Psychonomic Bulletin & Review
2001, 8 (3), 600-608

Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm

MICHAEL R. WALDMANN
University of Göttingen, Göttingen, Germany

Causal directionality belongs to one of the most fundamental aspects of causality that cannot be reduced to mere covariation. This paper is part of a debate between proponents of associative theories, which claim that learners are insensitive to the causal status of cues and outcomes, and proponents of causal-model theory, which postulates an interaction of assumptions about causal directionality and learning. Some researchers endorsing the associationist view have argued that evidence for the interaction between cue competition and causal directionality may be restricted to two-phase blocking designs. Furthermore, from the viewpoint of causal-model theory, blocking designs carry the potential problem that the predicted asymmetries of cue competition are partly dependent on asymmetries of retrospective inferences. The present experiments use a one-phase overshadowing paradigm that does not allow for retrospective inferences and therefore represents a more unambiguous test of sensitivity to causal directionality. The results strengthen causal-model theory by clearly demonstrating the influence of causal directionality on learning. However, they also provide evidence for boundary conditions for this effect by highlighting the role of the semantics of the learning task.

The capacity to learn about causal relations is central to our survival. Causal knowledge enables us to predict future events or explain the occurrence of present events. Traditionally, the competency needed to acquire predictive knowledge has been studied within the framework of associationist theories of classical and instrumental conditioning. According to this framework, learning can be characterized as the acquisition of associative weights, which express statistical covariations between cues and outcomes (see Chapman & Robbins, 1990; Cheng, 1997). In causal learning situations, the cues and outcomes may represent causes and effects, and, indeed, a number of psychologists have argued that there is no need for a special theory of causal learning since these tasks are perfectly handled by a domain-general associative learning mechanism (e.g., Shanks & Dickinson, 1987).

However, the reduction of causal learning to mere covariation detection between cues and outcomes carries the cost of neglecting important characteristics of causality (see also Wu & Cheng, 1999). For example, covariation detection mechanisms do not differentiate between true causal (e.g., a virus causing a flu symptom) as opposed to spuriously correlated event relations (e.g., two correlated flu symptoms). Another typical feature of causality neglected by these mechanisms is its inherent directionality (see also Waldmann, 1996). Causes generate effects but not vice versa. Honoring this aspect of causality is important for planning actions since we need to know in which events we should intervene to accomplish the desired effects. Because of its inherent symmetry, covariation knowledge does not provide us with information about whether an event represents a cause or an effect. Knowledge about causal directionality also points us to appropriate measures of causal strength (see also Waldmann & Hagmayer, in press; Waldmann & Martignon, 1998). For example, in a common-effect model, the causal processes emanating from multiple causes physically converge on their joint effect. Because of the possible confounding by the alternative causes, it is necessary to hold them constant when assessing a specific cause–effect relation within the common-effect model. In contrast, in a common-cause model a single cause independently generates different effects. In this type of model, it is not necessary to hold constant a collateral effect when a specific cause–effect relation is assessed. Associative theories are not sensitive to these distinctions since the postulated learning mechanisms generate weights that express covariations between cues and outcomes independent of whether the learning cues represent causes or effects. The majority of associative theories postulate a covariation detection process in which alternative cues are held constant regardless of whether they represent causes or effects (e.g., Rescorla & Wagner, 1972).

Sensitivity to Causal Directionality in Two-Phase Blocking Paradigms

The fact that causal relations cannot be reduced to mere covariations raises the question of whether people are

The research was supported by DFG Grant Wa 621/5-4. I thank Y. Hagmayer, M. Hildebrandt, and A. Prasse for helpful discussions and for help with preparing and running the experiments. Correspondence concerning this article should be sent to M. Waldmann, Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany (e-mail: michael.waldmann@bio.uni-goettingen.de) (www.psych.uni-goettingen.de/abt/1/waldmann/).

Copyright 2001 Psychonomic Society, Inc.

sensitive to the characteristics of causality. In regard to causal directionality there has been an ongoing debate in the past few years between theorists with an associationist background, who claim that human causal learning lacks sensitivity to this aspect of causality (e.g., Matute, Arcediano, & Miller, 1996; Price & Yates, 1995; Shanks & Lopez, 1996), and those who hold to causal-model theory (Waldmann, 1996; Waldmann & Martignon, 1998), which posits that knowledge about causal directionality is used during learning.

This debate was started by Waldmann and Holyoak's (1992) experiments in which a two-phase blocking paradigm was used to demonstrate the influence of knowledge about causal directionality on learning. The general idea of these experiments was to present participants with identical cues and identical outcomes in all learning conditions while the causal interpretation of cues and outcomes was manipulated by means of differential initial instructions. One recent example of this paradigm (Waldmann, 2000, Experiment 2) presented participants with information about the presence or absence of substances in animals' blood. All participants learned in Phase 1 that Substance 1 (the predictive cue) was a perfect indicator of Midosis, a disease of the blood (the outcome). In Phase 2, an additional substance (Substance 2), the redundant cue, was constantly paired with the predictive cue from Phase 1. Now both cues together were perfectly correlated with Midosis. Two different instructions were compared. In the *predictive condition*, the substances were characterized as causes of the disease (common-effect model), whereas, in the *diagnostic condition*, participants were told that the substances were effects of the disease (common-cause model). Otherwise both conditions were identical.

