

# Research Article

## TWO KINDS OF REASONING

Lance J. Rips

Northwestern University

**Abstract**—According to one view of reasoning, people can evaluate arguments in at least two qualitatively different ways: in terms of their deductive correctness and in terms of their inductive strength. According to a second view, assessments of both correctness and strength are a function of an argument's position on a single psychological continuum (e.g., subjective conditional probability). A deductively correct argument is one with the maximum value on this continuum; a strong argument is one with a high value. The present experiment tested these theories by asking participants to evaluate the same set of arguments for correctness and strength. The results produced an interaction between type of argument and instructions: In some conditions, participants judged one argument deductively correct more often than a second, but judged the second argument inductively strong more often than the first. This finding supports the view that people have distinct ways to evaluate arguments.

In thinking about arguments we read or hear, we sense a continuous range from arguments that are clearly worthless to ones we could never doubt. An argument can start from given information or premises that lend no support at all to its conclusion; another argument may have premises that, if true, force the truth of the conclusion; still other arguments vary freely between these extremes. At the worthless end are arguments like A, in which the truth of the premise (the top statement) is irrelevant to the truth of the conclusion (bottom statement):

A. Grizzlies hibernate during January.

Seven people fit in the back seat of a Ford.

Although the conclusion of A may be true, the premise provides no support for it, and it is hard to imagine anyone taking this argument seriously. Argument D, however, occupies the opposite end of this spectrum, because if the premise of D is true, the conclusion is obviously true as well:

D. Grizzlies hibernate during January, and black bears hibernate during January.

Grizzlies hibernate during January.

We can picture this variation as in Figure 1a, which represents individual arguments as points displayed on a single psychological dimension of *argument strength*. Worthless arguments, such as A, occupy the low end, arguments like D occupy the high end, and intermediate arguments, B and C, are in the middle ground. This picture encourages the thought that evaluating the worth of an argument is a

matter of locating the argument on this mental continuum. We could also try to make this picture more precise by specifying the nature of the dimension itself. For example, we might take as our measure of argument strength the conditional probability of the conclusion given the premises:  $P(\text{conclusion} | \text{premises})$ .<sup>1</sup>

Argument D is *deductively correct*: Its conclusion is true whenever its premise is true. Deductively correct arguments, however, are not the only sort of arguments that seem reasonable. For example, people might place some trust in argument C, even though it is not deductively correct:

C. Grizzlies hibernate during January.

Black bears hibernate during January.

The premise of C supports its conclusion to some degree, even though it is logically possible for its premise to be true and its conclusion false, and we can say that such arguments are *inductively strong*. On the mental scale of argument strength in Figure 1a, argument C might lie at the high end, though not at the maximum. Thus, the scale can accommodate the intuitions that argument C is stronger than argument A and that argument D is stronger than argument C.

If we take matters one step further, this *unitary view* might encompass all forms of argument evaluation. In some experiments, participants have to discriminate deductively correct arguments from those that are not correct; in others, they have to discriminate inductively strong arguments from those that are not strong. We can assume that decisions about a specific argument might differ in such experiments, with arguments like C being considered inductively strong but not deductively correct. Even so, we may be able to account for both types of assessment in terms of the criteria that participants set on the continuum shown in Figure 1a. In judging deductive correctness, participants may set a high criterion, as in Figure 1b, calling only the very strongest arguments correct. In judging inductive strength, however, they may set a less stringent criterion, as in Figure 1c, so more arguments will pass muster. If this unitary view is correct, the psychology of argument evaluation is a simple one-dimensional matter, akin to judgments of loudness or brightness.

