicted by the length of time taken to process the conclusion (but not Premise 3). In contrast, for consistent arguments we observed no relation between premise or conclusion reading time and magnitude of change in argument strength.

Discussion

The results of this experiment replicate and extend the results of Experiments 1 and 2. First, we have replicated the findings from Experiment 2 in support of our assumptions about what hypotheses participants may or may not entertain prior to encountering the third premise in our arguments; the first two premises of garden-path arguments rendered the garden-path conclusion likely, whereas the first two premises of consistent arguments rendered the consistent conclusion likely. Second, we have provided support for our assumption that the very same (third) premise leads to greater belief change (i.e., has greater cognitive effects) in a garden-path argument than in a consistent argument. Third, we have shown that, as predicted by relevance theory (Sperber & Wilson, 1995), cognitive effects are related to cognitive effort. For garden-path arguments, there was a clear association between cognitive effort and cognitive effect. With garden-path conclusions, people who took longer to read Premise 3—and to read and respond to the conclusion—tended to make larger reductions in their strength ratings. With consistent conclusions, people who took longer to read and respond to the conclusion tended to make larger increases in their strength ratings. Importantly, this effect was not about differences between slow and fast readers; reading time was unrelated to changes in argument strength ratings for consistent arguments. To the best of our knowledge, this is the first experimental demonstration of a link between cognitive effort and cognitive effects.

General Discussion

The relevance framework for category-based induction (Medin et al., 2003) suggests that when evaluating inductive arguments, people compare premises and conclusions to determine which relations may be relevant for evaluating the argument. Relevant relations are those that require little cognitive effort to process (i.e., are highly salient), or those that have large cognitive effect (i.e., result in marked belief change). In this article, we have demonstrated that the relevance framework may be differentiated from other approaches by virtue of its unique predictions about online measures of premise processing. Previous research has examined the relevance framework chiefly by manipulating cognitive effort—the salience of premise–premise or premise–conclusion relations—in order to predict and explain exceptions to standard category-based induction effects derived from similarity-based approaches. In contrast, by manipulating cognitive effect, we have shown that participants take longer to read a premise when it is inconsistent with a salient hypothesis, or establishes a novel hypothesis, than when the very same premise simply confirms an existing hypothesis. Moreover, we have shown that this increased processing was greatest among participants for whom inconsistent third premises tended to have a relatively large effect.

The relevance framework can easily accommodate our reading time findings as well as the link we have observed between reading times and belief change. When premises are presented in sequence to participants, they may update their hypotheses about the nature of the blank property upon receipt of each new premise. New premises that are inconsistent with a current hypothesis about the nature of the blank property have more cognitive effect than premises that merely confirm a hypothesis. According to relevance theory, people commit processing resources to new information in proportion to its likely cognitive effect. Therefore, the relevance framework predicts that people will spend longer in reading inconsistent premises than they will in reading confirming premises. Moreover, because of the relation between cognitive effort and cognitive effect, the relevance framework predicts that increased cognitive effort (as indexed via longer reading time) should only be observed in cases yielding potentially large cognitive effects (as indexed by relatively larger changes in belief). Taken together, the relevance framework for category-based induction is the only account that can predict and explain this set of findings.

In contrast, these reading time findings are difficult for similarity-based approaches to explain. Although similarity-based approaches differ in terms of the number of processes they invoke to explain category-based induction (for an empirical test of these predictions, see Feeney, 2007), no similarity-based approach that we know of makes processing predictions at the level of reading times for individual premises of an argument. In recent years, there have been attempts to model category-based induction in Bayesian terms (see Heit, 1998; Tenenbaum et al., 2007), and it might be possible to capture our data in these same terms. However, generating a Bayesian account to capture reading time effects may not be straightforward. Extant Bayesian models of induction attempt to capture people's judgments about the probability that the blank property extends to the category in the conclusion. Although some of our findings relate to judgments of argument strength, the most novel aspect of our findings relate to premise reading times, which occur before participants receive the conclusion to the argument. Thus, in order for our results to be captured in a Bayesian way, the process of testing hypotheses about the nature of the relevant relation would have to be formally modeled.