Since all participants saw identical cues (substances) and had to predict identical outcomes (Midosis), modern associative theories, such as the Rescorla-Wagner theory (Rescorla & Wagner, 1972), would predict blocking of the redundant cue (i.e., Substance 2). This is a consequence of cue competition built into associative learning mechanisms. In particular, it is predicted that, despite the fact that both the predictive and the redundant cue individually are perfect predictors of the outcome, participants should rate the association of the predictive cue with the outcome as higher than the association of the redundant cue with the outcome, as compared with a suitable control group.

In contrast to associative theories, causal-model theory predicts blocking only in the predictive but not in the diagnostic condition (see Waldmann, 2000; Waldmann & Holyoak, 1992). In the predictive condition, participants were confronted with a common-effect situation in which, due to the constant pairing of Substance 2 with the deterministic cause from Phase 1 (Substance 1), the causal impact of the redundant Substance 2 could not be assessed. In contrast, in the diagnostic condition, the learning trials are interpreted as referring to a common-cause model in which the common cause, Midosis, deterministically causes two effects, (i.e., the two substances). Since no alternative causes of these substances were mentioned, both should be seen as equally valid indicators of Midosis (i.e., absence of blocking). The results of the experiments (Waldmann, 2000) clearly bore out these predictions. Blocking interacted with causal status.

Causal Models and Mutual Overshadowing

Waldmann and Holyoak's (1992) conclusion that blocking interacts with the causal status of cues and outcomes was criticized by a number of researchers (e.g., Matute et al., 1996; Price & Yates, 1995; Shanks & Lopez, 1996). Some questioned the data, but Waldmann (2000) has recently presented four experiments in which the basic finding was replicated. Another line of criticism has been that the effect might be restricted to peculiarities of two-phase blocking paradigms. Williams, Sagness, and McPhee (1994) have shown that blocking is not always obtained in two-phase designs, because this paradigm often seems to encourage configural processing. This may be one of the reasons why most of the critics chose to test the role of causal status with paradigms that only used a single learning phase. The results of these attempts turned out to be inconsistent. Whereas some studies found that causal status interacted with cue competition (Matute et al., 1996, Experiments 1 and 2; Van Hamme, Kao, & Wasserman, 1993), others failed to find any differences between predictive and diagnostic learning (Matute et al., 1996, Experiment 3; Price & Yates, 1995; Shanks & Lopez, 1996). Waldmann (2000) discusses a number of possible reasons for the null effects (see also Waldmann & Holyoak, 1997). For example, in some studies (e.g., Shanks & Lopez, 1996), statistical relations between cues and outcomes have been presented that are not equally compatible with common-cause and common-effect structures.

One of the goals of the present experiments was to test the hypothesis that cue competition interacts with causal order by using a one-phase learning task that is equally compatible with the two contrasted causal structures. Overshadowing tasks fit this requirement. Whereas blocking paradigms consist of two learning phases, overshadowing paradigms present a single learning phase in which two simultaneously present cues (e.g., a tone and a light) are jointly predictive of the outcome (e.g., a shock). A typical finding in this kind of task is that the associative weight attached to each cue is substantially lowered, as compared with control conditions, in which each cue is presented individually as a predictor of the outcome (i.e., overshadowing).

In the present experiments, a simple overshadowing task in which two cues were perfectly correlated with each other as well as with the outcome was used. The basic manipulation involved the causal interpretation of these two cues. In the predictive condition, they represented potential causes; in the diagnostic condition, they represented potential effects (see Figure 1). Apart from the differential initial instructions, cues, outcomes, and

feedback information were identical in the two conditions. Therefore, associative theories predict mutual overshadowing of the two cues in both conditions. For example, the Rescorla-Wagner rule would predict that, given equal learning rates, each cue would gain half of the associative weight that is provided by the learning asymptote instead of the full strength that each cue could obtain individually. Another associative theory, Pearce's (1987) configural-cue account, explains overshadowing as a consequence of a generalization decrement between training and testing. Whereas the training and testing phases for single predictive cues are identical, the testing phase in which each redundant cue is shown by itself differs from the training phase in which they are always presented as a compound. According to this theory, this greater generalization decrement is viewed as responsible for the overshadowing effect. However, both theories predict overshadowing independent of whether the cues represent causes (predictive learning) or effects (diagnostic learning).

