Psychonomic Bulletin & Review
2003, 10 (3), 517-532

THEORETICAL AND REVIEW ARTICLES

A relevance theory of induction

DOUGLAS L. MEDIN
Northwestern University, Evanston, Illinois

JOHN D. COLEY
Northeastern University, Boston, Massachusetts

GERT STORMS
University of Leuven, Leuven, Belgium

and

BRETT K. HAYES
University of New South Wales, Sydney, New South Wales, Australia

A framework theory, organized around the principle of relevance, is proposed for category-based reasoning. According to the relevance principle, people assume that premises are informative with respect to conclusions. This idea leads to the prediction that people will use causal scenarios and property reinforcement strategies in inductive reasoning. These predictions are contrasted with both existing models and normative logic. Judgments of argument strength were gathered in three different countries, and the results showed the importance of both causal scenarios and property reinforcement in category-based inferences. The relation between the relevance framework and existing models of category-based inductive reasoning is discussed in the light of these findings.

One of the central functions of categorization is to support reasoning. Having categorized some entity as a *bird*, one may predict with reasonable confidence that it builds a nest, sings, and can fly, though none of these inferences is certain. In addition, between-category relations may guide reasoning. For example, from the knowledge that robins have some enzyme in their blood, one is likely to be more confident that sparrows also have this enzyme than that raccoons have this enzyme. The basis for this confidence may be that robins are more similar to sparrows than to raccoons or that robins and sparrows share a lower rank superordinate category (birds) than do robins and raccoons (vertebrates).

Recently, researchers have developed specific models for category-based reasoning and generated a range of distinctive reasoning phenomena (see Heit, 2000, for a review). These phenomena are quite robust when American college students are the research participants, but at least some of them do not generalize well to other populations. To address these limitations, we will offer not so much a specific model but rather a framework theory organized around the principle of relevance. This theory is more abstract than many of its predecessors, and one might imagine a number of implementations consistent with the relevance framework. Nonetheless, we will see that the relevance theory has testable implications.

The rest of the paper is organized as follows. First, we briefly review two of the most influential models for induction: the Osherson, Smith, Wilkie, López, and Shafir (1990) category-based induction model, and Sloman's (1993) feature-based induction model. Next, we turn to the question of the generality of reasoning phenomena and describe two, more abstract, approaches that may be able to address the question of generality. Then we offer a theory at an intermediate level of abstraction, the "relevance theory," and describe some tests of its implications. Finally, we summarize and argue that there are benefits from approaching induction from a number of levels of analysis.

The similarity-coverage model (SCM). The Osherson et al. (1990) model of induction is driven by two related notions: similarity and coverage. *Similarity* refers to the assumption that, all else being equal, people are

This research was supported by Grants NIH 55079 and NSF 9983260 to the first author. J.D.C., G.S., and B.K.H. contributed equally to this paper; their order of authorship was determined randomly. G.S. would like to thank the Flemish National Research Fund for the travel grant while visiting Northwestern University. B.K.H.'s contribution was supported by Australian Research Council Large Grant A00000851. We thank several colleagues who contributed to this work, including Scott Atran, Christian Luhmann, Christine Selmes, and Sally Hunt. Correspondence concerning this article should be addressed to D. L. Medin, Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208-2710 (e-mail: medin@northwestern.edu).

more likely to extend a predicate from a base premise to a target premise to the extent that the target category is similar to the base category. Given the premise that dogs have sesamoid bones, we are more likely to think that wolves have sesamoid bones than that cows do. The SCM also assumes that judgments may be partially based on the similarity of the premise category to examples of the lowest level superordinate category that spans the premise and conclusion categories. Consider, for example, the premise *bears have sesamoid bones* and the conclusion *therefore, all mammals have sesamoid bones.* According to the SCM, to evaluate this argument people would generate examples of the mammal category (e.g., dog, cow, wolves, horse, lion) and compute their similarity to the premise category, bear. In this example, the *coverage* would be the sum of the similarities of retrieved instances to bear. If the premise were that whales have sesamoid bones and the conclusion that all mammals do, then the same instance retrieval and similarity calculation process is assumed to operate. In this case, the summed similarities or coverage would be less, because whales are atypical mammals and less similar on average to other mammals than are bears. This example illustrates that the SCM predicts typicality effects in reasoning because typical examples have better coverage than atypical examples do.

Two of the best-studied phenomena associated with the SCM rely on the notion of coverage: *typicality* and *diversity.* As we have seen, typicality effects in reasoning follow directly from the definition of typicality in terms of similarity to other category members (Rosch & Mervis, 1975).

Diversity concerns coverage associated with multiple-premise arguments. Consider, for example, the relative strength of the premises that crows and blackbirds have property X versus the premise that crows and ducks have property X for the conclusion that all birds have property X. In the SCM, coverage is based on the average *maximal* similarity that examples of the category have to the premise examples. Crows and blackbirds are quite similar, and the coverage provided by each of them will be redundant to that provided by the other. In contrast, ducks are different from crows and will have substantially greater similarity to a number of birds (e.g., geese, swans, loons, pelicans, gulls) than will crows. This will produce better overall coverage. In short, the SCM predicts that two diverse premises will have greater induction strength for a category than will two similar premises. (Note, however, that two very different but atypical examples of a category, such as penguins and hummingbirds, may have poor overall coverage, and therefore, coverage cannot be equated with dissimilarity of premises; see Osherson et al., 1990, pp. 199–200).

The SCM is deceptively simple. It has only a single parameter reflecting the relative weight given to the similarity and coverage components. Given a set of category similarities, it can be used to generate a variety of both intuitive and counterintuitive predictions that have received considerable support (see Osherson et al., 1990).

Feature-based induction model (FBIM). Sloman's (1993) feature-based induction model also relies on the notions of (featural) similarity and (featural) coverage. The central idea is that similarity is driven by matching and mismatching features and that an argument is strong to the extent that the premise and conclusion categories share features. A distinctive property of the FBIM is that is does not use category information in the sense that it does not distinguish between different levels of categorization. Instead, it assumes that all categories are represented in terms of features and that argument strength is based on feature overlap.

It may seem that the FBIM is just the SCM with the notion of similarity decomposed into featural matches and mismatches. But the FBIM has no notion of generating category examples, and the fact that it treats a category as just a feature set leads to some unique predictions, predictions that have received support (e.g., Sloman, 1993, 1998). Although FBIM and SCM are distinct, for the present purposes we will treat them as providing more or less comparable accounts of typicality and diversity effects, phenomena to which we now turn.

Typicality and diversity are very robust phenomena in American undergraduate study populations. But these results do not generalize well to other groups. López, Atran, Coley, Medin, and Smith (1997) used local mammals as stimuli to study induction among the Itza' Maya. University of Michigan undergraduates' reasoning about mammals of Michigan provided a control or comparison condition. The Itza' Maya showed reliable typicality effects, but either no diversity effects or below chance diversity effects. Undergraduates displayed strong typicality and diversity effects. Proffitt, Coley, and Medin (2000) studied different types of tree experts' reasoning about trees. None of the groups showed typicality effects. Taxonomists showed reliable diversity effects, but parks maintenance workers responded below chance on diversity probes. Bailenson, Shum, Atran, Medin, and Coley (2002) studied Itza' Maya, U.S. bird experts', and Northwestern University undergraduates' categorizing and reasoning about birds of Illinois and birds of Guatemala. The Itza' and the bird experts were not reliably above chance on either typicality or diversity probes. Undergraduate responses and justifications strongly conformed to both typicality and diversity. In short, typicality and diversity effects are far from common in populations that have considerable knowledge concerning the domain of categories under study.

Why don't experts and Itza' (who are themselves biological experts) produce clear typicality and diversity responding? The most salient reason is that often they are instead employing causal and ecological reasoning about the kinds in question. For example, Proffitt et al. (2000) found that tree experts often reason about arguments involving novel tree diseases in terms of how widely planted different kinds of trees are, their susceptibility to disease, and so on. López et al. (1997) noted very similar reasoning strategies among the Itza' Maya. One might

argue that these informants were not treating these predicates as truly blank properties, but to take this stance artificially limits the potential scope of models of induction and risks a certain circularity (e.g., the SCM should apply only where the responses match its predictions). Furthermore, diversity-based reasoning is not absent in these populations, but rather seems to be one of several strategies employed. An alternative approach to capturing this range of results is to broaden the scope of induction models. We now turn to two models for induction that do just that.

Hypothesis-based induction. McDonald, Samuels, and Rispoli (1996) proposed what they refer to as a hypothesis-based model of induction. They argue that induction may be guided more by theories or explanations than by similarity itself. On this view, inductive strength may be based in part on whether the premises suggest alternative categories or hypotheses to the conclusion category given (these act as competing explanations; see also Sloman, 1994). They provide support for their framework by asking people to generate hypotheses or explanations and by showing that inductive confidence decreases when there are competing hypotheses (candidates for a conclusion category).