Along these same lines, in the first two experiments reported here we have failed to observe diversity effects when comparing strength ratings for arguments that varied in the taxonomic diversity of their premise categories. The absence of diversity effects is very difficult to explain under a purely similarity-based approach (e.g., Osherson et al., 1990; Sloman, 1993) but fall naturally out of the relevance framework. Because relations among premise categories are deemed relevant to evaluating the argument, people compare premise categories to generate hypotheses about the nature of the blank property. If an otherwise taxonomically diverse set of premise categories shares a salient nontaxonomic relation (e.g., racing pigeons, greyhounds, horses), the mismatch between that salient relation and a general conclusion may weaken the argument relative to a less diverse set of premises that offers no compelling alternative (see McDonald et al., 1996). Accordingly, our participants assigned low argument strength ratings to arguments whose premise categories were very diverse but shared a characteristic. Even stronger evidence for the relevance framework would have been a finding of significant nondiversity effects rather than an absence of diversity effects, and we consider our reading time findings to be more persuasive than participants' ratings of argument strength. While some of Medin et al.'s (2003) findings of exceptions to standard category-based induction effects have been replicated and extended (see Crisp, Feeney, & Shafto, 2010; Feeney et al., 2007), others have proven more problematic (see Heit & Feeney, 2005). It may be the case that online measures of processing effort will turn out to be more useful to researchers than offline measures of cognitive effects such as argument strength.

Although our initial interest in the relevance framework arose because it allowed us to derive predictions about premise processing in category-based inductive reasoning, because relevance theory was intended as a general cognitive theory, the existence of a link between cognitive effort and cognitive effect in a high-level cognitive task may be of interest to those working in areas as diverse as attention allocation, deductive reasoning, or linguistic pragmatics. For example, our findings are similar to recent work (Itti & Baldi, 2006) showing that a measure of surprise predicts where people look when inspecting moving images taken from television and video games. This work parallels our own in that it shows that the surprisingness of visual events at a particular location in a moving image predicts the attentional resource that is allocated to that location. Likewise, we have shown that a premise is processed for a longer period of time when it is inconsistent with a hypothesis about the nature of the relevant relation in an argument (and is therefore surprising) than when it confirms a hypothesis and is therefore less surprising. Future work might seek to discover whether this association between the attention allocated to new information and the subsequent effects of that information generalize from category-based inductive reasoning to other tasks involving belief revision and belief updating.

Conclusions

Our results strongly support the relevance framework for explaining category-based induction. Although Bayesian theories may also be able predict exceptions to phenomena such as the diversity effect (see Tenenbaum et al., 2007), only the relevance framework makes reading time predictions at the level of individual premises of an argument, or allows one to predict individual differences in reading times based on the cognitive effects of the premises.

Few, if any, other studies of category-based induction have employed online processing measures. This is in contrast to the literature on deduction where measures of reading time have been employed (e.g., Santamaria, Espino, & Byrne, 2005; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003). We hope that our methodology will encourage other investigators to derive and test predictions about online processing from a range of theoretical approaches to inductive reasoning. Although judgments of argument strength have been of central importance when evaluating theories of induction, increased use of online measures may encourage investigators to focus on the psychological process that underlie our ability to make inferences in the face of uncertainty.

References

Coley, J. D., & Vasilyeva, N. Y. (in press). Generating inductive inferences: Premise relations and property effects. *Psychology of Learning and Motivation.*

Coley, J. D., Vitkin, A. Z., Seaton, C. E., & Yopchick, J. E. (2005). Effects of experience on relational inferences in children: The case of folk biology. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 471–475). Mahwah, NJ: Erlbaum.

Crisp, A. K., Feeney, A., & Shafto, P. (2010). Testing dual process accounts of the category-based conjunction fallacy: When is decontextualised reasoning necessary for logical responding? Manuscript submitted for publication.

Feeney, A. (2007). How many processes underlie category-based induction? Effects of conclusion specificity and cognitive ability. *Memory & Cognition, 35,* 1830–1839.

Feeney, A., Shafto, P., & Dunning, D. (2007). Who is susceptible to conjunction fallacies in category-based induction? *Psychonomic Bulletin & Review, 14,* 884–889.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). New York, NY: Academic Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford, England: Oxford University Press.