In contrast, causal-model theory predicts different patterns of results for the predictive and the diagnostic conditions. According to this theory, learning is sensitive to causal direction and the structure of the underlying causal model. In the *predictive condition*, the relevant cues represent potential causes of a common effect. According to causal-model theory, the default assumption underlying common-effect models is that the different causes independently and additively influence the common effect. For probabilistic causes, the *noisy-or schema* has been proposed in the literature; it embodies the assumption that each cause independently contributes to the probability of the effect (Pearl, 1988; Waldmann & Martignon, 1998; see also Waldmann, Holyoak, & Fratianne, 1995, for empirical evidence). Furthermore, due to the possible confounding by the collateral cause, participants should attempt to hold the collateral cause constant when they assess the causal strength of the target cause. However, because information about the presence of each cause in the absence of the alternative cause is missing, it is expected that participants will lower their ratings for each potential cause. Previous research has shown that learners lean toward intermediate ratings when the causal status of a cue is uncertain (e.g., Waldmann, 2000; Waldmann & Holyoak, 1992). There is no way to decide whether a cue represents a deterministic or a probabilistic cause, a spurious correlate, or a part of a conjunctive cause. Since it is impossible to distinguish between the two cues, the most parsimonious guess is to equally divide causal strength between them, which would amount to mutual overshadowing. The strategy of dividing the influence among the different causal candidates is invited by the assumption that the causal impacts of multiple causes of a common effect combine additively (see also Waldmann et al., 1995).

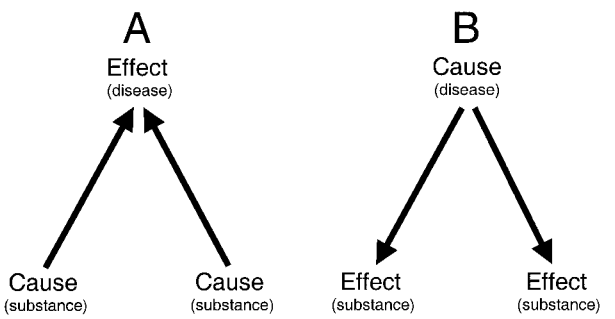

Figure 1. (A) Predictive common-effect model versus (B) diagnostic common-cause model. Cues (bottom) and outcome (top) are identical in both learning conditions.

By contrast, the diagnostic condition presents a situation with one common cause (outcome) that deterministically generates two effects (cues). In this situation, it is not appropriate to hold the collateral effect constant when the target effect is assessed. Thus, both cues should be viewed as deterministic effects of their common cause. Since no alternative causes for either effect were mentioned or observed, both cues should be seen as valid diagnostic indicators of the disease.

Another reason for using a one-phase overshadowing task comes from causal-model theory's analysis of the blocking paradigm. The most critical prediction for associative theories is the absence of cue competition in the diagnostic condition. Waldmann (2000, Experiment 2) has shown that this finding hinges on retrospective assumptions about Phase 1 that are a side effect of causal directionality. Let us assume, for example, that participants of a blocking study learn in Phase 1 that Midosis, a novel disease, always produces higher counts of T cells. Other cell lines are not measured so that it is uncertain whether Midosis has other effects. If, then, in Phase 2, it is discovered that, along with the T cells, the thrombocytes are also elevated, it is reasonable to infer that thrombocytes had been elevated in Phase 1 all along whenever Midosis was present. This is a consequence of the fact that, in Phase 2, Midosis is viewed as a cause of two effects—T cell and thrombocyte elevation. Unless there are reasons to assume that Midosis has changed its causal power with respect to one effect across the two phases, this retrospective inference is a natural consequence of the assumed common-cause model and the assumption of stability of causal power.

However, an associationist who is willing to model the differences between predictive and diagnostic learning may decide to forego the assumption that the observed effects are driven by sensitivity to causal directionality and, instead, focus on asymmetries of retrospective inferences. In the past few years, a number of extensions of associative theories have been put forward that address retrospective revaluations (e.g., Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994). So far, these theories have been unable to predict asymmetries of backward inferences for the predictive and the diagnostic conditions, in which identical cue-outcome structures were presented (see also Waldmann, 2000). However, it

is still desirable to test the impact of causal directionality in a paradigm, such as overshadowing, that is not potentially confounded with asymmetric retrospective inferences.

EXPERIMENT 1

The aim of this experiment was to test whether sensitivity to causal directionality can be demonstrated with an overshadowing paradigm. Participants learned about a fictitious disease, Midosis, that was characterized as either potentially being causally influenced by three substances in the blood (predictive common-effect condition) or as potentially causing the three substances (diagnostic common-cause condition). Apart from these different initial instructions, the learning trials were identical. In the overshadowing conditions, the participants learned about three substances, one of which was uncorrelated with the disease, and the other two were perfectly correlated with each other and the disease. Additionally, control conditions were run in which only one of the three substances was predictive of the disease.

Method

Participants and Design. Forty-eight students from the University of Göttingen, Germany, participated in this experiment. Half of them received diagnostic instructions (common-cause model), and half received predictive instructions (common-effect model). As a second factor, the number of predictive cues was manipulated (a single predictive cue vs. two predictive, mutually redundant cues).

