Cognitive Science 31 (2007) 233–256
Copyright © 2007 Cognitive Science Society, Inc. All rights reserved.

Combining Versus Analyzing Multiple Causes: How Domain Assumptions and Task Context Affect Integration Rules

Michael R. Waldmann

Department of Psychology, University of Göttingen, Germany

Received 25 April 2006; received in revised form 19 September 2006; accepted 20 September 2006

Abstract

In everyday life, people typically observe fragments of causal networks. From this knowledge, people infer how novel combinations of causes they may never have observed together might behave. I report on 4 experiments that address the question of how people intuitively integrate multiple causes to predict a continuously varying effect. Most theories of causal induction in psychology and statistics assume a bias toward linearity and additivity. In contrast, these experiments show that people are sensitive to cues biasing various integration rules. Causes that refer to intensive quantities (e.g., taste) or to preferences (e.g., liking) bias people toward averaging the causal influences, whereas extensive quantities (e.g., strength of a drug) lead to a tendency to add. However, the knowledge underlying these processes is fallible and unstable. Therefore, people are easily influenced by additional task-related context factors. These additional factors include the way data are presented, the difficulty of the inference task, and transfer from previous tasks. The results of the experiments provide evidence for causal model and related theories, which postulate that domain-general representations of causal knowledge are influenced by abstract domain knowledge, data-driven task factors, and processing difficulty.

Keywords: Causal reasoning; Domain specific and domain general; Learning; Bayes nets; Top down learning

1. Introduction

People rarely acquire knowledge about complex causal models at once. Often, people only learn about fragments of causal knowledge, which they later combine to more complex networks. For example, people may learn that they tend to get a stomach ache when they drink milk or when they take an aspirin. One may never have taken aspirin with milk but still

Correspondence should be addressed to Michael R. Waldmann, Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany. E-mail: michael.waldmann@bio.uni-goettingen.de

will have a hunch what the effect on the stomach might be. This example shows that people are using intuitive integration rules when combining multiple causes. Sometimes, people also want to analyze the influence of individual causes. A person may have eaten peanuts along with broccoli in a Thai dish that made her sick. Later, that person finds out that peanuts alone cause the stomach ache. Most people will now have an intuition about the probable causal impact of broccoli on their health status. This is also an example of the use of an integration rule, which in this case is used to analyze the probable causal role of broccoli. Because any integration rule is possible, there is no normatively correct answer to these questions. However, as the examples demonstrate, people nevertheless have strong intuitions. The main goal of my research was to explore the sources of these intuitions.

1.1. Domain theories and causal learning

In philosophy and psychology, there has been a long-standing debate about the relation between domain knowledge and causal learning. In philosophy, statistical approaches to causal induction, which can be traced back to Hume (Salmon, 1980; Suppes, 1970), have competed with the view that causal hypotheses refer to domain-specific physical processes that link causes and effects (Dowe, 2000; Salmon, 1984). In psychology, a related debate has centered on the question of whether causal learning is based on domain-specific or domain-general learning strategies (Ahn, Kalish, Medin, & Gelman, 1995; Carey, 1985; Gopnik & Meltzoff, 1997).

One way to look at this debate is by representing it as a dispute about the proper level of abstraction. Cartwright (1999, 2004), in her critique of causal Bayes nets, has argued that representing events simply in terms of causes and effects is too abstract to account for the many possible meanings of causation. For example, when people talk about devices, they hardly ever use the term *cause and effect* but rather use far richer ("thicker") terminology such as feeding, opening, sucking, allowing, or speeding. Similarly, scientists in physics, chemistry, or biology rarely think about their task as designing abstract cause-effect graphs but attempt to develop specific physical theories that are entrenched with content and based on existing domain theories. Thus, according to this analysis, domain theories and causal learning are not separated but are two sides of the same coin.

Causal Bayes nets are a recent development of the other extreme, the view that causal learning should be reconstructed as using abstract representations and induction strategies. Bayes nets were initially developed as domain-general approaches to causal induction that represent domain knowledge abstractly in terms of networks of causes and effects (see Gopnik et al., 2004). They provide statistical algorithms that recover causal structure from data irrespective of the domain that is being modeled.

Both sides have strengths and weaknesses. Against the domain-specific view, Cheng (1993) countered that these theories do not explain how this knowledge is acquired in the first place. Moreover, recent studies have shown that, although people often believe that they have detailed knowledge about mechanisms, in reality, this knowledge typically has holes and is skeletal at best (Rozenblit & Keil, 2002). People can manipulate their television set or know about the effects of smoking without having acquired deep knowledge about the underlying mechanisms. Often, people's knowledge is based on statistical input and additional fairly

abstract cues (e.g., temporal order) that provide them with some rough, pragmatically useful knowledge about cause-effect relations without giving them details about mechanisms (e.g., smoking disease; see Lagnado, Waldmann, Hagmayer, & Sloman, in press; Waldmann, 1996; Waldmann, Hagmayer, & Blaisdell, 2006).

However, it is an overstatement to say that people do not have any knowledge at all. Representing events as causes and effects already provides people with important information that is not captured by purely statistical or associative theories that only represent temporally ordered events (Waldmann, 1996; Waldmann et al., 2006). One is rarely in a situation in which there are not at least some cues that help to go beyond purely event-based representations. Moreover, there are also computational reasons why a purely abstract induction mechanism seems implausible. Such mechanisms require lots of data and very precise estimates for correct inductions. Tenenbaum and Griffiths (2003) pointed out that human learning is extremely efficient; often people acquire knowledge with very few observations, which casts doubt on the adequacy of learning models that require a large amount of data to constrain learning.

1.1.1. Causal-model theory and theory-based Bayesian models

Causal-model theory was one of the first theories that tried to overcome the deficits of purely domain-specific and domain-general theories (see Waldmann, 1996; Waldmann & Holyoak, 1992; see also Lagnado et al., in press). The main idea is that prior knowledge constrains the structure of the causal models that are used as initial hypotheses in learning. This knowledge may be specific but is more often fairly abstract. Waldmann and Martignon (1998) suggested Bayes nets whose structure is based on prior knowledge as a possible formalization of causal-model theory (see also Waldmann & Hagmayer, 2001).

In previous research on causal-model theory, we have focused on the distinction between causes and effects, which entails structural knowledge about the direction of the causal arrow. In virtually every psychological experiment on causal learning, participants know which events represent potential causes and which represent effects. This is an example of very abstract causal knowledge that nevertheless has an effect on learning. In my research, I and my collaborators (Waldmann, 2000, 2001; Waldmann & Holyoak, 1992) have shown that assumptions about causal directionality affect cue competition, the learning of linearly separable and nonlinearly separable categories (Waldmann, Holyoak, & Fratianne, 1995), and the estimation of causal strength (Waldmann & Hagmayer, 2001). Moreover, Hagmayer and Waldmann (2002) showed that assumptions about temporal delays between causes and effects also influence learning. Waldmann (1996) claimed that prior domain knowledge also may affect the choice of the rules that underlie the integration of multiple causes.

Tenenbaum and Griffiths (2003) and Tenenbaum, Griffiths, and Niyogi (in press) have also argued for theory-based mechanisms in causal learning. Tenenbaum and Griffiths and Tenenbaum et al. have developed formalized Bayesian models that capture the intuition that causal learning is guided by domain knowledge. According to this approach, there is a hierarchy of domain theories that constrain each other. This domain knowledge may be fairly abstract, as for example in the case of causal grammars that specify the possible causal roles of event classes (e.g., symptoms as effects of diseases). In domains with which people have more experience, this domain knowledge may also be very specific, capturing the fundamental theoretical concepts and laws of the domain. Causal Bayes nets are, according to this view,

specific theories that are generated on the spot to guide induction and reasoning and are heavily constrained by top-down knowledge.