A Bayesian model. Heit (1998, 2000) has taken a Bayesian approach to category-based induction. The idea is that people have expectations about the distributions of properties or features and that their judgments are based on these subjective distributions. Consider, for example, typicality effects in reasoning. The reasoner is assumed to consider the features that are unique to a premise, the features that might hold for the premise and categories that are subordinate to or overlap with the conclusion category, and the features that match the conclusion category. The advantage that a typical premise has over an atypical premise is that it may have relatively fewer distinctive features and fewer features shared with overlapping or subordinate categories. The Bayesian model is similar in spirit to Sloman's (1993) feature-based induction model, though they are far from equivalent (see also Griffiths & Tenenbaum, 2000).

A nice feature of the Bayesian approach (and the hypothesis-based model as well) is that it provides for more flexibility in induction. People's knowledge may lead them to have different expectations about the features relevant for induction. For example, Heit and Rubinstein's (1994) finding that physiological features or predicates trigger different patterns of induction than behavioral predicates follows naturally from this framework. Depending on expectations about feature distributions, the Bayesian framework, like Sloman's feature-based model, may provide an account for when diversity effects are or are not obtained.

The Bayesian model may be evaluated in a manner analogous to the hypothesis-based induction model. For a set of premise categories and predicates, one might obtain people's judgments about common and distinctive features and then use these distributions to make predictions about reasoning phenomena. A close correspondence supports the model. Lack of correspondence suggests either that the model is flawed or that the feature elicitation procedure is faulty.

Analysis. We think that each of the models for induction so far proposed contains valuable insights. All of the models constrain their predictions by obtaining predictor measures (e.g., similarity judgments, featural distributions, hypotheses) and using them to predict patterns of reasoning. The SCM is perhaps the most constrained in that the similarity judgments may be collected in a task remote from the reasoning task. To the extent that relevant features and hypotheses are thought to depend on the predicates and specific combinations of premises and conclusions, the predictor variables must be collected in a context very close to the actual reasoning task. Very likely there is a tradeoff—the closer the predictor task is to the predicted, the more accurate the predictions should be. But it is also true that the closer the tasks are, the more open the framework is to the criticism that its account has a circular flavor.

There seems to be something of a continuum. At one end, the SCM makes strong predictions but fails to capture some of the dynamic aspects of how people reason about categories. Bayesian and hypothesis-based models can address many of the more contextualized aspects of reasoning but are less able to make a priori predictions. In this paper, we offer an intermediate level framework; our goal is to provide an account of the dynamic and context-dependent components of category-based reasoning by postulating some processing principles that fall under the broad umbrella of relevance. We turn to that now.

Relevance Theory: An Overview

The lack of generality of typicality and diversity effects beyond undergraduate populations represents a serious limitation of most current models of induction, which generally predict that these phenomena will be more robust than they are. One of our test sessions with a tree expert provided the impetus for a shift toward a different framework theory. The expert was given typicality probes such as the following: "Suppose we know that river birch get Disease X and that white oaks get Disease Y, which disease do you think is more likely to affect all trees?" In this case, the expert said Disease X, noting that river birches are very susceptible to disease; so, "if one gets it they all get it." The very next probe involved the gingko tree, and the expert chose the disease associated with it as more likely to affect all trees on the grounds that "Gingkos are so resistant to disease that if they get it, it must be a very powerful disease." He then said that he felt as if he had just contradicted himself, but that nonetheless these seemed like the right answers.

Normatively, this expert's answers do not represent a contradiction. Instead, he appeared to be using the information that was most salient and accessible to guide his reasoning (on spontaneous feature-listing tasks, experts indicate that birches are notoriously susceptible to,

and gingkos notoriously resistant to, diseases). Simply put, the expert was using the knowledge that he considered most relevant.

We believe that Sperber and Wilson's (1986) relevance theory provides a good framework for understanding category-based induction. Furthermore, it leads to a number of novel predictions that contrast with those of other models of induction. In relevance theory, relevance is seen as a property of inputs to cognitive processes:

> An input is relevant to an individual at a certain time if processing this input yields *cognitive effects*. Examples of cognitive effects are the revision of previous beliefs, or the derivation of contextual conclusions, that is, conclusions that follow from the input taken together with previously available information. Such revisions or conclusions are particularly relevant when they answer questions that the individual had in mind (or in an experimental situation, was presented with). (Van der Henst, Sperber, & Politzer, 2002, p. 4)

In the Proffitt et al. (2000) studies, background knowledge about properties of trees and diseases presumably provides that basis for the sorts of contextual conclusions mentioned by our tree expert. Van der Henst et al. (2002) further elaborate:

> Everything else being equal, the greater the cognitive effects achieved by processing an input, the greater its relevance. On the other hand, the greater the effort involved in processing an input, the lower the relevance. . . . One implication of the definition of relevance in terms of effect *and effort* is that salient information, everything else being equal, has greater relevance, given that accessing it requires less effort. (p. 4)

Potentially, there are two problems with relevance theory that may limit its applicability to studies of induction. One is that it is not possible to maximize two functions at once. In general, more effort should lead to more effect, so it is not obvious how to trade off one for the other in determining relevance. The second, related, problem is that relevance theory appears to be subject to the same circularity criticism that we have raised with respect to Bayesian and hypothesis-based models.

Although it is not possible to simultaneously maximize (least) effort and (greatest) effect, one can experimentally manipulate effort and effect to determine whether they have the sorts of consequences predicted by relevance theory. In the present paper, we focus on undergraduates. They generally have little background knowledge to bring to bear on the sorts of reasoning tasks we have used. Consequently, it is not surprising that they rely heavily on more abstract reasoning strategies. However, it may be possible to select probes related to the limited biological knowledge that they have in order to vary what Sperber et al. call *effect*. As we shall see, it is also easy to experimentally manipulate effort. In the next few paragraphs, we will outline how relevance theory may apply to category-based induction and then develop specific predictions for our studies.

Relevance in category-based induction. The general idea is that the premises are assumed to be relevant to the conclusion(s). One motivation for this view is the fact that experiments take place in a social context and participants reasonably infer that the experimenter is being relevant and informative with respect to the inductive argument forms (cf. Grice, 1975). We also believe, however, that people may generally assume something like a principle of relevance or informativeness regardless of the source of observations.

How does the principle of relevance constrain induction? We suggest that when a blank property or predicate is associated with some premise category, people tend to associate that property with the most distinctive or informative features or categories associated with the premise. For example, immediate superordinate categories generally should be more salient and relevant than more remote superordinates, because immediate superordinates are more unusual (have lower base rates) and are therefore more informative (in an information-theoretic sense) than remote superordinates. Note that informativeness in this case follows a principle of parsimony and that it is concordant with Osherson et al.'s (1990) SCM in assuming that the lowest level superordinate capturing premise and conclusion categories is activated.

Another way of thinking about relevance is to suggest that, when given an argument to evaluate, participants ask themselves why this particular premise (and not some other one) is given for the particular conclusion under consideration. For example, suppose one is given the premise that "Skunks have property X." According to the relevance framework, good candidates for what property X might be related to are features that are distinctive of skunks; that is, features that skunks have that similar mammals such as squirrels or muskrats do not have. Two possibilities that immediately come to mind are that they are striped and that they can create a very strong odor. The conclusion category may act as a further important constraint on assumed relevance. For example, if the conclusion is that "Zebras also have property X," then it becomes plausible that property X is related to being striped and that the argument should be considered to be at least moderately strong. If the conclusion were instead, "Onions have property X," a participant who assumes that the experimenter is following a relevance principle should be more likely to assume that property X refers to odor rather than stripedness. Note also, that the argument going from zebras to skunks may be stronger than one going from skunks to zebras because skunks have two salient features and zebras may have only one, being striped (though perhaps being an African mammal is another one).

Summing up so far: The relevance framework suggests two processing principles. One is that distinctive properties of premise categories are candidates for providing the relevant basis for induction. (To be sure, particular predicates can support or undermine candidates for relevance; if the premise were "Skunks weigh more than

Martians," then stripedness and odor clearly would be irrelevant.) The second idea is that comparing the premise and conclusion categories acts as a further constraint on relevance by either reinforcing or undermining candidates for relevance on the basis of the premise categories considered by themselves. We further suggest that the same comparison process is used (for related ideas on the importance of comparison processes, see also Hahn & Chater, 1997; Medin, Goldstone, & Gentner, 1993) if there is more than one premise category (finding out that both skunks and onions have property X might make one fairly sure that property X is linked to having a strong odor) or even more than one conclusion category.