Procedure and Materials. The participants received initial written instructions (in German) in which they learned that a new allergic disease, *Midosis*, had been discovered in animals. In the diagnostic condition, it was mentioned that this disease might produce new types of substances in the blood, whereas, in the predictive context, participants were told that new types of substances in the blood that come from food items may be the cause of this disease.

During the learning phase, individual cases were presented on a computer screen. Three substances, *alpha*, *beta*, or *gamma*, were listed one above the other and described as either being present (*yes*) or absent (*no*). In the two-cue overshadowing conditions with two redundant predictive cues, these two substances were perfectly correlated with each other and the outcome, whereas the third substance was uncorrelated. This cue was always absent. There were two types of cases: If all substances were labeled as being absent, the disease was also absent; when the two predictive substances both were simultaneously present and the uncorrelated substance absent, the presence of Midosis was indicated. In the one-cue control conditions, only one predictive cue was present when Midosis was present, and was absent when the disease was absent. The other two cues were always absent. The assignment of the three substances to the three cue types was counterbalanced in both the one-cue and two-cue conditions.

Each of the two types of cases was presented seven times in a random order before the ratings were requested. The task in the learning phase was to indicate whether the disease was present or absent by pressing one of two keys on the keyboard. After pressing the respective key, feedback was displayed below the information about the status of the cues that indicated whether the judgment was correct or incorrect (*response correct* vs. *response incorrect*). This information was replaced after 1,500 msec by information about the disease (*Midosis* vs. *No Disease*). This information was shown for 3 sec, after which the participants were alerted to press a key to receive information about the next case.

The learning phase was followed by the ratings. The rating instructions were identical for all conditions. In this task, the participants were asked to rate how well each individual substance predicted Midosis by using a number between 0 and 100. It was pointed out that 100 meant that the substance perfectly predicts the disease, whereas 0 meant that the substance does not predict Midosis at all. The participants were encouraged to use the numbers in between to express intermediate assessments. In all conditions, the same sequence of ratings, *alpha*, *beta*, and *gamma*, was used.

Results and Discussion

Figure 2 displays the mean ratings for the average of the two mutually redundant predictive cues in the two-cue condition or the single predictive cue in the one-cue condition. The uncorrelated cues (i.e., one cue in the two-cue condition and two in the one-cue condition) received a rating of 0 by all but 1 participant from the diagnostic two-cue condition, who chose 10. A 2 (predictive vs. diagnostic learning) × 2 (one vs. two cues) analysis of variance (ANOVA), with the ratings of the predictive cues as the dependent variable, yielded a highly significant difference between the predictive and the diagnostic conditions [$F(1,44) = 14.8$, $MS_e = 172.5$, $p < .01$], as well as between the two- and the one-cue conditions [$F(1,44) = 75.4$, $MS_e = 172.5$, $p < .01$]. These two main effects were, however, moderated by a highly significant interaction [$F(1,44) = 14.8$, $MS_e = 172.5$, $p < .01$]. Overshadowing was substantially greater for predictive than for diagnostic learning.

Ignoring the difference between the predictive and the diagnostic condition, an associationist might argue that the experiment shows that overshadowing was observed in both the predictive and the diagnostic two-cue condition. In both conditions, the ratings for the two mutually redundant cues proved to be clearly lower than the ratings for the single predictive cue in the control conditions. In fact, all participants in the one-cue condition gave the predictive cue a rating of 100. However, associative theories would fail to account for the second important result of this experiment, the highly significant difference between the ratings of the redundant cues (two cues) in the predictive and the diagnostic conditions [$F(1,22) = 14.8$, $MS_e = 345.1$, $p < .01$] (see Figure 2). Associative theories would predict identical amounts of overshad-

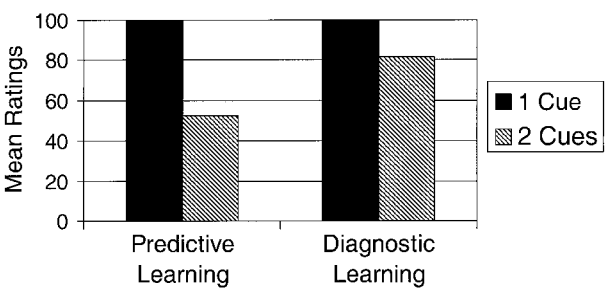

Figure 2. Mean predictiveness ratings for predictive and diagnostic conditions for the predictive cues in the one-cue and the two-cue conditions (Experiment 1).

owing in both conditions because identical trials had been presented.

How does causal-model theory account for the obtained results? According to the theory, the lowered ratings of the two redundant cues in the predictive-learning condition are a consequence of the fact that it is impossible to compute informative contrasts in which the alternative cause is kept absent. Furthermore, lowered ratings are invited by the assumption inherent in the semantics of common-effect models that the two causes should independently and additively influence the effect. In contrast, the one-cue condition presents a single deterministic cause that, accordingly, received maximal ratings in the predictive condition. The difference between the predictive and diagnostic conditions is also predicted by causal-model theory. In the diagnostic condition, participants should learn that the disease is a common cause of two deterministic effects. Thus, both redundant cues should be viewed as strongly related to the disease.