Tenenbaum and his colleagues (Tenenbaum et al., in press) also have addressed the role of functional forms (i.e., *integration rules* in this terminology). They have argued that domain knowledge not only constrains causal structure but also functional forms that specify the relation between multiple causes and effects. The important role of functional form in explaining efficient induction was empirically demonstrated by Sobel, Tenenbaum, and Gopnik (2004) in a simple learning task in which children were requested to learn to decide whether a block is a so-called blicket or not by observing it being placed on a blicket detector that can make a sound signal. Sobel et al. argued that children's performance can be best modeled if it is assumed that they enter the task with the prior assumption that individual blickets can activate the detector and that multiple blickets do not interact (i.e., "noisy-or" rule). Consequently, Tenenbaum and Griffiths (2003) developed a Bayesian model that explains learning by combining data-driven processes with top-down assumptions about causal structure and functional forms. Similar to the children, this model can master the task in a few learning trials.

1.2. Integration rules

There has been little empirical work on integration rules in causal learning, but most formal theories assume a tendency toward an additive integration rule.[1] For example, in most associative theories of causal learning, networks are used that add up the associative weights when making predictions for compounds of stimuli (e.g., Rescorla & Wagner, 1972). This assumption can be tested, for example, in tasks in which animals learn that a light as well as a tone are individually followed by shock with specific probabilities. According to most associative theories, animals should be more afraid when they experience the compound of both cues than when they experience each cue individually. This is an example of how cues are combined. The analysis of cues has also often been investigated. In the blocking paradigm, for example, an animal might experience that two cues together, tone and light, cause a shock but that the light by itself causes the same probability of shock. Most theories predict that the second cue is discounted in such a situation (i.e., blocking), which would also be evidence for an additive integration rule (see Waldmann, 2000). Additivity is also the typical default rule in probabilistic theories. A typical assumption is that multiple causes independently generate the common effect, which can be modeled by a noisy-or gate (see Cheng, 1997; Glymour, 2001; Tenenbaum & Griffiths, 2003).

Additive integration is not the only possibility of how multiple causes can be combined. A simple alternative linear rule is averaging (see Anderson, 1981; Busemeyer, 1991; Schlottmann & Anderson, 1993):

$$E = \frac{(w_1 \,.\, U_1 + w_2 \,.\, U_2)}{(w_1 + w_2)} + X. \tag{1}$$

Equation 1 gives the formula for weighted averaging of two causes. In a learning situation in which multiple (continuously varying) causes are always shown together, this rule cannot be distinguished from other linear additive rules. A simple linear regression model without

interaction terms would be capable of learning the weights for the two causes. However, weighted averaging can be distinguished from additive integration by comparing situations in which the number of cues is varied. A built-in constraint of Equation 1 is that the weights always sum up to one in every situation. In simple combination tasks in which learners observe each cause individually and then predict the effect of the never-observed compound of both causes, a weighted averaging model entails that the effect of the compound should lie in between the values of the effect for the two individual causes. How close the effect will be to the effect of either of the individual causes will depend on the weights. In contrast, additive integration rules imply that the effect will be stronger when both causes are present than when either one is present by itself (assuming that both causes have a positive weight).

Both tasks, combining causes and analyzing causal contributions, have been empirically studied. However, very few studies have specifically investigated causal learning. Animal learning studies have typically confirmed the use of additive rules, which are built into associative theories (Couvillon & Bitterman, 1982; Kehoe, 1986; Kehoe & Graham, 1988; Weiss, 1972). However, studies with human participants have presented a mixed picture. Birnbaum (1976), for example, used numbers as predictors of other numbers. In this task, Birnbaum (1976) found a strong tendency to average (see also Birnbaum & Mellers, 1983). Similarly, Downing, Sternberg, and Ross (1985) found a strong tendency to average with abstract material in which causes and effects were represented by letters. This tendency was weaker with more concrete material but still was used by many participants. In these tasks, participants combined the influence of multiple causes. Blocking studies, an example of analysis of causal contributions, have also been conducted. In most studies, the findings were consistent with the hypothesis that people used an additive rule (see e.g., Chapman & Robbins, 1990; Shanks & Dickinson, 1987; Sobel et al., 2004; Waldmann, 2000; Waldmann & Holyoak, 1992), although there are also exceptions (Williams, Sagness, & McPhee, 1994).

In summary, most learning theories solve the computational problems of learning with sparse data by assuming a fixed additive integration rule as the default. Although other rules, especially averaging, have been explored empirically, very little is known about the factors that determine the choice of integration rules in causal tasks.

1.2.1. Biasing integration rules: The effect of abstract prior knowledge

Additive integration of the probabilistic causal strength of generative binary causal events is a plausible default rule. A less constrained scenario, which I investigated in my studies, are causal situations with continuously varying causes and effects. Depending on the type of causal mechanism, various integration rules, including adding and averaging, may be appropriate. If, for example, the compound of two causes generates an effect of size +7 (on a scale) and one element causes the same effect, then the other cause should produce an effect of 0 if integration is additive. However, 7 is the correct prediction in domains in which the causes are averaged.

In my research, I focused on the question of how people select an integration rule. According to causal-model and related theories (Lagnado et al., in press; Tenenbaum & Griffiths, 2003; Tenenbaum et al., in press; Waldmann, 1996; Waldmann et al., 2006), more or less abstract domain knowledge may place constraints on the selection of the integration rule. One plausible

hypothesis is that people use their possibly rudimentary domain knowledge to decide between alternative integration rules. Often, simple cues that remind learners of classes of physical phenomena may influence which integration rule is chosen for the causal model. In some of the following experiments I discuss, I focus on two types of physical quantities as cues. *Extensive* quantities, such as the amount of a fluid, vary with volume. If two fluids are mixed together, the volume increases. The two volumes need to be added to predict the volume of the mix. By contrast, an *intensive* quantity, such as the color or the taste of a fluid, is not sensitive to the amount. A mix of two equally colored fluids will still have the same color. Intensive quantities are dependent on proportions, for example, the relation of color particles to the volume of a fluid. This property is the reason why the result of mixing intensive quantities is a weighted average of the components. Mixing two fluids that are light and dark blue will result in a blue that lies in between these two shades. The distinction between heat and temperature is another example. Whereas heat is an extensive quantity, temperature is an intensive quantity (see Wiser & Carey, 1983).

A number of studies have investigated what people, especially children, know about intensive and extensive quantities. Often young children misrepresent intensive quantities such as sweetness or temperature as an extensive quantity (see Moore, Dixon, & Haines, 1991; Strauss & Stavy, 1982). Reed and Evans (1987) investigated students. Reed and Evans' experiments show that these students knew how to predict temperatures but had difficulties with acids unless they were presented with the analogous domain of temperatures first. In general, these studies have shown that adults have knowledge about the difference between intensive and extensive quantities at least in some more familiar domains.

Whereas research on intuitive physics investigates already acquired domain knowledge, in my studies, I focused on learning about novel domains for which no prior knowledge about physical mechanisms is available. For example, participants learned that a novel blue fluid causes an increase of the heart rate of +3 (on a scale that ranges between 0 and +12), whereas a yellow fluid increases the heart rate to the level +7. The crucial test question was what heart rate people expect when both fluids are mixed together. Will they go over the value of +7, which would indicate an additive strategy, or will they rate the heart rate between the two values, which would indicate averaging? My main manipulation was to present additional cues that should have influenced the integration strategy. I expected cues that indicate an extensive quantity (e.g., fluids as drugs) should have cued learners into assuming additive integration, whereas intensive quantities (e.g., taste of fluids) should bias them toward averaging.