Although relevance often may involve categories, unlike the SCM, relevance theory is not restricted to them. Instead, nontaxonomic categories, properties, and even thematic relationships may form the basis for categorical induction. For example, a premise statement that kangaroos have some property may trigger mammals as the relevant category but it may also lead to *Australian animals* or *mammals with pouches* as the relevant superordinate. Another difference from the SCM is that, for a given rank or level, some superordinates may be more informative (salient) than others. To continue the prior example, *kangaroo* should be more likely to activate Australian animals than *muskrat* should be to activate North American animals (at least for participants from universities in the United States). That is, the fact that Australia is more distinctive with respect to the animals that inhabit it should make it more likely that Australian animals would be seen as a relevant category for induction. A third difference from the SCM (and the feature-based and Bayesian approaches as well) is that premises and conclusions may be linked through causal reasoning. Shortly we will amplify this point.

The principle that premises are compared with each other to determine relevant categories and properties is similar in spirit to the McDonald et al. (1996) hypothesis assessment model (see also Gentner & Medina, 1998, and Blok & Gentner, 2000, for related ideas concerning premise comparison). McDonald et al.'s efforts and experiments were directed at linking category-based induction with other research in the hypothesis-testing tradition. They view premises of arguments as triggering hypotheses that fix the scope for induction. Our goals are tied more directly to manipulating effect and effort, in most instances through comparison processes used to fix relevance. Although one can certainly cast the outcome of such comparison processes as hypotheses, the relevance framework leads to a new set of predicted induction phenomena and a different slant on the effects described by Osherson et al. (1990). Before bringing out these predictions, we turn first to the role of causal reasoning in induction.

Causal relations. Consider the following inductive argument: "Grass has Enzyme X, therefore cows have Enzyme X." Osherson et al.'s SCM would assess this argument in terms of the similarity of grass to cows and the coverage of grass in the lowest level superordinate category that includes cows and grass (living things). Consequently, the argument strength should be low, according to the SCM. As mentioned earlier, our relevance framework employs a notion of similarity constrained by comparison processes and allows for thematic or causal relations to affect induction. For this example, people are likely to retrieve a linkage between cows and grass, namely that cows eat grass. This knowledge invites the causal inference that Enzyme X might be transmitted from grass to cows by ingestion. Consequently, the argument about grass and cows should seem to be strong (and relevant). In brief, by selecting categories (and properties) about which undergraduates may have relevant background information, we may be able to vary what Van der Henst et al. (2002) call effect. Biological experts or Itza' Maya have a great deal of background knowledge such that arguments involving biological categories will naturally produce large effects, often in terms of causal relations

Manipulating effort. The relevance framework suggests some straightforward ways of varying effort to affect inductive confidence. First, with respect to comparison processes, additional premise (and conclusion) categories can be used to reinforce or undermine the ease and likelihood of seeing some property as relevant. Consider again an argument going from skunks to zebras. Adding the premise that striped bass also have the property in question should make it easier to conclude that the property in question is linked to having stripes and therefore applies to zebras. In fact, relevance may even override normative considerations. Suppose we compare an argument going from skunks to zebras with an argument going from skunks to striped bass and zebras. It is possible that the comparisons of conclusions and premise will so boost confidence that the relevant basis for induction has been identified that the argument with the conjunctive conclusion will be seen as stronger than the one with a single conclusion category.

A similar contrast involving effort is readily available for causal scenarios. Consider the argument that "Grass has Enzyme X and therefore humans also have Enzyme X." A potential causal linkage may be less transparent than for the case with the same premise but where the conclusion is that "therefore cows and humans have Enzyme X." The addition of cows (and the accessible knowledge that humans drink the milk of cows) may make it easy to create or retrieve a causal linkage from grass to humans and lead to the conjunctive conclusion's being evaluated as stronger than the single conclusion (obviously this prediction has to be evaluated in a between-participants design). Work on the availability heuristic in relation to causal schemas (e.g., Tversky & Kahneman, 1974) also suggests that causal relations will more readily affect inductive confidence when the cause is the premise and the effect is the conclusion than for the reverse order. In short, the relevance framework leads to a number of novel, and in some cases nonnormative, predictions.

A note on blank versus nonblank properties. Osherson et al. (1990) define "blank" properties as those for which participants have few beliefs and which are unlikely to evoke beliefs that cause one argument to have more strength than another. For example, most people have no a priori opinion about whether robins or ostriches "require biotin for protein synthesis." The SCM works best in explaining induction phenomena that involve blank properties. Indeed, in order to account for arguments with nonblank predicates, Smith, Shafir, and Osherson (1993) showed that a number of additional processing assumptions needed to be added to the similarity coverage framework.

The distinction between blank and nonblank properties, however, is not always clear-cut. Heit and Rubinstein (1994), for example, showed that undergraduates generalized abstract behavioral properties in a different pattern than they did abstract physiological properties (*behavioral* similarity had a greater effect in the former condition).

The relevance framework suggests that interactions between premise and conclusion categories or between premise categories may evoke beliefs about even the blankest of blank properties. Suppose we modify our earlier argument to the more abstract form "Grass has some Property X, therefore cows have Property X." It seems likely that people will still entertain the idea that X may be something that can be transmitted from grass to cows. Even an isolated premise may evoke certain beliefs. For example, the premise "Penguins have Property Y" is likely to trigger expectations about Property Y that render penguins a relevant, informative premise category. In this case, people might expect that Property Y is an adaptation to an antarctic environment or linked to swimming and waddling rather than flying, or they may even entertain the abstract belief that the property must be unusual because penguins are unusual birds.

The Osherson et al. (1990) strategy of using abstract, unfamiliar properties is very effective for seeing what other information people bring to a task to determine relevance and draw inferences. In addition, the absence of a strong borderline between blank and nonblank properties suggests that we should be able to develop models of induction that address a range of specificity and familiarity of predicates. One advantage of the relevance framework is that it does not require blank predicates (nor do the feature-based Bayesian or hypothesis models).

Summary of relevance framework predictions. The specific assumptions we have been developing can be seen as implementing the general principle of relevance for the case of category-based induction. The main ideas are that induction involves a search for relevance and that candidates for relevance are salient properties and (causal) relations. Most important for our present purposes is the idea that effect and effort can be manipulated by using undergraduates' background knowledge and by introducing additional premises and/or conclusions that increase or decrease effort.

The key experimental manipulations in our studies are as follows: (1) the strengthening and weakening of candidates for relevance via property reinforcement, and (2) scenario (causal) instantiation and manipulations of effort designed to increase or decrease access to causal associations. So far we have kept our descriptions general, rather than adopted a specific, quantitative model, mainly because the framework leads to a number of qualitative predictions and would be consistent with a large set of specific instantiations. In addition, the determination of relevance may require fairly flexible processing principles. For example, rather than occurring in a fixed order, comparisons may be guided by the strength of correspondence between the representations associated with a comparison (as Goldstone & Medin, 1994, assume), which could alter the comparisons themselves. Consider the following argument: "Polar bears have CO3 and walruses have CO3, therefore polar bears have CO3." In this case, the excellent correspondence between the first premise and the conclusion (viz. identity) might well preempt the comparison of premises to each other and lead directly to the inference that the argument is perfectly strong. For the present, we will restrict ourselves to these general ideas about property weakening and strengthening. In the next section, we will present predictions/phenomena tied to the relevance framework.

Predictions of the Relevance Framework

In this paper, we focus on phenomena associated with effort and effect by varying causal scenarios and property reinforcement. In general, our strategy is to develop items for which relevance-based reasoning makes predictions that either run counter to those of other models (e.g., nondiversity) or that contradict normative judgments (e.g., conclusion conjunction fallacy).

We now examine five phenomena involving *causal scenarios*. The first of these, *causal asymmetry*, predicts that an inference from A to B will be rated as stronger than an inference from B to A when a relevant causal scenario about the transmission of a property from A to B is salient. For example, GAZELLES/LIONS should be stronger than LIONS/GAZELLES, because it is easier to imagine a property being transmitted from gazelles to lions via the food chain than vice versa. In other words, premise order affects the effort needed for activation of a causal scenario. Next, *causal violation of similarity* and *causal nondiversity* pit causal relations directly against predictions derived from other models. We will examine whether causal relations might override similarity by strengthening inferences between dissimilar premise and conclusion categories (as in the GRASS/COWS example discussed above), and whether causal relations might override diversity by weakening otherwise diverse premises. For example, (ROBINS + WORMS)/GOLDFISH may be stronger than (ROBINS + IGUANAS)/GOLDFISH in terms of the sheer coverage of the inclusive category *animals*, but the salient fact that robins eat worms might make it plausible that they would share a property not generally shared, and

this would therefore weaken the former argument. Such effects of causal scenarios have been found frequently in expert populations (see, e.g., Proffitt et al., 2000); our main new contribution is to show that these effects can be predicted in advance and that they can be demonstrated in undergraduates, as long as relevant background knowledge is pinpointed (in the relevance framework, one is selecting items for which background knowledge will produce larger effects).