The difference between the one-cue and the two-cue conditions with diagnostic cover stories presents a problem for causal-model theory, however. In both the one- and two-cue conditions, the participants were presented with effect cues that are deterministically related to their cause. The common-cause model, which encodes the assumption that the effects are independent of each other conditional upon the common cause, allows it to infer the state of the common cause from either of the observed effects. Given that no alternative causes were mentioned in the instructions, maximal and equal ratings should be obtained in both the one-cue and the two-cue conditions. In contrast to this prediction, only 5 out of 12 participants gave a rating of 100 for the redundant cues in the diagnostic two-cue condition. The rest chose ratings between 50 and 90. Although these ratings are clearly higher than in the two-cue predictive-learning condition, in which 11 out of 12 participants gave ratings of 50 or below for these cues, they are also clearly below the ratings in the one-cue condition [$F(1,22) = 8.9, MS_e = 225.8, p < .01$].

A closer look at the one-cue and the two-cue tasks reveals that this contrast provides a particularly strict test of causal-model theory. In the one-cue condition, a single cue is deterministically related to the outcome. Regardless of whether this cue is viewed as a cause or as an effect, maximal ratings are predicted for this cue. Maximal ratings are also predicted if learners ignore the cover stories and treat the learning events as associated cues and outcomes. Thus, this condition is extremely robust. By contrast, the two-cue condition is sensitive to causal manipulations. Only the participants who adopt the instructed common-cause model with conditionally independent effects and who do not assume alternative hidden factors are expected to give maximal ratings. The participants who ignore or misrepresent the causal status of the cues, impose a predictive-learning frame on the task, assume violations of conditional independence, or bring to bear additional pre-experimental assumptions to the task are expected to lower their ratings.

This analysis predicts that noise in the data should not equally affect the one-cue and the two-cue conditions but should tend to lead to a lowering of the ratings in only the two-cue condition. This prediction is supported by the fact that, in the previous experiments, in which diagnostic learning was studied within a blocking paradigm, the redundant cue (which corresponds to the ratings of the predictive cues in the two-cue condition) was rated either slightly lower or equal but never higher than the single predictive cue (corresponding to the one-cue condition; Waldmann, 2000; Waldmann & Holyoak, 1992).

Thus, according to causal-model theory, complete absence of overshadowing in the diagnostic condition is to be expected only if every single participant adopts the causal model in the intended fashion. The present results showed that participants' ratings were strongly influenced by causal models, but they also indicate that a subset of participants in the diagnostic condition were influenced by additional factors.

EXPERIMENT 2

Experiment 1 provided clear evidence for sensitivity to causal directionality in a one-phase overshadowing paradigm. However, the lowered ratings in the two-cue conditions also signified that not all the participants in the diagnostic condition gave ratings optimally conforming to the predictions of causal-model theory. Although, on average, the ratings were clearly higher in this condition than in the predictive conditions, they still were lower than the ratings obtained in the one-cue condition. These results raise the question as to why, in the previous blocking experiments, no significant blocking effect was observed in the diagnostic conditions. These experiments also compared ratings for a single predictive cue with ratings for a redundant cue.

Revisiting the results of the blocking experiments reveals that some experiments indeed showed a small (although nonsignificant) effect in the direction of a blocking effect (see Waldmann, 2000; Waldmann & Holyoak, 1992). Moreover, in most experiments, the ratings deviated from 100—that is, the ratings predicted by causal-model theory for deterministic relations. Part of the reason for this pattern of results may again be that some of the participants did not adopt the instructed common-cause model in a fashion predicted by causal-model theory, or that some participants interpreted the rating instructions differently. The optimal result for causal-model theory requires that *all* participants behave according to the predictions of causal-model theory, which certainly is unlikely to be observed. Interestingly, however, nearly optimal behavior, in which almost all participants gave ratings of 100 for the predictive and the redundant cues, was indeed seen in one experiment of Waldmann (2000, Experiment 1). In that experiment, participants learned about a box with buttons on one side (the cause side) and lights on the other side (the effect side). The task was to learn how the buttons and lights were causally linked.

What is the difference between this cover story and the cover stories used in the present Experiment 1?

One important difference between the two domains is that simple devices represent closed worlds with salient causal structures. If, for example, the task is to learn that two observed lights (potential effects) are switched on by a specific button on the other side of the box, and all learning trials are consistent with this causal model, it is plausible to assume that there are no other invisible causes hidden in the box that were not mentioned and that would generate novel effect patterns in the future. In this situation, the presence of the common cause can be inferred by observing one of the two redundant effects of this common cause, which in turn implies the presence of the other redundant effect.

However, almost all the participants in Experiment 1 who lowered their ratings in the diagnostic two-cue condition mentioned in an informal interview at the end of the experiment that the request to assess each symptom individually made them think of the possibility that at least in some cases this symptom could occur by itself. Since the learning trials showed that the disease Midosis consistently generated both symptoms simultaneously, the potential presence of one of the symptoms by itself may rather be viewed as evidence for an alternative cause.