Unlike in the research on intuitive physics, there is no prior knowledge about the influence of the colored fluids on heart rate, and it is far from obvious that real fluids would have these characteristics. Different drugs often interact in unpredictable ways, and it is not easy to predict the taste of a compound on the basis of two differently tasting components (see De Graaf & Frijters, 1988). Still, it seems plausible that people associate some abstract physical features with differently characterized fluids, which should bias them toward different parameterizations of the underlying causal model. The integration rules activated by abstract prior knowledge represent rough guesses based on hunches, not stable knowledge. Therefore, it seems likely that learners are also influenced by additional cues.

1.2.2. Biasing integration rules: The effect of learning data

Integration rules are not only affected by prior knowledge but also by the learning data. With feedback, people can easily learn different integration rules (Koh & Meyer, 1991). In my studies, I investigated the more interesting case that learning data does not fully constrain the learning rule. If people predict the effect of a combination of causes they have never observed before, different integration rules are consistent with the data. Nevertheless, the kind and sequence of data may bias the rule. I investigated this possibility by comparing two different tasks within the same domain that both require a choice of an integration rule. In the *combination tasks*, participants learned the causal impact of two causes of a common effect separately and subsequently, I asked them to predict the outcome of combining the two causes. Thus, in this task, people had information about individual causes as data, and I requested them to reflect about a combination they had never seen. I contrasted this task with an *analysis task* in which the sequence is reversed. For example, participants may have learned that the blue fluid causes a heart rate of +3 and the blue and yellow fluid combined cause a heart rate of +5. Then I asked participants to infer the causal strength of the yellow fluid, which they had never observed individually.

If only domain related cues drive the choice of the integration rule, identical rules should be invoked in both tasks. Thus, if people assume an additive rule in combination tasks, they should choose a rating below +5 in the analysis task. If an averaging rule is assumed in the combination task, then their ratings should be above +5 in the analysis task. The way causes interact in the real world is not affected by the kind of inference one draws.

Indeed, studies on the representation of temperature mixtures have shown that at least adults can make the correct inferences independent of the task context (Stavy, Strauss, Orpaz, & Carmi, 1982). This is a familiar domain to most adults. However, in other areas, the invariance not always holds. There are studies that have demonstrated that people have difficulties when the combination rule implies multiplication. For example, Anderson and Butzin (1974) found that people multiply ability and motivation when predicting performance but subtract ability from performance when predicting motivation. Similar findings have been obtained for predicting the behavior of balance scales (Surber & Gzesh, 1984).

These results indicate that the choice of the integration rule is not solely driven by prior knowledge but also by the data and the required task. Apparently, prior knowledge dominates in domains in which people have rich and stable domain knowledge such as in the domain of temperatures. In these domains, the tasks are based on a multitude of experiences so that responses inconsistent with the familiar rule can be recognized as false and in need of a correction. In my study, I was concerned with learning tasks in which no stable prior knowledge is available. Although abstract domain knowledge, such as assumptions about the kind of physical quantity, may bias a person's choice, the selected rule is not entrenched in rich domain knowledge and supported by prior experiences so that other factors may also play a role.

Previous research (Anderson & Butzin, 1974; Surber & Gzesh, 1984; see also Anderson & Wilkening, 1991) has suggested that people indeed often tend to invoke different integration rules in the combination and the analysis task. How can this be explained? In the combination tasks in the following experiments, the inferences were based on knowledge about the strength of two separate causes, with one cause being stronger than the other. Without any prior

assumptions, there are no clear biases this task exerts on the inference for the combined cause. If the strength of both causes is graphically indicated on a continuous rating scale (lower values on the left side) the combined effect could either be placed in the middle zone between the causes (i.e., averaging) or to the right of the stronger cause (i.e., adding). The data does not clearly bias any of these two solutions, which suggests that prior knowledge will have a strong biasing influence in this task. As the experiments I discuss in the following sections showed, the source of these biases can either be abstract domain knowledge or prior experiences with similar tasks (i.e., transfer).

In the analysis task, I provided information about one cause A along with information about the combination of this and a novel cause X (i.e., $A + X$). Again, this task can be visualized by imagining a scale on which A and $A + X$ are placed on the same scale. The fact that the combination contains the component cause A along with an additional unfamiliar cause X highlights the additive structure underlying the compound of A and X (A vs. $A + X$). If A generates a specific size of the effect, then the salient additive relation between the component and the compound highlights the possibility that the difference between the effect of A and the effect of $A + X$ should be attributed to X. Thus, in this case, the additive structure of the relation between A and $A + X$ may bias the way the relation between the two outcomes is processed.

In fact, the strength of this data-driven bias can also be seen in the history of causality research. For many years, both philosophers and psychologists have argued that the difference between the probability of the effect in the presence and the absence of the cause (i.e., contingency or delta-p rule) is the best measure of causal strength (Cheng & Novick, 1992; Salmon, 1980). The intuitive reasoning behind this proposal is that the presence of the cause adds this factor to the constantly present background. The comparison between the cause plus background to background alone highlights the possibility that the observed difference of the effect probability should be attributed to the cause. For many years, this rule was proposed as normative until Cheng (1997) showed that the rule is based on faulty reasoning. It neglects the fact that there is an upper and lower ceiling for probabilities that should be taken into account. Thus, even philosophers and scientists are not immune to the biasing effect of data.

2. Experiment 1

In this experiment, I focused on the combination of different causes. Participants learned about the effects of different colored fluids on animals' heart rates. This was a learning task because no prior knowledge about this relation was available. I presented both causes and effects as continuous variables. More specifically, participants learned that blue fluids cause a heart rate of $+3$ and yellow fluids a heart rate of $+7$. In the test question, I asked learners to predict the heart rate if both fluids are filled into a large container and mixed. Because there was no feedback in the test phase, any answer was possibly correct. However, my hypothesis was that participants would use cues that reminded them of physical knowledge to choose an integration rule.

The crucial manipulation was whether the cues reminded participants of extensive or intensive quantities. In the extensive quantity condition, I told participants that the fluids

represented drugs that can have different strengths. In this condition, I expected people to favor an additive integration rule in which the strength of the drugs that were mixed together would amount to a weighted sum (i.e., ratings $> +7$). In the intensive quantity condition, I also introduced the fluids as drugs, but I told participants that it is assumed that the heart rate is sensitive to the taste of the drugs. Thus, I expected a preference for choosing a weighted average rule (ratings $< +7$) when predicting how the taste of the mixture would affect the heart rate.

Especially in the latter case, it was clear that this can only be a hypothesis. In people's everyday experience, mixtures of differently tasting fluids may result in all kinds of tastes, most of them probably horribly tasting. Nevertheless, taste is an intensive quantity, and this type of quantity is typically associated with proportional reasoning. Furthermore, I assumed that this kind of hypothetical knowledge is less stable than physical knowledge one has about familiar domains (e.g., temperatures). This should lead to sensitivity to other kinds of cues. In this experiment, I added a second condition in which an additional cue was given in both conditions. In this condition, participants were told that the strength ("extensive quantity condition") or the taste ("intensive quantity condition") were based on the quantity of a specific substance called "corium," which was dissolved in the fluids. Because corium was introduced as an extensive quantity, this may have led at least some participants to represent the mechanism as based on an extensive quantity. Thus, I expected an increase of the use of the adding rule in both the strength and the taste conditions.

2.1. Method

2.1.1. Participants and design

A total of 96 students from the University of Frankfurt/Main, Germany, participated in the experiment. We randomly assigned participants to one of the four conditions generated by the factors type of cue (intensive vs. extensive quantity) and substance (corium vs. no corium). The research assistant tested all participants individually.