The fourth causal phenomenon involves a case in which salient causal scenarios lead to logically nonnormative judgments: the *causal conjunction fallacy*. Normatively, adding a conclusion category to an argument should never strengthen it. In contrast, the causal conjunction fallacy predicts that adding a conclusion category that strengthens a causal link between premise and conclusion might strengthen the argument. For example, GRAIN/(MICE + HAWKS) might be considered stronger than GRAIN/HAWKS (a logical fallacy), because it may foster a causal link from grain to mice to hawks. In terms of the relevance framework, the addition of the MICE premise reduces the effort needed for one to develop the causal linkage from grain to hawks.

Finally, *causal nonmonotonicity* predicts that adding a premise category might weaken an argument if it highlights a causal relation between premise categories not shared by the conclusion category. For example, HUMANS/OAKS might be considered stronger than (HUMANS + MOSQUITOES)/OAKS, because mosquitoes might plausibly transmit a property to humans but not to oaks. The SCM allows for weakening only if the additional premise increases the abstractness of the lowest level superordinate category that covers premises and conclusion and the (original form of the) FBIM does not allow for nonmonotonicities at all.

We also investigate three phenomena involving *property reinforcement* that parallel the phenomena presented above for causal scenarios. *Nondiversity via property reinforcement* suggests that if an otherwise diverse set of premises shares a salient property not shared by the conclusion category, the reinforcement of the property might weaken that argument relative to a related argument with less diverse premises. This is not unlike Tversky's (1977) well-known diagnosticity principle for similarity judgments. For instance, in the SCM framework, the argument (PIGS + CHICKENS)/COBRAS is assumed to be stronger via coverage than the argument (PIGS + WHALES)/COBRAS (because pigs are mammals and chickens are birds, and therefore cover the inclusive category *animal* better than pigs and whales, two mammals). However, pigs and chickens—unlike cobras—are farm animals and are raised for food; these properties might weaken the argument. *Conclusion conjunction fallacy via property reinforcement* predicts that adding a conclusion category that reinforces a property shared by premise and conclusion might strengthen the argument. For example, DRAFT HORSES/(RACE HORSES + PONIES) might be considered stronger than (DRAFT HORSES)/PONIES. As we shall see, the significance of this effect is less that it is nonnormative than that it contrasts the relevance framework with alternative models.

Nonmonotonicity via property reinforcement predicts that adding premise categories might weaken an argument if the added categories reinforce a property shared by all premise categories but not by the conclusion category. For example, (BROWN BEARS)/BUFFALO might be considered stronger than (BROWN BEARS + POLAR BEARS + GRIZZLY BEARS)/BUFFALO because in the latter case participants may be thinking that the relevant conclusion categories is bears.

In general, our strategy is to present participants with sets of arguments in which relevance theory makes predictions that run counter to one or more other models. To the degree that relevance-based arguments are rated as stronger, the relevance framework is supported.

METHOD

Participants

The study was carried out at three different international sites. At the U.S. site, the participants were 30 male and female psychology students enrolled in introductory psychology courses who received course credit for their involvement. The participants at this site were administered items relating to all of the eight phenomena under investigation. At the Australian site, the participants were 138 male and female undergraduates from 18 to 46 years old, enrolled in introductory or senior-level psychology courses. All received course credit for their involvement. Ninety-three participants were administered items relating to causal asymmetry, causal violation of similarity, causal conjunction fallacy, and conjunction fallacy via property reinforcement; the remaining 45 completed items relating to causal nondiversity, causal nonmonotonicity, nondiversity via property reinforcement, and nonmonotonicity via property reinforcement. At the Belgian site, 36 first-year students of the Faculties of Law and of Economics, from 18 to 20 years old, participated for course credit. All participants at this site were administered items relating to causal nondiversity, causal nonmonotonicity, nondiversity via property reinforcement, conjunction fallacy via property reinforcement, and nonmonotonicity via property reinforcement. At both the Australian and the Belgian sites, equal numbers of participants were randomly assigned one of three forms of an induction questionnaire containing different versions of the induction items as described below. At the U.S. site, all participants completed all versions of each item, presented in random order with the constraint that related items were presented with at least two intervening unrelated arguments.

Materials

A series of inductive reasoning items was constructed; these were thought to tap a number of novel phenomena that follow from the relevance principle. In all cases, the properties attributed to the premises were fictitious and assumed to be "blank" in that they were unlikely to evoke beliefs that would cause the selective strengthening of a particular premise or conclusion argument. The fictitious properties were therefore labeled with either a "nonsense" property (e.g., "contains retinum") or an uninformative symbol (e.g., "Property X12"). A different blank property was used for each item. Two or three versions of each item type were constructed, depending on the specific phenomenon being tested, as is shown in Table 1.

Procedure

The participants were tested individually in a quiet room. They were told that they would be asked to judge the strength of a num-

Table 1
Mean Ratings for Hypothesized Stronger and Weaker Versions of Each Relevance-Based Phenomenon Showing Premise/Conclusion Categories From Each Testing Site

Item No.	Strong	*M*	*SD*	Weak	*M*	*SD*
Note: All Belgian items have been translated from the original Flemish.						
Causal Asymmetry: U.S.						
1.	GAZELLES/LIONS	5.03	2.76	LIONS/GAZELLES	4.63	2.61
2.	FLOWERS/BEES	5.00	2.82	BEES/FLOWERS	4.90	2.66
3.	CARROTS/RABBITS	4.57	2.92	RABBITS/CARROTS	3.90	2.95
Causal Asymmetry: Australia						
1.	FLOWERS/BEES	5.83	2.17	BEES/FLOWERS	4.50	2.02
2.	CARROTS/RABBITS	4.26	2.61	RABBITS/CARROTS	3.87	2.32
3.	BEETLES/CROWS	3.71	1.99	CROWS/BEETLES	2.57	1.63
4.	ANTELOPES/LIONS	5.07	2.24	LIONS/ANTELOPES	3.78	2.06
Causal Violation of Similarity: U.S.						
1.	WATER/TULIPS	5.23	2.67	(SPRUCE TREES)/TULIPS	5.50	2.26
2.	ACORNS/SQUIRRELS	4.47	2.76	ELK/SQUIRRELS	4.40	2.34
3.	BANANAS/MONKEYS	4.93	2.74	MICE/MONKEYS	3.87	2.32
Causal Violation of Similarity: Australia						
1.	BANANAS/MONKEYS	5.19	2.49	MICE/MONKEYS	4.10	2.01
2.	DAFFODILS/BEES	4.77	2.73	DAFFODILS/GUM TREES	3.88	2.03
3.	MOSQUITOS/PEOPLE	3.55	2.29	MOSQUITOS/BEES	5.23	2.43
Causal Nondiversity: U.S.						
1.	(ROBINS + IGUANAS)/GOLDFISH	3.33	2.02	(ROBINS + WORMS)/GOLDFISH	3.67	2.44
2.	(CATS + RHINOS)/LIZARDS	3.90	2.48	(CATS + SPARROWS)/LIZARDS	3.50	2.17
3.	(FLEAS + BUTTERFLIES)/SPARROWS	4.23	2.54	(FLEAS + DOGS)/SPARROWS	3.50	2.27
Causal Nondiversity: Australia						
1.	(SPARROWS + DOGS)/(LIVING THINGS)	5.88	2.29	(SPARROWS + SEEDS)/(LIVING THINGS)	3.65	2.32
2.	(KOALAS + WOLVES)/(LIVING THINGS)	4.74	2.02	(KOALAS + GUM TREES)/(LIVING THINGS)	4.18	1.94
3.	(RABBITS + ZEBRAS)/(LIVING THINGS)	5.06	2.30	(RABBITS + LETTUCE)/(LIVING THINGS)	6.00	2.59
Causal Nondiversity: Belgium						
1.	(SPARROWS + DOGS)/GOLDFISH	4.24	2.28	(SPARROWS + SEEDS)/GOLDFISH	2.88	2.32
2.	(HORSES + ANTS)/STARLINGS	4.17	2.38	(HORSES + GRASS)/STARLINGS	3.06	2.29
3.	(RABBITS + ZEBRAS)/BUTTERFLIES	2.61	1.91	(RABBITS + CARROTS)/BUTTERFLIES	2.22	1.48
Causal Conjunction Fallacy: U.S.						
1.	FLEAS/(DOGS + HUMANS)	4.47	2.46	FLEAS/DOGS	5.47	2.74
				FLEAS/HUMANS	5.53	2.49
2.	GRAIN/(MICE + OWLS)	4.10	2.58	GRAIN/MICE	3.50	2.58
				GRAIN/OWLS	2.70	2.15
3.	GRASS/(COWS + HUMANS)	4.17	2.57	GRASS/COWS	4.70	2.45
				GRASS/HUMANS	2.97	2.33
Causal Conjunction Fallacy: Australia						
1.	LEAVES/(DEER + WOLVES)	2.81	1.49	LEAVES/DEER	4.90	2.44
				LEAVES/WOLVES	1.63	0.89
2.	GRAIN/(MICE + HAWKS)	5.06	2.34	GRAIN/MICE	3.66	2.47
				GRAIN/HAWKS	3.80	2.35
3.	GRASS/(COWS + HUMANS)	4.90	2.44	GRASS/COWS	5.20	2.43
				GRASS/HUMANS	2.31	1.42