It is plausible to assume that this interpretation of the test questions is particularly likely in such domains as diseases, which are characterized by the fact that symptoms may be caused by several factors, including factors that are currently unknown. Diseases are typically diagnosed on the basis of whole patterns of symptoms because each symptom individually may be a sign of a different disease. For example, hepatitis is likely when a specific pattern of symptoms is observed, which may include fatigue, nausea, vomiting, pain in the liver area, dark urine, or fever. However, in case only one of these symptoms is observed by itself, other diagnoses seem more plausible. In contrast, in a closed-world domain, such as with the simple devices mentioned, it is unlikely that people would imagine that one effect light goes on by itself as an indicator of an alternative, unknown cause when so far the switch had always turned on the two lights together.

Finally, devices represent salient causal models, which allow for a graphical representation of the underlying structure, such as buttons and lamps linked by simple causal mechanisms (e.g., electrical wires). Such graphical mental models accessibly represent conditional independence between the two effects of the common cause, which is one of the requirements for maximal ratings in deterministic common-cause situations. By contrast, most people only have skeleton knowledge of the causal models underlying diseases.

The aim of Experiment 2 was to investigate whether the lowering of the ratings in the two-cue condition relative to the one-cue condition would be reduced if the task involved learning about a device similar to the one used in Waldmann (2000, Experiment 1). As in Experiment 1, a single predictive cue was compared with two, mutually redundant cues. However, in the present experiment, overshadowing was tested on the basis of a within-subjects comparison.

Method

Participants and Design. Twenty-four students from the University of Göttingen, Germany, participated in this experiment. Half of them received diagnostic instructions, and half predictive instructions. As a within-subjects factor, ratings for a single predictive cue were compared with the average ratings of the two mutually redundant predictive cues.

Procedure and Materials. Participants in the predictive condition were told that they were going to learn about a box. It was pointed out that on the front side of the box three colored lights (red, blue, and green) could be seen that could be switched on by pressing the respective button below the light. Furthermore, it was mentioned that there were two lights on the back side, light 1 and light 2 (the effects). The task was to predict whether light 1, light 2, or none of the lights would be on, based on information about the states of the visible lights on the cause side of the device. It was emphasized that either light 1 or light 2 but not both could be on at the same time. After the instructions, the participants were requested to explain the causal situation by using a picture model that was handed to them. The picture showed the three colored lamps on one half of a sheet of paper and the two numbered lights on the other half. The paper was folded in the middle to represent that only the three colored lights were visible during learning.

The diagnostic condition was modeled closely after the predictive one. Again the three colored lamps from the front side were mentioned, which, in this condition, represented potential effects and therefore had no attached buttons. The lights on the back side, lights 1 and 2, represented indicators of potential causes of the three lights visible on the front side. The instructions stated that below each indicator light on the back side a button was placed, which could be pressed. Pressing the button would cause the attached light, light 1 or 2, to go on. Again the task was to say whether light 1 or light 2 was on, based on information about the states of the visible effect lights on the front side. Thus, this task involved diagnosing whether one of the indicator lights on the back side was on as a consequence of the attached button's having been pressed. As in the predictive condition, it was pointed out that on every trial only one of the two lights on the back side could be on, and again the participants were shown a paper model to recapitulate the task.

After the instructions, all participants received identical learning input. They were handed 12 index cards on which they could see patterns of the three visible lights. No buttons were visible in either of the two causal conditions. Four cases were shown in which all lamps were off (i.e., gray), and the correct answer was to say that both lamps on the back side were also off. Four further cases showed one lamp being on by itself, and the correct answer was, for example, that light 1 was also on. This lamp represented the single predictive cue (one-cue condition). The assignment of this cue to the three colored lamps on the front side and the two lamps on the back side was counterbalanced. Finally, the set included 4 more cards in which the remaining two colored lights on the front side were on simultaneously, and the correct answer was that the other light on the back side, for example, light 2, also was on. These two lights represent the mutually redundant cues (two-cues condition).

During the learning phase, the participants were presented with the 12 index cards one after another in random order and had to judge whether light 1, light 2, or none of the lights was on, and after making their judgments received feedback from the experimenter. In the test phase, 3 different index cards were laid out in front of the participants one after another, which showed only one of the three lights as being on. The other two lights were covered by a big grated square indicating that the current state of the lights below the squares could not be presently observed. The rating instructions emphasized that

the index cards still represented the same box, which functioned exactly the way participants had learned. They were then requested to rate how well each of the three colored lamps predicted the state of the two lamps on the back side by using a number between 0 (*not predictive at all*) and 100 (*perfectly predictive*) (i.e., six ratings).