2.1.2. Materials and procedure

Participants first received written instructions that mentioned new drugs. In the intensive quantity condition, the instructions stated that the researchers intended to improve the taste of the drugs to increase sales. To test the drugs, they were given to animals. The researchers knew that animals' heart rate would rise the better the drug tasted. Neutrally tasting drugs would leave the heart rate unaffected. We then introduced participants to the rating scale that ranged from 0 (*normal heart rate*) to 12 (*very strong heart rate*). The tick marks were all numbered. In the extensive quantity condition, the instruction stated that the researchers had the goal to develop a drug that increases performance and would excite the physiology of the body. Neutral drugs would leave the heart rate unaffected, but the more the drugs excited the body, the higher the heart rate would be. Otherwise, the instructions and rating scales were identical in both conditions.

The conditions in which the substance corium was mentioned were almost identical except that in both conditions, participants were told that corium underlies the taste (intensive quantity condition) or the strength (extensive quantity condition) of the drugs. Moreover, the

instructions stated that the more corium (in milligrams) was dissolved the better the drug would taste or the stronger it would be.

As learning materials, we presented participants with cards on which colored half circles (diameter: 3.5 cm) could be seen. We mentioned that these pictures represented containers filled with colored fluids (the drugs). We used the colors green, yellow, blue, and purple. We counterbalanced the choice of the colors and the sequence. We then told participants that an animal drank the whole container of the colored fluid, which caused a heart rate of +3. To remind participants of the instruction, we mentioned again that the heart rate is sensitive to the taste or the strength of the drug. Then we presented a second item, which caused a heart rate of +7. We reminded all participants that the heart rate can go up to +12.

In the test phase, we repeated these two trials, and participants had to predict without feedback the heart rate the two fluids would produce. We then told participants that now the two fluids were filled into a bigger container without leaving any of the rest in the smaller containers. Furthermore the instructions stated that new test animals drank the whole container of the mixed fluid. We emphasized that animals drank everything to make sure that they did not assume that only a portion of the drug was drunk, which should affect the extensive quantity condition. For half of the groups, we reversed the test questions for the elements and the compound so that they were asked about the compound first.

2.2. Results and discussion

Table 1 shows the mean ratings in the four conditions. In general, the ratings in the extensive quantity conditions (strength) were higher than in the intensive quantity conditions (taste). Also, there were higher ratings when corium was mentioned. A 2 (type of cue) × 2 (substance) analysis of variance revealed a significant effect for cue, $F(1, 92) = 25.3$, $p < .001$, mean square error (MSE) = 3.64, as well as substance, $F(1, 92) = 6.59$, $p < .05$, $MSE = 3.64$. The interaction was not significant ($F < 1$). This pattern was consistent with the hypothesis that there would be more averaging when cues were given that suggest intensive quantities and less averaging when an extensive quantity was offered (i.e., corium) as an explanation of both types of causes (strength vs. taste).

The analysis of the differences of mean ratings provides information about the size of the rating differences between the conditions but not about which strategies were chosen. Therefore, I additionally classified participants according to their strategy. If the rating for the effect of the compound was between +3 and +7, I diagnosed an averaging strategy. If the rating was higher than +7, I classified the behavior as indicating adding. Ratings of +3 indicated a minimum (min) strategy, and ratings of +7 a maximum (max) strategy. These

Table 1
Mean ratings of causal strength of the compound

	Without Corium	With Corium
Taste	5.46	6.63
Strength	7.58	8.42

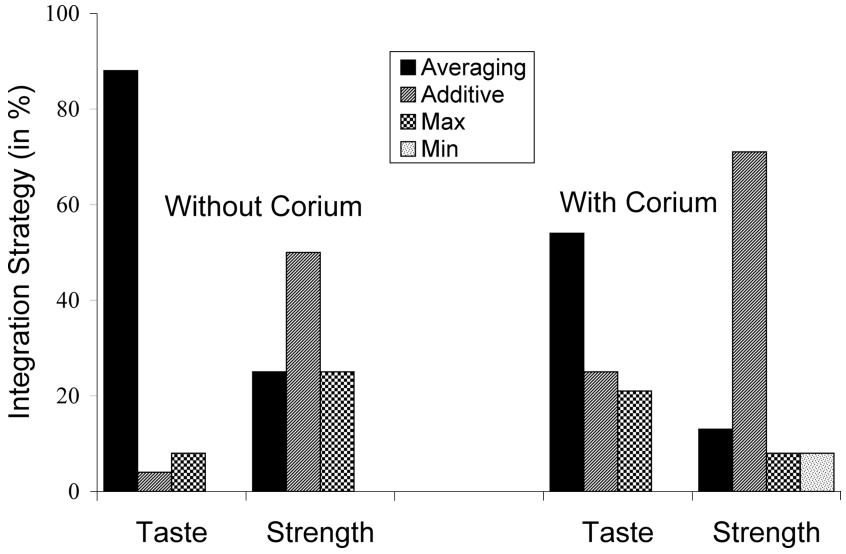

Fig. 1. Selected integration strategy in the intensive (taste) and extensive (strength) quantity conditions in Experiment 1. Max = maximum; Min = minimum.

participants seemed to believe that the combined effect was solely determined by either the weaker or the stronger cause.

Figure 1 shows the results. It can be seen that averaging dominated in the intensive quantity condition, and adding dominated in the extensive quantity condition. However, mentioning corium had a clear effect in both conditions. The number of people who added increased. The min strategy was observed rarely, but there were some participants who used the max strategy, which may be seen as a special case of either averaging or adding. The choice of strategies significantly varied in both the conditions without corium, $\chi^2(2, N = 48) = 19.6$, $p < .001$, in which nobody used the min strategy, and the conditions in which corium was mentioned, $\chi^2(3, N = 48) = 14.8$, $p < .01$. In this condition, all four strategies were observed (two participants chose the min strategy).

The results clearly confirm my hypotheses. It is interesting to see how sensitive participants were to the extensive cue corium, which led many to assume additive integration, although the effect of the drug was, according to the instructions, still based on the taste of the fluid. This reveals that people only have unstable assumptions about the underlying mechanisms. Because participants did not have firm prior knowledge about the relation between the colored drugs and the heart rate, they could not rely on intuitive domain knowledge. Nevertheless, it is clear that there was no general preference for adding. People chose integration rules on the basis of cues that reminded them of abstract characteristics of familiar domains.

3. Experiment 2

In this experiment, I focused on the conditions in which no corium was mentioned. Instead of a combination task, I used an analysis task. Here, I asked participants to infer the causal

impact of one element when they had learned about the compound and the other elemental cause. The question was which integration rule participants would choose in this analysis task. If participants were solely driven by the cues that suggested different causal mechanisms, they should have chosen the same integration rule as in the combination task. However, the corium condition in Experiment 1 already provided evidence for the instability of the underlying knowledge. Thus, I expected that other factors, such as the highlighting of the difference between compound and element in the presentation of the data for the analysis task, may affect the choice of an integration rule.

3.1. Method

3.1.1. Participants and design

A total of 48 students from the University of Frankfurt participated, with half of this group randomly assigned to the intensive and half to the extensive quantity condition.

3.1.2. Materials and procedure

The instructions and materials were taken from Experiment 1. The only difference was that we changed the sequence of the learning and test trials. First, participants learned that the strength (extensive quantity condition) or the taste (intensive quantity condition) of one of the fluids caused a heart rate of +3. Then the experimenter showed the second card with the other fluid and said that participants should imagine that both fluids are filled into a large container and mixed. It was stated that the animals drank the whole container, which caused a heart rate of +5. After two learning trials, the two fluids were shown individually, and participants were asked about the effects of either fluid. The answer for one fluid was already known from the learning phase, but the answer for the second fluid had to be inferred (without feedback). We counterbalanced the sequence of these test trials.