Table 1 (Continued)

Item No.	Strong	*M*	*SD*	Weak	*M*	*SD*
Causal Nonmonotonicity: U.S.						
1.	HUMANS/OAKS	2.77	1.96	(HUMAN + MOSQUITOS)/OAKS	2.63	1.75
	MOSQUITOS/OAKS	2.53	2.21			
2.	SAND/CLAY	5.67	2.54	(SAND + GLASS)/CLAY	5.27	2.43
	GLASS/CLAY	4.27	2.63			
3.	CHICKENS/PELICANS	5.20	2.46			
	(CHICKEN HAWKS)/PELICANS	5.43	2.19	(CHICKENS + CHICKEN HAWKS)/PELICANS	5.30	2.07
Causal Nonmonotonicity: Australia						
1.	ANTEATERS/SNAKES	2.94	1.71	(ANTEATERS + ANTS)/SNAKES	3.65	1.97
	ANTS/SNAKES	2.35	1.22			
2.	MICE/GIRAFFES	3.06	1.95	(MICE + CATS)/GIRAFFES	3.53	1.66
	CATS/GIRAFFES	3.41	1.77			
3.	CHICKENS/PELICANS	3.94	1.82	(CHICKENS + CHICKEN HAWKS)/PELICANS	4.94	1.71
	CHICKEN HAWKS/PELICANS	5.29	2.23			
4.	LIONS/MICE	3.35	2.09	(LIONS + ANTELOPES)/MICE	4.65	1.97
	ANTELOPES/MICE	2.88	1.32			
5.	FOXES/ELEPHANTS	2.53	1.77	(FOXES + RABBITS)/ELEPHANTS	3.41	2.37
	RABBITS/ELEPHANTS	2.82	1.78			
Causal Nonmonotonicity: Belgium						
1.	ANTEATERS/SNAKES	3.42	1.83	(ANTEATERS + ANTS)/SNAKES	3.08	1.73
	ANTS/SNAKES	3.25	1.71			
2.	MICE/GIRAFFES	5.16	2.12	(MICE + CATS)/GIRAFFES	5.50	2.11
	CATS/GIRAFFES	5.25	1.60			
3.	CHICKENS/PELICANS	4.75	1.42	(CHICKENS + CHICKEN HAWKS)/PELICANS	4.17	1.85
	CHICKEN HAWKS/PELICANS	3.08	2.35			
Nondiversity via Property Reinforcement: U.S.						
1.*	(PIGS + WHALES)/COBRAS	3.57	2.31	(PIGS + CHICKENS)/COBRAS	3.03	2.19
2.*	(BATS + ELEPHANTS)/ALLIGATORS	4.20	2.51	(BATS + ROBINS)/ALLIGATORS	2.93	1.84
3.	(PENGUINS + EAGLES)/CAMELS	3.43	2.53	(PENGUINS + POLAR BEARS)/CAMELS	3.50	2.36
*Item not analyzed because taxonomically more similar pair was rated as less similar, thereby invalidating item.						
Nondiversity via Property Reinforcement: Australia						
1.	(SKUNKS + DEER)/ANIMALS	5.82	1.81	(SKUNKS + STINK BUGS)/ANIMALS	3.68	1.97
2.	(KANGAROOS + ELEPHANTS)/ANIMALS	5.76	2.55	(KANGAROOS + FROGS)/ANIMALS	6.76	0.83
3.	(POLAR BEARS + ANTELOPES)/ANIMALS	4.94	2.01	(POLAR BEARS + PENGUINS)/ANIMALS	4.41	2.16
4.	(CAMELS + RHINOS)/MAMMALS	4.47	2.12	(CAMELS + DESERT RATS)/MAMMALS	4.38	2.27
5.	(CHIMPANZEES + COWS)/MAMMALS	4.71	1.96	(CHIMPANZEES + DOLPHINS)/MAMMALS	6.38	2.16
Nondiversity via Property Reinforcement: Belgium						
1.	(PIGS + WHALES)/COBRAS	4.60	1.67	(PIGS + CHICKENS)/COBRAS	3.36	2.06
2.	(CAMELS + RHINOS)/TOUCANS	4.44	2.38	(CAMELS + DESERT RATS)/TOUCANS	2.94	1.73
3.	(PENGUINS + FROGS)/GIRAFFES	3.89	1.94	(PENGUINS + POLAR BEARS)/GIRAFFES	2.83	1.89
Conjunction Fallacy via Property Reinforcement: U.S.						
1.	CHICKENS/(COWS + PIGS)	6.33	2.48	CHICKENS/COWS	4.90	2.16
				CHICKENS/PIGS	5.97	2.40
2.	PASTA/(RICE + POTATOES)	7.07	2.16	PASTA/RICE	6.43	2.19
				PASTA/POTATOES	6.47	2.30
3.	FERRARIS/(ROLLS ROYCES + BMWS)	6.83	1.93	FERRARIS/(ROLLS ROYCES)	6.93	2.21
				FERRARIS/BMWS	6.77	1.89

Table 1 (Continued)

Item No.	Strong	*M*	*SD*	Weak	*M*	*SD*
Conjunction Fallacy via Property Reinforcement: Australia						
1.	ANDEAN PEOPLE/(HIMALAYAN PEOPLE + ALPINE PEOPLE)	6.35	1.97	(ANDEAN PEOPLE)/(HIMALAYAN PEOPLE)	4.88	2.09
				(ANDEAN PEOPLE)/(ALPINE PEOPLE)	4.65	2.12
2.	KANGAROOS/(WOMBATS + KOALAS)	6.53	1.66	KANGAROOS/WOMBATS	5.88	2.47
				KANGAROOS/KOALAS	6.00	1.80
Conjunction Fallacy via Property Reinforcement: Belgium						
1.	CHICKENS/(COWS + PIGS)	7.25	1.29	CHICKENS/COWS	6.17	2.12
				CHICKENS/PIGS	6.67	2.06
2.	(DRAFT HORSES)/(RACE HORSES + PONIES)	4.92	1.68	(DRAFT HORSES)/(RACE HORSES)	3.75	1.96
				(DRAFT HORSES)/PONIES	3.75	1.66
3.	(ANDEAN PEOPLE)/(HIMALAYAN PEOPLE + ALPINE PEOPLE)	7.08	1.44	(ANDEAN PEOPLE)/(HIMALAYAN PEOPLE)	6.75	1.06
				(ANDEAN PEOPLE)/(ALPINE PEOPLE)	7.33	1.30
Nonmonotonicity via Property Reinforcement: U.S.						
1.	(MARKETING MAJORS)/(ENGLISH MAJORS)	5.20	2.31	(MARKETING MAJORS + FINANCE MAJORS + MANAGEMENT MAJORS)/(ENGLISH MAJORS)	5.03	2.57
2.	(POISON IVY)/DANDELIONS	5.03	2.36	(POISON IVY + POISON OAK + POISON + SUMAC)/DANDELIONS	4.23	2.66
3.	(BROWN BEARS)/BUFFALO	5.57	2.34	(BROWN BEARS + POLAR BEARS + GRIZZLY BEARS)/BUFFALO	4.83	2.45
Nonmonotonicity via Property Reinforcement: Australia						
1.	(RED GUM TREES + GHOST GUMS)/(MAPLE TREES)	4.76	1.89	(RED GUM TREES + GHOST GUMS + BLUE GUMS + FLOODED GUMS)/(MAPLE TREES)	3.76	2.25
2.	(INDONESIANS + VIETNAMESE)/NORWEGIANS	3.94	2.27	(INDONESIANS + VIETNAMESE + MALAYSIANS + CAMBODIANS)/NORWEGIANS	2.82	1.94
3.	(BROWN BEARS + POLAR BEARS)/GOATS	3.24	1.46	(BROWN BEARS + POLAR BEARS + BLACK BEARS + GRIZZLY BEARS)/GOATS	1.88	1.22
4.	(SWEDES + FINNS)/ITALIANS	4.50	2.14	(SWEDES + FINNS + DANES + NORWEGIANS)/ITALIANS	5.53	2.43
5.	VIOLINISTS/DRUMMERS	5.06	2.45	(VIOLINISTS + DOUBLE-BASS PLAYERS + CELLISTS)/DRUMMERS	6.12	2.23
6.	(GERMAN SHEPHERDS)/POODLES	4.53	2.09	(GERMAN SHEPHERDS + DOBERMANS + ROTTWEILERS)/POODLES	3.41	2.27