1. Argument D has a conditional probability of 1 and so has maximum strength, in line with this view. The conditional probability of argument A, however, is not 0, but is instead equal to the probability of the conclusion. Thus, we would need to adjust the conditional probability if we wanted all worthless arguments to occupy the same minimum value on the scale of Figure 1a. As one such measure, we might consider

$$\text{argument strength} = \frac{P(\text{conclusion} | \text{premise}) - P(\text{conclusion})}{1 - P(\text{conclusion})},$$

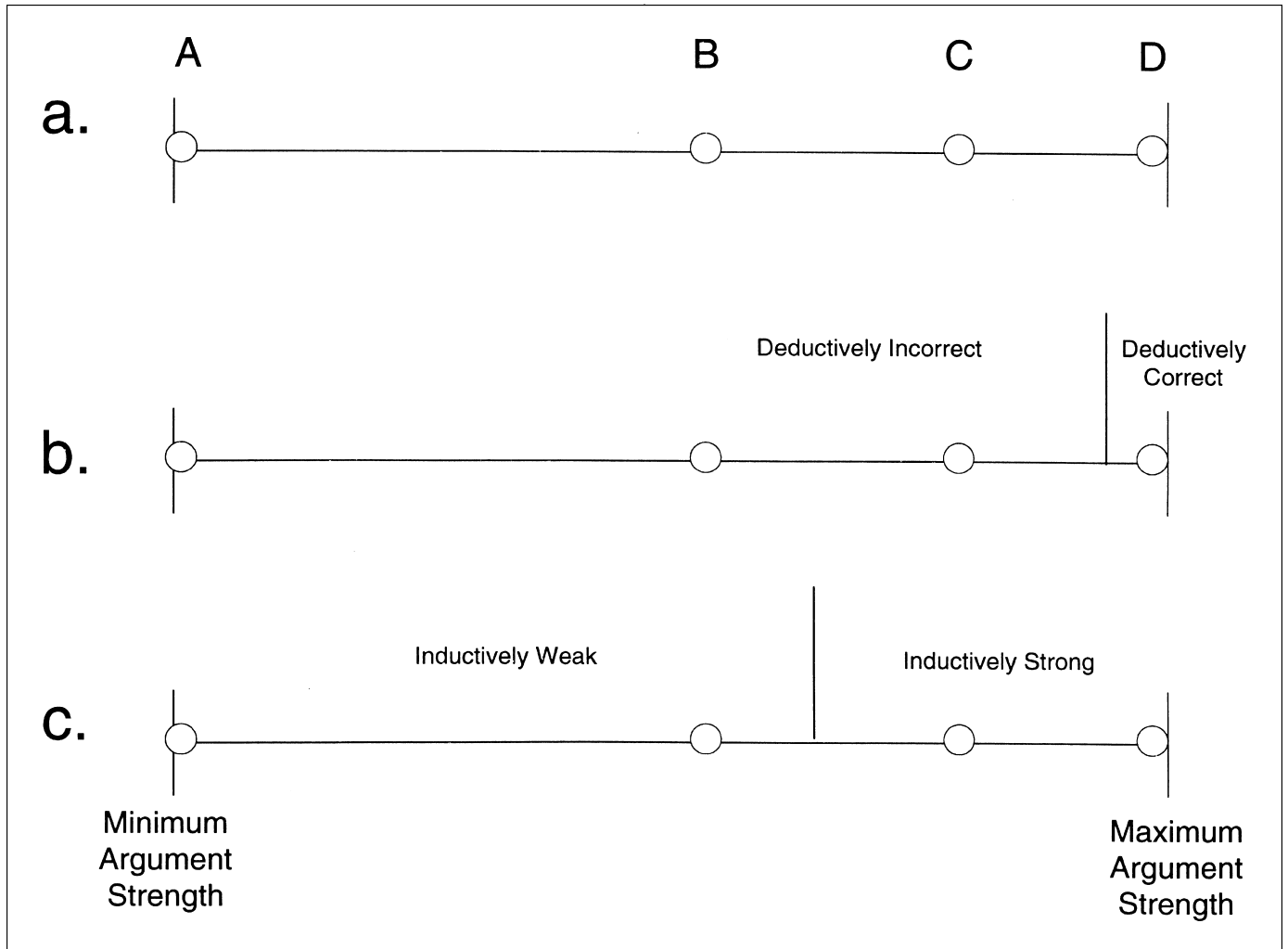
if  $P(\text{conclusion} | \text{premise}) > P(\text{conclusion})$

= 0, otherwise.