Results and Discussion

Figure 3 shows the mean ratings for the single predictive cue and the average of the two redundant cues (which received equal ratings by all participants). The results are perfectly in line with the predictions of causal-model theory. All participants gave a rating of 100 for the single cue (one-cue condition). Also, *all* participants in the diagnostic condition rated the redundant cues (two-cues condition) at 100.[1] In contrast, the ratings for the redundant cues in the predictive condition were substantially lowered despite the fact that these participants received the same learning input as the ones in the diagnostic condition. Seven out of the 12 participants rated the redundant cues at 50. Finally, all participants in both conditions chose 0 for the ratings that referred to the noncausal relations (e.g., the relation between a colored light and the light on the back side that was not switched on [predictive condition] or that was not the cause of this light [diagnostic condition]). A 2 (predictive vs. diagnostic learning) $\times$ 2 (one vs. two cues) ANOVA, which analyzed the ratings for the causally linked, predictive relations, yielded a highly significant interaction between the two factors [$F(1,22) = 22.2, MS_e = 135.6, p < .01$].

These results replicate the findings of Waldmann (2000, Experiment 1) with a one-stage overshadowing paradigm. They indicate that judgments perfectly corresponding to the predictions of causal-model theory can indeed be observed with cover stories that present salient causal models and that suggest closed-world assumptions, thus eliminating the potential impact of assumed hidden causal factors.

Although the obtained numerical ratings fully correspond to the predictions of causal-model theory for deterministic relations, a possible concern raised by one commentator is that the uniformly high ratings in the diagnostic condition may have created a ceiling effect. One problem of this critique is that no theory has been proposed in the literature that would account for the highly significant difference between the ratings of the redundant cues in the predictive, as compared with the diagnostic conditions, in which identical trials had been presented. Furthermore, previous research with more complex tasks has shown that the interaction predicted by causal-model theory is also seen when, possibly due to noise, the mean ratings are not maximal (e.g., Waldmann, 2000, Experiment 3).

GENERAL DISCUSSION

Previous demonstrations of the interaction between causal status and cue competition were mainly based on two-phase blocking tasks. Some critics have argued that the effect may be restricted to this type of task, and indeed, sometimes did not find any evidence for sensitivity to causal directionality with other tasks (e.g., Matute et al., 1996; Price & Yates, 1995; Shanks & Lopez, 1996). The present experiments provide clear evidence that the influence of the causal status of cues and outcomes can also be observed in a one-phase overshadowing task. The two experiments demonstrated substantial differences in identical learning tasks, depending on whether the cues were described as causes of a common effect or effects of a common cause. In general, substantially more overshadowing was observed in a predictive task, in which the cues represented causes, than in a diagnostic task, in which the cues represented effects.

However, the experiments also demonstrate the role of additional factors influencing participants' judgments. The predictions of causal-model theory are extremely strict. They are based on the assumption that *all* participants adopt the intended causal model, do not bring to bear additional background assumptions, and reason normatively. For example, in the diagnostic conditions of our experiments, behavior fully corresponding to the predictions of causal-model theory can only be expected if all participants adopt the instructed common-cause model with conditionally independent effects and do not consider the possibility of alternative, hidden causal factors.

Interestingly, it is indeed possible to observe optimal or nearly optimal behavior with task domains that are easily mapped onto common-cause models and that rule out the influence of hidden factors. The present Experiment 2 along with Waldmann's (2000) Experiment 1 shows that artificial devices represent such domains. Artificial devices lend themselves to being represented in a fashion similar to the graphical representations favored by Bayesian causal models (see Glymour & Cooper, 1999; Pearl, 2000). These graphical models provide a particularly intuitive way of representing conditional independence assumptions inherent in complex causal models. Furthermore, artificial devices typically represent closed worlds, which discourages the assumption of additional causal factors not mentioned in the instructions. Accordingly, in this experiment, no overshadowing was observed in the diagnostic condition.

By contrast, Experiment 1 showed slight but clear deviations from the predictions of causal-model theory. Al-

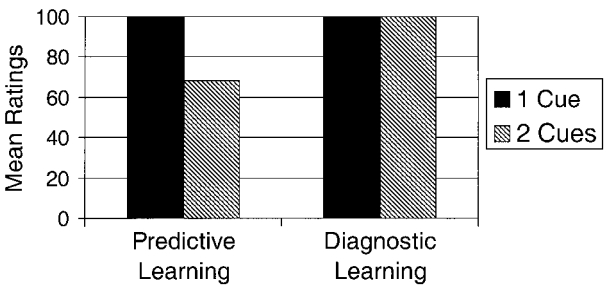

Figure 3. Mean predictiveness ratings for predictive and diagnostic conditions for the single predictive cue (one-cue conditions) and the two redundant cues (two-cues conditions; Experiment 2).

though the difference between predictive and diagnostic learning was once again observed, the results also revealed a small, but significant overshadowing effect in the diagnostic condition. This effect can be interpreted as evidence for additional factors affecting participants' judgments. The major difference between the two experiments was the semantics of the task domains. Whereas Experiment 2 mapped the overshadowing paradigm onto artificial devices, Experiment 1 used a fictitious disease. Although the disease and the symptoms used in this experiment were novel, it seems plausible that learners brought to bear general world knowledge about diseases on this task. Diseases are typically characterized by complex, partly unknown underlying causal networks, and by the presence of unknown hidden causes. This opaqueness might have led some participants in the diagnostic condition of Experiment 1 to deviate from the evidence provided by the learning input and to consider the possibility that the request to assess single symptoms might refer to situations in which the symptom occurs by itself. Although this case had never been observed in the learning phase, it is certainly a common possibility with diseases in the real world.