ber of arguments and were given a practice example of a strong and a weak inductive inference. They were then given a questionnaire containing all versions (U.S.) or one version (Australia, Belgium) of each of the test items. For each item, the participants were presented with an argument involving the projection of a blank property from one or more premise categories to one or more conclusion categories. These arguments followed the general form "[Premise category(ies)] have Property X, therefore, [Conclusion Category(ies)] have Property X." The specific premise and conclusion categories used for each of the relevance phenomena are given in Table 1. For example, the first item listed for the U.S. version of causal asymmetry in the table was presented as "Gazelles have Property X12, therefore, Lions have Property X12." The participants were asked to rate how "strong or convincing" they thought each argument was on a 9-point scale (1 = *weak/not very convincing*, 9 = *strong/very convincing*). Table 1 shows which version of each item was predicted to be rated as stronger according to the relevance principle. The order of presentation of all items within each questionnaire was randomized over the different participants at all testing sites.

The participants were also asked to provide a written justification for each of their ratings of argument strength. At the U.S. and Belgian sites, these justifications were required immediately after each item rating, whereas at the Australian site, the participants provided justifications after they had completed all the item ratings. There was no time limit on the completion of the item ratings or justifications. The entire procedure took 30–40 min per participant.

Item Validation

Data were collected from a further sample of 20 participants (all of them research assistants in the Department of Psychology, University of Leuven) to confirm our intuitions about similarity for the violation of similarity and the nondiversity items. Each of these participants was shown 21 sets of pairs of premise categories, corresponding to each of the causal violation of similarity and diversity items in Table 1, except for the first item from the Australian nondiversity via property reinforcement, which was dropped because of the unfamiliarity of Belgian participants with the categories involved. Their task was to choose which pair of categories was most similar, basing judgments on "general similarity, not on associations." The items were presented in two different random orders, each presented to 10 subjects.

RESULTS

The mean ratings for the different versions of each item are presented in Table 1. For every phenomenon, the rated argument strength for all items in all sites where the phenomenon was included were analyzed together in an analysis of variance, with items and the different item versions as independent variables. The data were analyzed using both the items variable and the item versions variable as between-subjects variables. Note that similar items were sometimes used in different sites. However, owing to linguistic and cultural differences, every item in every site is considered a separate item. For the phenomena with a significant interaction between items and item versions, all items were also analyzed individually in separate one-way analyses of variance with the different item versions as the independent variable. In these follow-up analyses, the power to reject the null hypothesis was considerably lower than in the overall analyses.

Justifications were coded by simply noting whether or not the participants, when asked to explain their responses, mentioned the target relation for each item. For example, for GAZELLES/LIONS and LIONS/GAZELLES, if participants explained their responses by saying "Lions eat gazelles," they were scored as having mentioned the target relation. If they said "Both live in Africa," they were not scored as having mentioned the target relation. A summary of the results of this coding is presented in Table 2. The Relevance column represents the proportion of times the target relation was explicitly mentioned for the items hypothesized to highlight that relation; the Competitor column represents the proportion of responses mentioning the target relation for the comparison or control items. For each phenomenon, relatively high frequencies of mentioning the target relations for items where we attempted to vary effort and/or effect suggest that target causal relations or properties did influence reasoning.

Table 2
Proportion of Justifications Mentioning Target Relation for Relevance and Competing Items

Phenomena	Relevance	Competitor
Causal Scenarios		
Asymmetry	.70	.49
Violation of similarity	.71	.09
Nondiversity	.32	.03
Conjunction fallacy	.57	.40
Nonmonotonicity	.20	.04
Mean	.50	.22
Property Reinforcement		
Nondiversity	.46	.08
Conjunction fallacy	.75	.58
Nonmonotonicity	.57	.09
Mean	.60	.25

First, the results from the causal scenario items will be discussed. Next, the data from the property reinforcement items will be presented. For phenomena involving similarity and diversity, the validity of the items depends on the validity of our intuitions about similarity among premise and conclusion categories. In these cases, we also report results of the auxiliary similarity task described above. We simply tallied the number of participants out of 20 who chose the predicted pair of categories as more similar for each item, and computed the corresponding binomial probability (see Table 3). Justifications were not analyzed statistically, but will be discussed with their respective phenomena.

Causal Scenario Items

Causal asymmetry. For these items, the prediction was that arguments would be rated as stronger when a salient causal link flowed from premise to conclusion than when causal direction flowed from conclusion to premise. As predicted, the overall analysis yielded a significant difference between the two versions of the items [$F(1,413) = 5.81$, $p < .05$]. The mean rating of the versions in line with the causal scenario was 4.69. The mean rating of the reversed version was 4.12. Moreover, examination of Table 2 reveals that although the causal relations were salient in the competitor items (49% of the

Table 3
Item Analysis for Selected Phenomena: Number of Participants Choosing Each Category Pair as More Similar

Causal Violation of Similarity Items			
Similar Pair		Causal Pair	
spruce tree & tulip	20*	water & tulip	0
elk & squirrel	19*	acorn & squirrel	1
mouse & monkey	19*	banana & monkey	1
daffodil & gum tree	18*	daffodil & bee	2
mosquito & bee	18*	mosquito & people	2
Causal Nondiversity Items			
Similar Pair		Diverse Pair	
robin & iguana	16*	robin & worm	4
cat & rhino	15*	cat & sparrow	5
flea & butterfly	20*	flea & dog	0
sparrow & dog	19*	sparrow & seeds	1
koala & wolf	20*	koala & gum tree	0
rabbit & zebra	18*	rabbit & lettuce	2
horse & ant	18*	horse & grass	2
rabbit & zebra	18*	rabbit & carrot	2
Nondiversity via Property Reinforcement			
Similar Pair		Diverse Pair	
pig & whale	7	pig & chicken	13
bat & elephant	2	bat & robin	18*
penguin & eagle	15*	penguin & polar bear	5
skunk & deer	–	skunk & stink bug†	–
kangaroo & elephant	15*	kangaroo & frog	5
polar bear & antelope	15*	polar bear & penguin	5
camel & rhino	19*	camel & desert rat	1
chimpanzee & cow	17*	chimpanzee & dolphin	3

*Value differs from chance (.50) by binomial $p < .05$. †Similarity comparisons were not collected for this item because of lack of familiarity of Belgian participants with skunks and stink bugs.

participants mentioned the target relation for the reversed items), they were even more so for the items in which cause flowed from premise to conclusion (70% mentioned target relations). The item × item version interaction was not significant.

Causal violation of similarity. For these items, we pitted an argument with a salient causal link between a dissimilar premise and a conclusion against an argument with a premise that was much more similar to the conclusion but lacked a salient causal link, in effect pitting causal connections versus similarity (e.g., BANANAS/MONKEYS vs. MICE/MONKEYS). Item analysis confirmed our intuitions; similar premise–conclusion pairs were chosen as more similar than causally related premise–conclusion pairs for all items (see Table 3). Although the mean rating of the causal scenario version (4.69) was higher than the mean rating for the noncausal version (4.50) (where the similarity of the premise and the conclusion categories presumably was larger), the overall analysis showed that this difference was not significant. An examination of Table 3 confirms that causal information was salient, as 71% of the participants mentioned target relations.

The item × item version interaction was significant [$F(5,353) = 2.99, p < .05$], suggesting that the phenomenon did not work equally well in all items. Four out of the six individual items (Items 2 and 3 from the American data and Items 1 and 2 from the Australian data) yielded higher mean ratings for the causal scenario versions, but the difference never reached significance. Item 3 of the Australian data resulted in a significant difference ($p < .05$) between the two versions, but this difference was opposite to the predictions. Overall, responses to these items suggest that causal connections were as inductively potent as similarity but not reliably more so.