After the analysis task, we presented participants with a third colored fluid that caused a heart rate of +9. Then we asked participants to rate the strength of the effect when this fluid is mixed with the first fluid (+3). This was a standard combination task, which served as an additional test.

3.2. Results and discussion

Table 2 shows the mean ratings for both the analysis and combination tasks. First, I analyzed the analysis task (first row). The mean values showed significantly higher values for

Table 2
Mean ratings of causal strength of the compound (combination) or the second element (analysis)

	Taste	Strength
Analysis	3.71	2.21
Combination	8.21	11.33

the intensive than the extensive quantity condition, $F(1, 46) = 8.57$, $p < .01$, $MSE = 3.15$. This indicates sensitivity to the physical cue. Higher values suggest that more people used an averaging rule. However, both means were clearly below +5, which is consistent with a tendency to add in both tasks.

By contrast, the very same participants showed a different pattern in the final combination task (second row in Table 2). Again, there was a significant difference that suggested more adding in the extensive than the intensive quantity condition, $F(1, 46) = 24.1$, $p < .001$, $MSE = 4.85$. However, now the mean value for the intensive quantity condition fell below +9, which suggested averaging, and the mean for the extensive quantity condition was above +9, which suggested adding. Thus, I found a within-subjects dissociation. Whereas participants were sensitive to the physical cues in the final combination task, they seemed to gravitate toward additive integration in the initially presented analysis context.

Figure 2 breaks down the strategies of participants. I observed two strategies in the analysis and three in the combination context. I used the same method as in Experiment 1 to classify participants (except that +9 was then the reference point for the combination task). The figure shows that the majority of participants chose an additive integration rule in the analysis task, although also a small effect of the type of cue (taste vs. strength) can be seen, $\chi^2(1, N = 48) = 5.4$, $p < .05$.

The subsequent combination task showed a similar pattern as Experiment 1, with a prevalence of adding in the extensive quantity condition and of averaging in the intensive quantity condition, $\chi^2(2, N = 48) = 17.5$, $p < .001$. There were more people who used an additive strategy in this experiment, which might be due to transfer effects between the two tasks. Nevertheless, it is interesting to see that a significant number of people switched strategies between the two tasks.

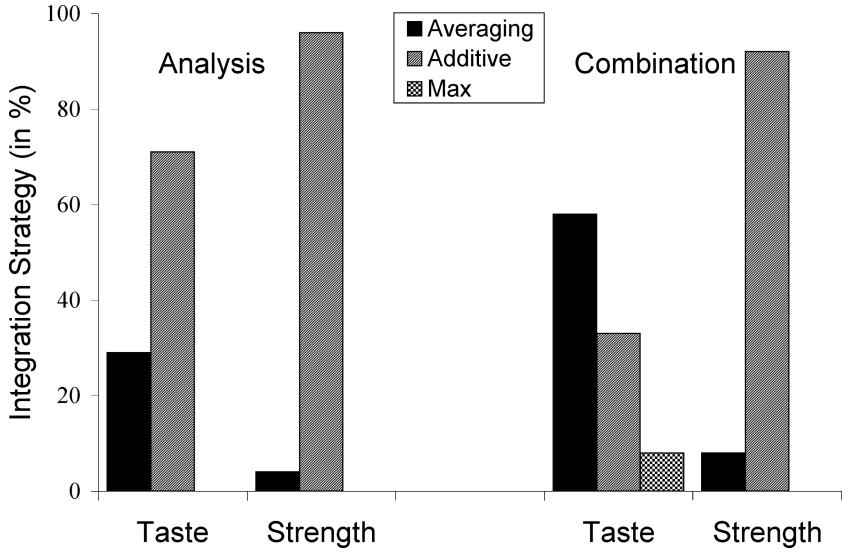

Fig. 2. Selected integration strategy in the intensive (taste) and extensive (strength) quantity conditions in the analysis versus the combination task in Experiment 2. Max = maximum.

The dissociation demonstrates again that participants could not rely on stable knowledge but were affected by cues. Apart from the influence of domain-related cues, this experiment demonstrated that the kind of learning data, the task, and transfer from previous tasks also affect the selected integration rule.

4. Experiment 3

In Experiment 2, I used a within-subjects design to compare analysis and combination, with the latter task always following the first task. In Experiment 3, I used a between-subjects design to compare these two tasks. Half of the group of participants started with a combination task and half with an analysis task. Additionally, I presented a different second combination task in both conditions to be better able to test whether there is transfer between the two consecutive tasks.

A second goal of this experiment was to test my hypotheses in a different causal domain. In this experiment, I chose liking as a mechanism underlying the causal relations. Participants saw the same stimuli as in the previous experiments but were told that the heart rate of the test animals was sensitive to how much they liked the colored figures. I chose liking because Anderson (1981) found a preference for averaging and proportional reasoning in such judgments (with more natural stimuli). A prevalence of averaging in the combination task is a precondition of a possible dissociation between combining and analyzing.

However, whereas previous studies have focused on judgments of familiar domains, I used this mechanism in a novel learning context. There are arguably even less constraints on the generation of liking judgments than in the domain of tastes. There is no rational reason to assume that people who like blue and yellow half circles differently would average when presented with a full circle that is half yellow and half blue. All kinds of outcomes seem possible. Also, if the heart rate is dependent on viewing colored circles, it may very well seem plausible that a full circle would generate a faster heart beat than each of its components (i.e., adding). Nevertheless, based on previous research, I expected that liking would cue participants into a preference for averaging causal influences but that this knowledge might also be unstable and might therefore be affected by additional cues.

4.1. Method

4.1.1. Participants and design

A total of 48 students from the University of Frankfurt participated; we randomly assigned half of this group to the analysis and half to the combination condition.

4.1.2. Materials and procedure

The materials and the task were largely identical to the ones we used in the previous experiment. Moreover, we used similar instructions except that we told participants that the colored figures represented geometric figures (instead of liquids), and that animals' heart rates were sensitive to how much they liked the figures. Otherwise we used the same colored half-circles and numbered rating scales as in the previous experiments.

In the combination condition, we told participants that one of the half circles causes a heart rate of +3 and the other a heart rate of +7 (with the colors being counterbalanced). These learning trials were repeated once. In the subsequent test phase, we presented full circles that had the same radius as the half circles. The full circles combined the two differently colored half circles. Otherwise, the test phase was similar to Experiment 1; we asked half of the group about the full circle first prior to the two half circles. At the end, we presented a second combination task in which, according to the learning feedback, a 3rd half circle caused a heart rate of +9.

The analysis condition was analogous to Experiment 2. Participants learned that one half circle causes a heart rate of +3 and the full circle a heart rate of +5. We counterbalanced the order of these two trials. We repeated this sequence before the test phase started. In the test phase, we asked participants about the causal effect of each component of the full circle (i.e., the half circles). We again counterbalanced the sequence. Subsequently, we presented a 3rd half circle that caused a heart rate of +9. In the test phase, participants had to predict the effect for the full circle that contained this new half circle and the one that caused a heart rate of +3.