Heit, E. (2007). What is induction and why study it? In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 1–24). Cambridge, England: Cambridge University Press.

Heit, E., & Feeney, A. (2005). Relations between premise similarity and inductive strength. *Psychonomic Bulletin & Review, 12,* 340–344.

Heit, E., Hahn, U., & Feeney, A. (2005). Defending diversity. In W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside of the lab: Essays in honor of Douglas L. Medin* (pp. 87–100). Washington, DC: American Psychological Association.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 411–422.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Proceedings of Neural Information Processing Systems (NIPS 2005)* (pp. 547–554). Cambridge, MA: MIT Press.

López, A., Atran, S., Coley, J. D., Medin, D., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folk biological taxonomies and inductions. *Cognitive Psychology, 32,* 251–295.

McDonald, J., Samuels, M., & Rispoli, J. (1996). A hypothesis-assessment model of categorical argument strength. *Cognition, 59,* 199–217.

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

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278.

Murphy, G. L., & Ross, B. H. (1999). Induction with cross-classified categories. *Memory & Cognition, 27,* 1024–1041.

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

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 811–828.

Santamaria, C., Espino, O., & Byrne, R. M. J. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1149–1154.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naïve similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 641–649.

Shafto, P., Coley, J. D., & Vitkin, A. (2007). Availability in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 114–136). Cambridge, England: Cambridge University Press.

Sloman, S. A. (1993). Feature based induction. *Cognitive Psychology, 25,* 231–280.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford, England: Blackwell.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 167–204). Cambridge, England: Cambridge University Press.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin & Review, 10,* 184–189.

Appendix

Problem Sets Used in Experiments 1–3

Argument types			Conclusions		
Neutral	Consistent	Garden path	General	Specific	Garden path
Mice, dolphins, walruses	Tuna, dolphins, walruses	Chimps, dolphins, walruses	All animals	All sea animals	All intelligent animals
Monkeys, greyhounds, horses	Racing pigeons, greyhounds horses	German shepherds, greyhounds, horses	All animals	All racing animals	All dogs
Sparrows, ducks, herons	Frogs, ducks, herons	Chickens, ducks, herons	All animals	All pond animals	All domestic fowl
Centipedes, mosquitoes, leeches	Vampire bats, mosquitoes, leeches	Flies, mosquitoes, leeches	All animals	All blood-sucking animals	All flying insects
Greek people, British people, Australian people	American people, British people, Australian people	French people, British people, Australian people	People of all nations	People of all English-speaking nations	People of all European nations
Donkeys, polar bears, snow tigers	Doves, polar bears, snow tigers	Grizzly bears, polar bears, snow tigers	All animals	All white animals	All bears
Tigers, rats, cockroaches	Pigeons, rats, cockroaches	Squirrels, rats, cockroaches	All animals	All urban pests	All rodents
Sheep, lions, cobras	Scorpions, lions, cobras	Cats, lions, cobras	All animals	All dangerous animals	All cats
Cats, zebras, skunks	Striped bass, zebras, skunks	Horses, zebras, skunks	All animals	All striped animals	All hooved animals
Moose, lions, wolves	Sharks, lions, wolves	Elephants, lions, wolves	All animals	All predatory animals	All large African animals
Chipmunks, cows, pigs	Chickens, cows, pigs	Buffalo, cows, pigs	All animals	All farm animals	All grazing animals
Vultures, penguins, puffins	Polar bears, penguins, puffins	Ostriches, penguins, puffins	All animals	All cold-climate animals	—
Hondas, Ferraris, jet planes	Express trains, Ferraris, jet planes	Fiats, Ferraris, jet planes	All modes of transport	All fast modes of transport	All cars
Foxes, whales, dolphins	Cod, whales, dolphins	Elephants, whales, dolphins	All animals	All sea animals	All large animals
Otters, panda bears, zebra	Magpies, panda bears, zebra	Brown bears, panda bears, zebra	All animals	All black-and-white animals	All bears
Badgers, hamsters, dogs	Budgies, hamsters, dogs	Squirrels, hamsters, dogs	All animals	All pet animals	All rodents

Note. The dash indicates that this item was not used in Experiment 2.

Received February 25, 2009
Revision received February 11, 2010
Accepted February 17, 2010 ■