Causal nondiversity. The diversity phenomenon predicts that the more diverse the premises, the stronger the argument. In contrast, we predicted that a more diverse pair of premises might be rated weaker if there existed a salient causal link between the premises (e.g., HORSES + GRASS) that made it plausible that they would share a salient property that was not likely to be shared by other members of the superordinate category. Item analysis again validated the items; nondiverse premise pairs were chosen as more similar than diverse premise pairs for all items (see Table 3). In the overall analysis, the nondiverse version yielded a significantly higher mean (4.10) than the diverse but causally linked version (3.89) [$F(1,423) = 4.68, p < .05$], suggesting that causal reasoning can undermine diversity. And although not overwhelming, 32% of participants did mention the target relations in their justifications, as opposed to 3% in the control.

The interaction between item and item version was also significant [$F(8,423) = 2.23, p < .05$]. In the follow-up analyses of the individual items, the mean rating for the nondiverse version was higher for six out of nine items. The two versions were significantly different only for Item 1 of the Australian data. Overall, it appears that causal relations can lead to a preference to reason from nondiverse premises.

Causal conjunction fallacy. For these items, we predicted that arguments with two-category conclusions might be rated stronger than arguments with one-category conclusions (a conjunction fallacy) if the added conclusion category reinforced a salient causal chain connecting all three (e.g., because of the salient causal food chain, GRAIN/(MICE + HAWKS) might be seen as stronger than GRAIN/MICE or GRAIN/HAWKS). The overall analysis yielded a significant difference between the three versions of the causal conjunction fallacy items [$F(2,531) = 27.61, p < .01$]. However, in this analysis, the difference between the two single-conclusion versions is irrelevant. Therefore, a contrast that compared the mean of the single-conclusion versions with the double-conclusion versions was formulated. This is a conservative test, in that it is nonnormative for the double-conclusion to be stronger than either of the single-premise arguments. (Testing the double-conclusion against the weaker of the two single-conclusion arguments would introduce a [modest] potential bias that could capitalize on random error if there were no true difference.) As hypothesized, the mean rating for the single-conclusion versions (3.65) was found to be significantly lower than the mean for the double-conclusion versions (4.25) [$F(1,531) = 8.29, p < .01$]. Again, justifications suggest that these items worked via the hypothesized mechanism; the additional conclusion

category led to an increase from 40% to 57% target justifications.

The item × item version interaction was again significant [$F(10,531) = 2.78$, $p < .01$], which requires more detailed analyses at the level of the individual items. In the analyses of the individual items, the rating for the double-conclusion version was higher than the rating for one of the single-conclusion items in Items 1 and 3 of the American data, but the difference was not significant. For Item 2 of the American data, the rating for the double-conclusion version was higher than that for both single-conclusion versions, but again the difference was not significant. Items 2 and 3 of the Australian data yielded significant differences between the mean single-conclusion and the mean double-conclusion ratings ($p < .05$). In the first Australian item, the double-conclusion ratings were significantly higher than the ratings for one of the single-conclusion versions.

Causal nonmonotonicity. For these items, we predicted that adding a category to the premise of an argument that was causally related to the original premise but not the conclusion might weaken the argument (e.g., [HUMANS + MOSQUITOS]/OAKS might be weaker than HUMANS/OAKS because of the causal scenario linking humans and mosquitos). In an overall analysis, the three versions of the item differed significantly [$F(2,600) = 3.23$, $p < .05$]. However, a contrast that compared the mean rating for the single-premise category version (3.79) with that for the double-premise category version (4.19) showed the latter to be significantly higher [$F(1,600) = 4.89$, $p < .05$], which is contrary to the hypothesized result. The item × item version interaction was not significant. Perhaps these items failed because the causal scenario was not sufficiently salient. Table 2 reveals that only 20% of the participants mentioned the target relation for the two-premise items.

In summary, the hypothesized causal asymmetry, causal conjunction fallacy, and causal nondiversity phenomena were clearly supported by the data. The results of the causal violation of similarity items were as predicted, but the difference was not significant. Finally, we found no evidence for the hypothesized nonmonotonicity. Moreover, justifications suggest that—as predicted—the success of these phenomena was due to the salience of causal relations among categories.

Property Reinforcement Items

The rationale behind these items is that providing a number of instances as premises or conclusions that reinforce a particular property as relevant would influence perceived argument strength.

Nondiversity via property reinforcement. For these items, we predicted that an argument with less diverse premises might be rated as stronger than an argument with more diverse premises if the premise categories of the latter reinforced a property not shared by the conclusion category. Item analysis validated most of the items; nondiverse premise pairs were chosen as more similar than diverse premise pairs for all but two of the items (see Table 3). Because diversity and property reinforcement made the same predictions for these items, they were excluded from analysis. As predicted, in the overall analysis, the nondiverse version yielded a significantly higher mean (4.70) than did the diverse version (4.37) [$F(1,405) = 8.84$, $p < .01$]. Table 2 reveals that 46% of the participants mentioned the target property for the diverse premises.

The item × item version interaction was significant [$F(8,405) = 13.24$, $p < .01$], which again required analyses at the level of the individual items. The analyses of the individual items showed that the mean rating for the nondiverse version was higher for six out of nine items, and the difference was significant for Item 1 of the Australian data and Item 2 of the Belgian data. Only Item 5 of the Australian data yielded a significantly higher mean rating for the diverse version. Overall, the predicted effect appears to have been readily demonstrated.

Conjunction fallacy via property reinforcement. For these items, we predicted that arguments with two-category conclusions might be rated stronger than arguments with one-category conclusions (a conjunction fallacy) if the added conclusion category reinforced a salient property shared by all three. The overall analysis yielded a significant difference between the three versions of the conjunction fallacy via property reinforcement items [$F(2,456) = 5.96$, $p < .01$]. As for the causal conjunction fallacy phenomenon, in this overall analysis the difference between the two single-conclusion versions was not relevant. Again, to be conservative, a contrast that compared the mean of the single-conclusion versions with the mean of the double-conclusion versions was formulated. As hypothesized, the mean rating for the single-conclusion versions (5.83) was found to be significantly lower than the mean for the double-conclusion versions (6.55) [$F(1,456) = 11.00$, $p < .01$]. Justifications revealed that (as with causal conjunction fallacy) whereas the target property was salient for the one-conclusion items (58% target relations), it was nevertheless rendered more salient by the additional conclusion category (75% target relations). The item × item version interaction was not significant.

Nonmonotonicity via property reinforcement. For these items, we predicted that adding premise categories might weaken the argument (nonmonotonicity) if the added items reinforced a property shared by the premise categories but not the conclusion. The overall analysis showed that the mean rating for the single-premise version (5.17) was significantly higher than the mean rating for the multiple-premise version (4.63) [$F(1,570) = 8.53$, $p < .01$]. Justifications revealed that, unlike with causal nonmonotonicity, the target property was rendered much more salient in the multipremise items (57% vs. 9% mention of target property). The item × item version interaction was not significant. In short, this phenomenon was robust.

In summary, for the property reinforcement items, the hypothesized conjunction fallacy and nondiversity phenomena were clearly supported by the data. Also, despite

the lack of evidence for nonmonotonicity in the causal scenario items, evidence supporting nonmonotonicity was found with property reinforcement items. In general, as predicted by the relevance framework, targeted properties were rendered salient by our manipulations.

DISCUSSION

Overall, the responses robustly demonstrated the importance of causal scenarios and property reinforcement in category-based induction and provide support for the relevance framework. In addition to the ratings of argument strength, examination of justifications revealed that, as predicted, target properties and causal relations were often explicitly mentioned by participants when they explained their rationales. The demonstration of causal asymmetries can be seen as a straightforward variation of effort, and it parallels the original observations by Tversky and Kahneman (1974) that correlations are construed as leading to better predictions when going from cause to effect than from effect to cause. Causal relations led to reliable preferences for the less diverse pair of premises. In addition, arguments with salient causal relations between dissimilar premise and conclusion categories were rated just as strong as arguments with more similar premise-conclusion pairs lacking any causal links. Overall, causal relations were of equal salience as, or greater salience than, similarity in the evaluation of arguments.

Causal relations also led to a conjunction fallacy whereby arguments with a wider conclusion were deemed stronger by virtue of emphasized causal relations between premise and conclusion. From the perspective of relevance theory, the conjunctive conclusion leads to greater confidence, because the additional category provides a link between the premise and the other conclusion category (as in *cows* acting as a mediator between *humans* and *grass*). The one prediction that failed concerned causal nonmonotonicity. Perhaps for these items, causal relations were not sufficiently compelling to overcome the strength of two premises as opposed to one. Indeed, the justification data suggest that causal relations for these items were not particularly salient.