4.2. Results and discussion

In this experiment, my goal was to compare the integration strategies that are used in combination and analysis tasks. Therefore, my analysis focused on the classification of the strategies the individual participants chose. Figure 3 shows a breakdown of the strategies used in the two different tasks. The left side depicts the results of the first combination task in which the majority of participants chose an averaging integration rule. Some participants preferred the max rule. In contrast, the analysis task (also shown on the left side of Fig. 3) showed a

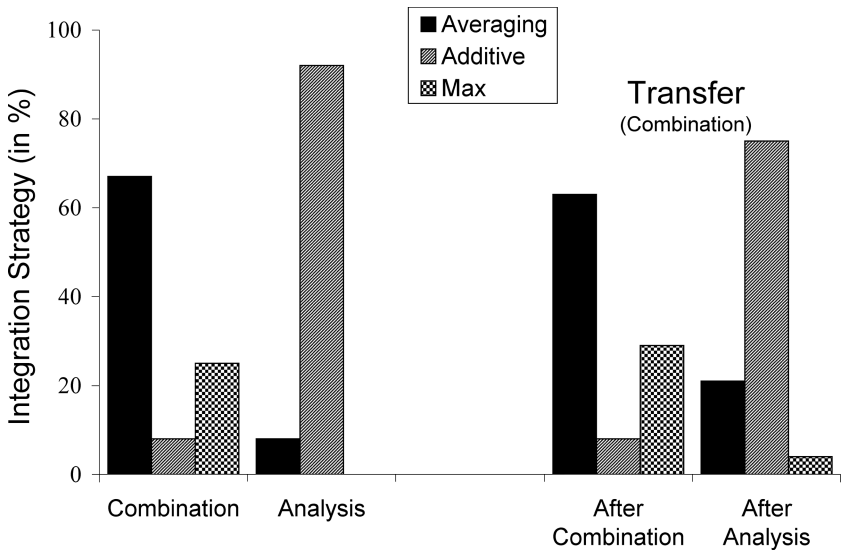

Fig. 3. Selected integration strategy in different task contexts in Experiment 3 (left). The right side shows the results of the final combination task. Max = maximum.

different pattern. About 90% of the participants preferred an additive integration rule. The difference in the choice of strategies between the two conditions was highly significant, $\chi^2(2, N = 48) = 33.56$, $p < .001$. Thus, the asymmetry between combination and analysis turned out to be even stronger than in the previous experiment. This may be due to the fact that domain knowledge underlying liking judgments may be even less stable than knowledge about factors affecting taste, which could strengthen the role of additional factors.

The right side of Fig. 3 demonstrates the role of transfer between the initial task (combination or analysis) and the subsequent different combination task. The strategy choices differed depending on the type of initial task, $\chi^2(2, N = 48) = 22.3$, $p < .001$. Whereas participants preferred an additive integration in the second combination task when the initial task required an analysis, averaging dominated when the initial task cued into averaging. Only 17% switched from adding in the first analysis task to averaging in the subsequent combination task. Thus, there were stronger transfer effects than in the previous studies, which may reflect the greater uncertainty of participants about the proper integration rule in liking judgments.

5. Experiment 4

Experiment 4 consists of two separate studies, Experiments 4a and 4b. In Experiment 4a, I used taste, and in Experiment 4b, I used liking as causes of animals' heart rate. Otherwise, both studies were identical. The main goal of Experiment 4 was to replicate previous results using a different dependent measure. In the previous experiments, I had used numbered rating scales. This raises the question whether the obtained asymmetry may be a consequence of differences in the difficulty of symbolic arithmetic operations. The inversion of averaging in the analysis task may be mathematically more difficult than the inversion of adding, which may have led to a switch to a simpler rule in the analysis task. In my view, this hypothesis was not very plausible as the sole explanation of the dissociations because I had used university students as my participants who certainly have a grasp of averaging (and its inversion). This was shown in the study by Stavy et al. (1982). In the highly familiar domain Stavy et al. studied, participants had no problem with inversions of averaging.

Nevertheless, Brunswik (1956) proposed the hypothesis that analytic reasoning, which is based on number representations, may generally differ from intuitive reasoning, which is invoked in domains that one does not represent with analytic tools. Therefore, it seems interesting to study the effect of a more intuitive measure in one's tasks. Research on the difference between analytic and intuitive reasoning in intuitive physics has presented a mixed picture, with none of the two modes being clearly superior in all domains (see Ahl, Moore, & Dixon, 1992; Anderson, 1987; Budescu, Weinberg, & Wallsten, 1988; Hammond, Hamm, Grassia, & Pearson, 1987).

To test whether the effects were solely due to analytic reasoning with numbers or whether they can also be found with more intuitive causal reasoning, I switched to a nonnumeric, graphic rating scale. These scales had neither numbers nor tick marks so that counting or other processes of symbolic measurements were not invited. Intensity was signaled by an arrow that pointed to a specific point on a line, which was introduced as graphically representing the

intensity of the effect (i.e., heart rate) starting with normal values on the left side of the line. This manipulation is motivated by findings that have shown that symbolic manipulations with real numbers are ontogenetically and phylogenetically preceded by the capacity to approximate precise numerical reasoning using mental manipulations of numerosities on an analog "number line" (Dehaene, 1997). Recent research has shown that these analog mental manipulations are based on brain circuitry different from the ones activated in symbolic arithmetic reasoning (Dehaene, Molko, Cohen, & Wilson, 2004).

My main goal was to test whether the asymmetry between combination and analysis tasks in identical domains can be replicated with nonnumeric scales that discourage algebraic operations but rather encourage analog reasoning on a number line. This way, I intended to investigate causal reasoning independent of possible differences in difficulty of symbolic numeric operations.

5.1. Method

5.1.1. Participants and design

In each of the two experiments (Experiment 4a and 4b), we investigated 96 students from the University of Frankfurt. We assigned half of these groups to the combination and half to the analysis condition.

5.1.2. Materials and procedure

For Experiment 4a, we used the instructions and materials from Experiments 1 (combination task) and 2 (analysis task). We only used the instruction that the causal effect (heart rate) depended on the taste of the drug. In Experiment 4b, we used the cover stories and materials from Experiment 3. Thus, participants learned that the heart rate was sensitive to how much the animals liked the geometric figures. The key difference to the previous experiments was that instead of numeric rating scales, we used graphic scales in which the left end mark of the scale was labeled "normal" and the right end mark "very strong." There were neither numbers nor tick marks, just a continuous line. We told participants that the points between the end points represented different degrees of strength.

The learning task was similar to the ones used in the previous experiments. The main difference was that the intensity of the heart rate was expressed by arrows pointing to specific points on the scale. These points corresponded to the positions of the numbers (+3, +5, +7) in an equidistant scale. In the combination task, we counterbalanced the sequence of the stronger and weaker causes. In the analysis task, we counterbalanced whether the compound or the element was presented first. Each learning task consisted of two trial types, which were repeated three times. Participants had to point to positions on the scale and received immediate feedback (scales with arrows) about the correct position.

In the test phase, we handed out rating sheets with three scales that did not show any arrows. Participants' task was to mark the positions with a pen that corresponded to the effect the two elemental causes and the combined cause would generate. The critical response was the rating of the compound in the combination task and the rating of the second element in the analysis task. We counterbalanced the sequence of these two responses in the analysis task.

5.2. Results and discussion

Again, I was interested in the differences in the choice of strategies. I coded the rating sheets analogous to the previous studies. The crucial question was how the three arrows in the test sheets were located relative to each other. If the arrow for the compound cause was pointing at a segment in between the two other arrows, averaging was diagnosed; when it was pointing to an intensity that was higher than the two others, then this was seen as evidence for adding. Similarly, the analysis task required a decision as to whether the second element was rated more intense (averaging) or less intense (adding) than the compound. If the arrows for the element and compound cue were the same, the max or min rule was diagnosed.