It is interesting to see that the outcomes of diagnostic analyses, at least in the type of tasks discussed in this article, superficially become more similar to predictive strategies when the task domains are complex. Similar to the reduced individual predictiveness of single causes that always have been part of a compound of causes, single effects also tend to be less diagnostic of a specific cause than the compounds they usually are part of when the task domain includes hidden causal factors of the effects. This can be exemplified by an example. It is a known fact that the HIV virus has a causal influence on the result of the ELISA test and of the Western-Blot test, two tests commonly used for the diagnosis of an infection. Although both tests are highly reliable, false alarms are possible. Assuming that only one of the two tests was conducted, observing a positive test result raises the likelihood that the patient is infected. From this result, the prediction can be derived that this patient will also test positive in the other test. However, the possible observation that the second test turned out negative reduces the likelihood of an infection relative to what was expected after the single test, whereas the observation that both tests are positive would increase this likelihood. This pattern is a normative implication of the assumed common-cause model if diagnostic judgments are consistent with normative Bayesian diagnostic inference strategies (see Waldmann & Martignon, 1998). Thus, in complex domains with hidden factors, diagnostic overshadowing actually is a normative implication of causal-model theory, provided participants have a tendency to interpret the instruction to rate individual effect cues as requests to consider the possibility of these effect cues occurring by themselves in the absence of other effects of the common cause.

Predicting overshadowing in both predictive and diagnostic learning raises the question of whether it is not more parsimonious to postulate uniform, possibly associative learning mechanisms, at least for complex domains. The present experiments, however, do not support this conclusion. Both experiments show clear differences between predictive and diagnostic learning with identical learning trials. Although a small diagnostic overshadowing effect was observed in Experiment 1, the high ratings in the diagnostic, relative to the predictive condition, show, nevertheless, that, on average, participants proved sensitive to the statistical structure of the learning events, which suggests deterministic relations between the cause and its two effects in the diagnostic condition. The theoretical possibility that the data is generated by a mixture of participants following causal-model theory and participants following associative theories is also not supported by the results, because this would predict a bi-modal distribution in the diagnostic condition rather than the uniform shift towards maximal ratings, which was observed in this condition relative to the predictive condition. Thus, the results in the diagnostic condition are better accounted for by a theory that postulates the use of a deterministic common-cause model along with additional assumptions that are based on prior knowledge about the learning domain.

In summary, the present research adds to the growing body of recent research showing that humans do not simply associate cues with outcomes in causal learning but are capable of acquiring knowledge about causal models in a normative fashion. The results also show that in complex domains learners do not always fully rely on the observed statistical structure of the learning input, but in addition bring to bear prior assumptions about the learning domain on the task.

REFERENCES

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, **18**, 537-545.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, **104**, 367-405.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology*, **49(B)**, 60-80.

Glymour, C. N., & Cooper, G. F. (Eds.) (1999). *Computation, causation, and discovery*. Cambridge, MA: MIT Press.

Matute, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 182-196.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, **94**, 61-73.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.

Price, P. C., & Yates, J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1639-1655.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of

causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 21, pp. 229-261). New York: Academic Press.

Shanks, D. R., & Lopez, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory & Cognition*, **24**, 511-522.

Van Hamme, L. J., Kao, S. F., & Wasserman, E. A. (1993). Judging interevent relations: From cause to effect and from effect to cause. *Memory & Cognition*, **21**, 802-808.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning & Motivation*, **25**, 127-151.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol 34. Causal learning* (pp. 47-88). San Diego: Academic Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 53-76.

Waldmann, M. R., & Hagmayer, Y. (in press). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, **121**, 222-236.

Waldmann, M. R., & Holyoak, K. J. (1997). Determining whether causal order affects cue selection in human contingency learning: Comments on Shanks and Lopez (1996). *Memory & Cognition*, **25**, 125-134.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, **124**, 181-206.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1102-1107). Mahwah, NJ: Erlbaum.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 694-709.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, **10**, 92-97.

NOTE

1. One participant in the diagnostic condition was replaced because she was apparently confused about the task. In the final ratings, she chose equal ratings of 0 (instead of 100) for the deterministic causal relations, in which the single and the redundant cues were involved, while she rated the relation between the single predictive cue and the light on the back side that was *not* causally linked at 100 and between the redundant cues and the other light at 50. Under the post hoc assumption that this participant confused the two alternative causes, her ratings may be viewed as the single instance of behavior observed in the diagnostic condition of this experiment that was inconsistent with causal-model theory.

(Manuscript received September 16, 1999;
revision accepted for publication December 20, 2000.)