Neither the SCM nor the FBIM addresses causal relations, and therefore they do not predict the preceding effects. It is also not clear how a Bayesian model could handle causal relations. One might posit that Bayesian calculations are made over the sets of features activated and treat the activation of features or properties as a separate issue requiring an independent theory. In the present study, however, it is precisely the activation of knowledge that is driving the phenomena of interest. The hypothesis-based model fares somewhat better, in that it views induction as a selection of (a limited number of) candidate bases for induction. In that respect, it is similar to the relevance framework. Before drawing any overall conclusions, we turn first to the phenomena associated with property reinforcement. Here the idea is that premises and conclusion categories are compared in an attempt to determine the relevant basis for induction.

Property reinforcement probes led to a conjunction fallacy and to a negative diversity effect. A striking example of the latter was that penguins and frogs produced greater confidence that some property would be true of giraffes than of penguins and polar bears. A model based on overall similarity would be committed to the opposite prediction, because polar bears are presumably more similar to giraffes than are frogs, which are not even mammals. On our account, this effect is mediated by the idea that the most relevant (or informative) relation between premises is that penguins and polar bears share living in a cold environment, something that is not true of giraffes. Moreover, unlike causal relations, property reinforcement produced nonmonotonicities such that fewer premises led to stronger arguments in cases where added premises reinforced a property not shared by the conclusion.

Again, the SCM does not address these effects. The original form of the FBIM also does not address our observed premise nonmonotonicities. Sloman (1993, pp. 267–269) offers a version in which features may be weighted by the number of categories that they are consistent with in such a way that features shared by all premise categories would have the greatest weight. His particular instantiation of this idea relies on weight decay, but the notion of selective attention and feature weighting is widespread in models of categorization (Tversky, 1977). Consequently, there is considerable reason to believe that selective attention will prove necessary and useful in category-based induction. In short, failure of diversity and premise nonmonotonicity should be within the scope of similarity-based models that allow for selective attention.

What makes the relevance framework distinctive is the conclusion conjunction fallacy via property reinforcement. Consider, for example, a version of feature-based induction in which multiple premises lead to some features' being weighted more heavily than others. The most straightforward way of applying such a model to conjunctive conclusion categories is to compute the similarity of the premise representation to each of the conclusion categories separately, with the assumption that argument strength is a function of whichever similarity is smaller (a min rule). Only when we add the idea that participants are trying to determine what the proper basis for induction is (and that the experimenter is cooperating in this enterprise) does it become plausible to assume that participants compare conclusion categories as an additional source of information about relevance. Computational models could be developed that incorporated conclusion comparison, but in so doing, they would represent implementations of, rather than alternatives to, the relevance framework.

The Bayesian and hypothesis-based induction models may fare better in addressing property reinforcement effects in the sense that they do not generate obviously incorrect predictions. In the case of the Bayesian model,

one would need to collect data on people's notions of featural distributions to generate predictions. In their absence, the Bayesian model is not committed to specific predictions. As in the case of causal relation phenomena, the factors driving the effects would be tied to feature activation, so that the Bayesian part of the model would be doing little explanatory work.

The relevance framework is most closely related to hypothesis-based models in that one could see the relevance framework as a basis for predicting which hypotheses people would tend to generate. In that sense, relevance is more powerful than the hypothesis-based model because it can generate predictions about induction phenomena without simply relating one kind of dependent variable to other dependent variables. In our view, the key variable that both kinds of models identify is the notion that induction is not computed over all potential features or associations but rather that judgments are based on a tiny (presumably relevant) subset of them. The justifications for judgments suggest that our manipulations of effect and effort were successful in modifying participants' ideas about relevance.

Other observations seem generally consistent with the relevance framework, and other predictions derive naturally from it. For example, Sloman (1998) found striking inclusion similarity and premise specificity effects (e.g., some participants rate *all animals* as providing a weaker basis for induction to *sparrows* than they rate *all birds*). However, these effects essentially disappeared when he added the inclusion relation as a premise (e.g., *All birds are animals*). Even varying the order of rating tasks to make the inclusion relation more accessible appeared to diminish the inclusion similarity effect (see, e.g., Sloman's Experiment 5 vs. Experiment 2).

The cross-cultural study of induction by Choi, Nisbett, and Smith (1997) suggests that chronic differences in category accessibility affect induction. They compared induction by Korean and American students for arguments involving both biological and social categories. Choi et al. review evidence suggesting that American students tend to categorize much more readily than Korean students, except perhaps in the social domain, where Koreans may have greater propensity to categorize. In agreement with this idea, they found that manipulations aimed at increasing category salience were effective for Korean participants for biological categories and for American students for social categories. From the perspective of relevance theory, manipulations aimed at increasing category salience affect effort and chronic differences in accessibility determine the efficacy of such manipulations.

Similarly, the degree to which taxonomic similarity guides category-based reasoning may be dependent on the reasoner's expertise in or familiarity with the domain in which the relevant premise and conclusion categories are located. Domain experts like those studied by López et al. (1997) and Proffitt et al. (2000) appear to have a range of strategies available for linking premises and conclusions, including both causal and functional relations as well as taxonomic similarity. The domain novices in these studies, on the other hand, appear to have had a more restricted repertoire, with taxonomic similarity serving as a default induction strategy.

A further consideration of the effort component of relevance theories suggests that we should not prematurely sell the repertoire of undergraduate reasoning strategies short. In addition to using salient causal relations in our study, we also tested participants individually and asked them to justify their answers. Implicit in these procedures was the request to apply more effort than one might observe if one tested participants in groups, with many, many items and without asking for justifications (with many probes, participants may look for a strategy that can be applied on every item). In one of our labs, we have obtained some preliminary evidence that testing participants individually rather than in groups leads to less use of abstract strategies such as diversity and more use of causal reasoning, even when the causal relations are not salient. Of course an alternative possibility is that asking for justifications biases participants toward strategies that are easy to justify.

Relevance theory has some of the positive and negative qualities that are associated with any framework theory. The notions of effort and effect can seem frustratingly vague, especially in an area where computational models are more the norm. Perhaps the best way to evaluate a framework theory is to do so on terms of its usefulness, and we hope that we have demonstrated its utility here.

There are other ways to test the relevance framework. For example, if participants were made to believe that the premises had been randomly selected, the effects associated with the present study should weaken or disappear. For example, consider our finding that (POISON IVY)/DANDELION was stronger than (POISON IVY + POISON OAK + POISON SUMAC)/DANDELION. If participants were led to believe that they were seeing just a subset of potential premises from a data set that had been alphabetized, they should be much less sure that being poisonous was the most relevant property. (The alphabetizing scenario is based on a suggestion provided by Dan Sperber, personal communication, June 7, 2001.)

Recognizing a role for relevance does not necessarily imply that similarity-based models are wrong or misguided. Rather, we argue that they are in principle incomplete, and that they may miss much of what is essential in human reasoning. The use of relevant background knowledge is central to induction and by no means exhausted by judgements of similarity, as our results clearly demonstrate. Nor does a relevance framework mean that more specific models cannot be successfully formulated, although such models would likely have to be more abstract than the SCM or the FBIM. (See, e.g., the premise probability principle of Lo, Sides, Rozelle, & Osherson, 2002, which may be able to represent the effects of causal relations among premises in a natural way.) What is clear from the present studies is that humans use all knowledge available to them when reasoning about the world. Causal

knowledge and specific relations among categories, as well as overall similarity, can be seen as critically relevant to the reasoning process and therefore cannot be ignored. We have attempted here to take a small step toward giving these other kinds of information their due.

REFERENCES

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., & Coley, J. D. (2002). A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition*, **84**, 1-53.

Blok, S. V., & Gentner, D. (2000, August). *Reasoning from shared structure*. Poster presented at the 22nd Annual Conference of the Cognitive Science Society, Philadelphia.

Choi, I., Nisbett, R. E., & Smith, E. E. (1997). Culture, category salience, and inductive reasoning. *Cognition*, **65**, 15-32.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, **65**, 263-297.

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 29-50.

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: Academic Press.

Griffiths, T. L., & Tenenbaum, J. B. (2000). Teacakes, trains, taxicabs, and toxins: A Bayesian account of predicting the future. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 202-207). Mahwah, NJ: Erlbaum.

Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 43-92). London: Psychology 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: Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, **7**, 569-592.

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

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. [N.] (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science*, **26,** 181-206.

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological 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., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.

Osherson, D. N., Smith, E. E., Wilkie, O., López, 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, & Cognition*, **26**, 811-828.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

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

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition*, **52**, 1-21.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, **35**, 1-33.

Smith, E. E., Shafir, E., & Osherson, D. [N.] (1993). Similarity, plausibility, and judgments of probability. *Cognition*, **49**, 67-96.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

Van der Henst, J.-B., Sperber, D., & Politzer, G. (2002). When is a conclusion worth deriving? A relevance-based analysis of indeterminate relational problems. *Thinking & Reasoning*, **8**, 1-20.

(Manuscript received July 17, 2001;
revision accepted for publication June 19, 2002.)