Figure 4 breaks down the selected strategies in Experiment 4a (left) and 4b (right) in the two tasks. The most important result is that the difference between combination and analysis was replicated in both experiments. The strategies varied depending on the task in both Experiment 4a (taste), $\chi^2(3, N = 96) = 12.87$, $p < .01$, and Experiment 4b (liking), $\chi^2(3, N = 96) = 9.41$, $p < .05$. As in the previous experiments, I observed more averaging in the combination task than in the analysis task. This shows that this effect cannot be attributed to asymmetries in the difficulty of arithmetic operations. In these experiments, we used a graphic rating scale that neither contained numbers nor tick marks.

The effect was somewhat smaller in these two experiments than in the previous ones. In Experiment 4a, one can see for the first time that averaging dominated in both conditions, although there was still a clear, highly significant difference. In Experiment 4b, I found the usual disordinal interaction as in Experiment 3. The weakening of the effect, especially in Experiment 4a, may indicate that at least some participants in the previous experiments may have used numeric representations and arithmetic procedures and that difficulty of the arithmetic procedure may also have affected the choice of integration rules. However, the

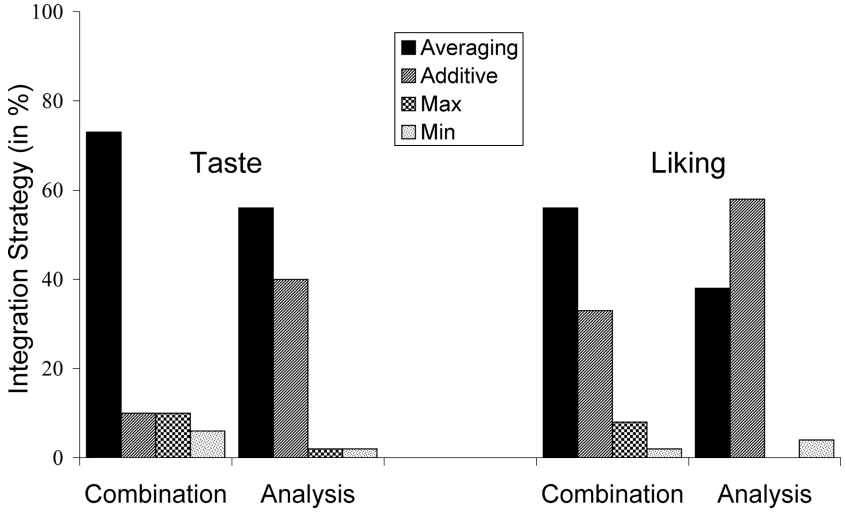

Fig. 4. Selected integration strategy in different task contexts in Experiment 4a (taste; left) and 4b (liking; right). Max = maximum; Min = minimum.

replication of the dissociation between combining and analyzing with nonnumeric scales suggested that the difficulty of arithmetic operations was at best responsible for portions of the effect. The assumption that the obtained dissociation was not entirely caused by asymmetries in the difficulty of arithmetic procedures was also suggested by the different patterns of the obtained interactions in Experiments 4a and 4b. If the dissociation was solely caused by some learners trying to estimate numeric measures from the continuous scale and using arithmetic procedures, the same deviations from the patterns from the ones obtained in the previous studies should have been observed in both experiments. Whereas I got a switch to an ordinal interaction in Experiment 4a, the previous disordinal interaction was largely replicated in Experiment 4b. Moreover, no participant spontaneously reported to have tried to measure the distances on the scale using number estimates. It seems more likely that most participants reasoned using an analog representation of the size of the effects (along the lines of Dehaene's, 1997, number line).

The difference between Experiment 4a and 4b was consistent with the observation that the domain of liking was less constrained by prior domain assumptions than the domain of taste so that domain assumptions dominated in the taste condition, whereas the liking condition was more strongly affected by other task factors (see also the discussion of Experiment 3).

6. General discussion

The causal texture of the world contains complex networks in which effects are influenced by several causes. Learning all these relations and possible interactions would require samples of observation that present all empirically possible patterns of causes and effects. In everyday life, people typically only observe events relating to fragments of causal networks. From this knowledge, people have to infer how novel patterns of causes probably would behave. The research I report on in this article addresses the question of how people intuitively integrate multiple causes with respect to a joint effect.

There is little work on this topic in the literature, but most theories of causal induction assume a bias toward additivity and linearity of causal influences (see Introduction section). Our experiments have shown that people draw on analogies from prior knowledge when selecting an integration rule. For example, it was shown that people tended to average causal influences when the causes were described in terms of intensive physical quantities (taste) or when a psychological preference (liking) was the basis of the effect. In contrast, extensive quantities as causes generated a preference for adding.

Unlike research on intuitive theories, which focuses on existing knowledge, my interest was in learning of novel relations (fictitious colored drugs that cause heart rates). Thus, participants could not draw on stable knowledge. In these situations, prior knowledge only provides hunches, which might very well turn out to be wrong. For example, it is far from necessary that the taste of a mixed fluid corresponds to the average taste of its components or that animals' liking of a full circle with two colors corresponds to the average of its components. Nevertheless, lacking more substantial knowledge, people seem to guess the most appropriate integration rule. These guesses are easily overturned though if additional cues are present that are associated with different rules.

Experiment 1 provided first evidence for the instability of these guesses. Mentioning an extensive quantity (i.e., corium) generated an increase of additive responses in all conditions, although the general difference between extensive and intensive quantities stayed intact.

The instability of the knowledge underlying the choices may also be the reason for the surprisingly large context effects I observed. I compared a combination task in which participants learned about individual causes first and then predicted the effect of the compound with an analysis task in which participants inferred the probable effect of an element after having learned about the compound and the other element. In general, I found a preference for adding in the analysis task even when there was a preference for averaging in the corresponding combination task. I found this dissociation within and across groups of participants and across different measurement methods (numbered vs. graphic rating scales). Inconsistencies have often been found when knowledge is unstable (see Reed & Evans, 1987; Surber, 1987), whereas in more familiar domains, people are better at being consistent across different tasks (Stavy et al., 1982).

This asymmetry shows that participants' selections of integration rules, apart from being influenced by domain-related cues, were also influenced by other factors such as the way the data are presented or previous tasks. This is consistent with the assumption that domain knowledge is only one of many fallible cues learners use to select an integration rule. For the analysis task, my hypothesis was that the way the data are presented in this task highlights the difference between the effect of the compound and of one of its components, which suggests that the second component is probably responsible for this difference. The most convincing evidence for this hypothesis comes from the experiments in which I requested nonnumeric ratings. Especially in the liking task, I observed a strong disordinal interaction between combining and analyzing tasks with a task that did not encourage numeric, symbolic reasoning.

The contrast between the experiments requesting numeric and nonnumeric ratings suggested arithmetic difficulty as an additional factor being responsible for the asymmetry between combination and analysis tasks. Although previous experiments have shown that in some domains people are capable of correctly inverting averaging (e.g., Stavy et al., 1982), it may well be that in domains with less stable domain intuitions, even small differences in the difficulty of arithmetic operations affect the choice of integration rules. One reason for the differences in performance between domains might be that in familiar domains, people might immediately recognize that their intuitive response is wrong. If people learn that the fluid in one glass has a temperature of 20°C and the mixture a temperature of 26°C, people might be immediately aware of the fact that it cannot be possibly true that the second component has only a temperature of 6°C. Such control processes are less likely in the less familiar domains I chose. The fact that difficulty of procedures might also affect people's causal reasoning strategies was also confirmed in the studies by Waldmann and Hagmayer (2001). In Experiment 2 of this set of studies, we showed that the difficulty of the task influenced whether people controlled for a cofactor when assessing causal strength.

Other cues affecting the integration rule can come from previous tasks (i.e., analogical transfer). Both Experiments 2 and 3 provided evidence for the influence of the previous task. In both experiments, the choice of the integration rule in a final combination task was affected

by the initial task. When this task required analyzing, a greater tendency to add was observed regardless of the domain.

The research on domain knowledge (e.g., intuitive physics) suggests that domain assumptions should be the normative basis for responses. For example, if temperatures should be averaged, this rule should be used regardless of the task and the context. According to this view, the obtained differences between combining and analyzing appear irrational. However, in the tasks I chose, there was no clear correct response in any of the tasks. For example, combining two differently tasting liquids may have all kinds of effects on the taste of the compound (imagine mixing fish and ice cream). Similarly, it seems unlikely that the global assessment of the aesthetic value of a painting can be explained by averaging how much people like sections of the painting. Although it seems reasonable to use domain cues for rough estimates, learners are probably aware of the fallibility of these cues. This explains why other cues, apart from domain-related ones, also affect the choice of integration rules and may even override a person's domain assumptions.

In this article, I started by discussing the current debate between domain-specific and domain-general theories of causal induction (Cartwright, 2004; Gopnik et al., 2004). The results of our experiments suggest a middle position that is captured by causal-model theory (Lagnado et al., in press; Waldmann, 1996; Waldmann et al., 2006) and theory-based Bayesian theories (Tenenbaum & Griffiths, 2003; Tenenbaum et al., in press). The results of the experiments demonstrate that it is necessary to differentiate between different levels of prior knowledge. For some domains, people may have stable, thick (Cartwright, 2004) prior knowledge that can directly be the medium of causal reasoning without having to translate the tasks into more abstract causal representations. In such domains, knowledge dominates and reasoning is little affected by other task-related factors (e.g., Stavy et al., 1982). This kind of knowledge has probably little to do with everyday learning but may underlie reasoning processes of scientists in domains for which precise, very specific theories are already available. In contrast, learning tasks, especially in everyday contexts, often present domains for which no stable domain theories are available. People rarely have thick knowledge about the causal structures surrounding them (Rozenblit & Keil, 2002). I explored examples of such domains with sparse, fragmentary prior knowledge in our studies. In such tasks, people typically needed to focus on statistical data, which suggested the presence or absence of causal relations. Nevertheless, purely bottom-up learning would require amounts of data and information-processing capacity that are not available (Tenenbaum & Griffiths, 2003). Therefore, in such tasks, learners tend to use multiple fallible cues to simplify the induction process. Such cues include abstract knowledge of domain characteristics, learning data, difficulty of operations, and transfer from previous experiences, which all collaborate or compete to aid the induction process.

Note

1. In this article, I focus on generative causes (as opposed to preventive ones) and use the term *additivity* whenever people attribute more strength to a compound of causes than to each cause individually. This characterization acknowledges that the contributions of the individual causes may be weighted.

Acknowledgment

The experiments were conducted while Michael Waldmann was affiliated with the University of Frankfurt, Germany.

References

Ahl, V. A., Moore, C. F., & Dixon, J. A. (1992). Development of intuitive and numerical proportional reasoning. *Cognitive Development, 7,* 82–108.

Ahn, W.-K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54,* 299–352.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Anderson, N. H. (1987). Function knowledge: Comment on Reed and Evans. *Journal of Experimental Psychology: General, 116,* 297–299.

Anderson, N. H., & Butzin, C. A. (1974). Performance = motivation × ability: An integration-theoretical analysis. *Journal of Personality and Social Psychology, 30,* 598–604.

Anderson, N. H., & Wilkening, F. (1991). Adaptive thinking in intuitive physics. In N. H. Anderson (Ed.), *Contributions to information integration theory* (Vol. 3, pp. 1–42). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Birnbaum, M. H. (1976). Intuitive numerical prediction. *American Journal of Psychology, 89,* 417–429.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with the opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792–804.

Brunswik, E. (1956). *Perception and the representative design of experiments.* Berkeley: University of California Press.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 281–294.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory* (Vol. 1, pp. 187–215). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Cartwright, N. (1999). *The dappled world. A study of the boundaries of science.* Cambridge, England: Cambridge University Press.

Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science, 71,* 805–819.

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

Cheng, P. W. (1993). Separating causal laws from casual facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 30, pp. 215–264). San Diego, CA: Academic Press.

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

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382.

Couvillon, P. A., & Bitterman, M. E. (1982). Compound conditioning in honeybees. *Journal of Comparative and Physiological Psychology, 96,* 192–199.

De Graaf, C., & Frijters, J. E. R. (1988). Assessment of the taste interaction between two qualitatively similar-tasting substances: A comparison between comparison rules. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 526–538.

Dehaene, S. (1997). *The number sense.* New York: Oxford University Press.

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology, 14,* 218–224.

Dowe, P. (2000). *Physical causation.* Cambridge, England: Cambridge University Press.

Downing, C. J., Sternberg, R. J., & Ross, B. H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General, 114,* 239–263.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–32.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts and theories.* Cambridge, MA: MIT Press.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30,* 1128–1137.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions, SMC-17,* 753–770.

Kehoe, E. J. (1986). Summation and configuration in conditioning the rabbit's nictating membrane response to compound stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 12,* 186–195.

Kehoe, E. J., & Graham, P. (1988). Summation and configuration: Stimulus compounding and negative patterning in the rabbit. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 320–333.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 811–836.

Lagnado, D. A., Waldmann, M. A., Hagmayer, Y., & Sloman, S. A. (in press). Beyond covariation. Cues to causal structure. In A. Gopnik, & L. E. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation.* New York: Oxford University Press.

Moore, C. F., Dixon, J. A., & Haines, B. A. (1991). Components of understanding in proportional reasoning: A fuzzy set representation of developmental progression. *Child Development, 62,* 441–459.

Reed, S. K., & Evans, A. C. (1987). Learning functional relations: A theoretical and instructional analysis. *Journal of Experimental Psychology: General, 116,* 106–118.

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

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26,* 521–562.

Salmon, W. C. (1980). Probabilistic causality. *Pacific Philosophical Quarterly, 61,* 50–74.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Schlottmann, A., & Anderson, N. H. (1993). An information integration approach to phenomenal causality. *Memory & Cognition, 21,* 785–801.

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.

Sobel, D., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303–333.

Stavy, R., Strauss, S., Orpaz, N., & Carmi, C. (1982). U-shaped behavioral growth in ratio comparisons, or that's funny I would not have thought you were U-ish. In S. Strauss & R. Stavy (Eds.), *U-shaped behavioral growth* (pp. 11–36). New York: Academic.

Strauss, S., & Stavy, R. (1982). U-shaped behavioral growth: Implications for theories of development. In W. W. Hartup (Ed.), *Review of child development research* (Vol. 6, pp. 547–599). Chicago: University of Chicago Press.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North-Holland.

Surber, C. F. (1987). Formal representation of qualitative and quantitative reversible operations. In J. Bisanz, C. J. Brainerd, & R. Kail (Eds.), *Formal methods in developmental psychology* (pp. 115–154). New York: Springer-Verlag.

Surber, C. F., & Gzesh, S. M. (1984). Reversible operations in the balance scale task. *Journal of Experimental Child Psychology, 38,* 254–274.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems, 15,* 35–42.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (in press). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. New York: Oxford University Press.

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.

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

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychological Bulletin & Review, 8,* 600–608.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82,* 27–58.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*, 307–311.

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., & 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: Lawrence Erlbaum Associates, Inc.

Weiss, S. J. (1972). Stimulus compounding in free-operant and classical conditioning: A review and analysis. *Psychological Bulletin, 78,* 189–208.

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

Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 267–297). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.