This expression is identical to the index  $K_i$  for conditional agreement (see Bishop, Feinberg, & Holland, 1975, p. 397). For purposes of this report, however, it is not crucial exactly how the dimension of Figure 1 is defined.

Address correspondence to Lance Rips, Psychology Department, Northwestern University, 209 Sheridan Rd., Evanston, IL 60208; e-mail: rips@northwestern.edu.

Two Kinds of Reasoning



**Fig. 1.** A one-dimensional view of argument evaluation. The continuum in (a) represents argument strength, and Points A, B, C, and D indicate individual arguments of differing strength. The same continuum can distinguish deductively correct from incorrect arguments when a hypothetical high criterion is used (b) and can distinguish inductively strong from weak arguments when a lax criterion is used (c).

There are reasons to think, however, that the unitary view is too simple. For one thing, if the Figure 1 continuum is a measure of conditional probability, then difficulties arise from examples in which  $P(\text{conclusion} \mid \text{premise}) = 1$  but the argument is not deductively correct. The following argument is one such example:

Calvin randomly chooses a real number between 3 and 4.  
 \_\_\_\_\_  
 Calvin does not choose  $\pi$ .

Because there is an infinite number of real numbers between 3 and 4, the likelihood of choosing  $\pi$  is 0; hence, the likelihood of not choosing  $\pi$  is 1. However, the conclusion of this argument does not follow deductively from its premise.<sup>2</sup> Although all deductively correct argu-

ments in classical logic have conditional probabilities of 1, the converse is not true. If both deductively correct arguments (e.g., D) and some arguments that are not deductively correct (e.g., the argument about  $\pi$ ) occupy the maximum value on the psychological dimension in Figure 1, then we can no longer count on this dimension to explain how people discriminate the deductively correct items. We must either deny that people have this ability (i.e., say that they lump these arguments together) or acknowledge that people have other ways of evaluating goodness of arguments.

The idea that people have more than one method for evaluating arguments receives support from two recent neuroimaging studies (Goel, Gold, Kapur, & Houle, 1997; Osherson et al., 1998). In these studies, participants viewed a series of arguments, deciding in one condition whether the premises entailed the conclusions and in a second condition whether the premises made the conclusions more plausible or probable. Although the findings of these studies differ in particulars, especially with respect to the brain areas that support the deduction judgments, both found differences in areas activated by the

2. Gilbert Harman (personal communication, April 29, 1998) suggested this example. A similar one appears in Priest (1999). See also Carnap's (1950) discussion of "almost L-true sentences."

**Table 1.** *Sample stimulus arguments*

Deductively correct, causally consistent:	
a.	If car X10 runs into a brick wall, it will stop. Car X10 runs into a brick wall.
	Car X10 will stop.
Deductively correct, causally inconsistent:	
b.	If car X10 runs into a brick wall, it will speed up. Car X10 runs into a brick wall.
	Car X10 will speed up.
Deductively incorrect, causally consistent:	
c.	Car X10 runs into a brick wall.
	Car X10 will stop.
Deductively incorrect, causally inconsistent:	
d.	Car X10 runs into a brick wall.
	Car X10 will speed up.

induction and the deduction tasks. Are these dissociations explainable within the unitary Figure 1 framework? We have seen that this view can explain differences between deduction and induction judgments through changes in criterion setting, as in Figure 1b versus Figure 1c. Although we might not expect criterion shifts to cause the sorts of disparities in activation that Goel et al. (1997) and Osherson et al. (1998) observed, it is hard to be certain without detailed knowledge of the role of the activated brain regions. It might be possible to shed light on these findings by systematically varying the properties of arguments that participants judge for deductive correctness and inductive support.

The unitary view of reasoning implies that evaluating an argument is always a matter of assessing its argument strength. If people decide that one argument is deductively correct and a second is not, then that must be because the first is stronger than the second. For example, if people decide that argument D in Figure 1b is correct and argument C is incorrect, then argument D must lie to the right of C on the strength continuum. This fact provides a potential test of the unitary view. Consider, for example, the four arguments in Table 1. These items vary in terms of their deductive correctness, with arguments a and b being correct and arguments c and d being incorrect. The arguments also vary in how plausible they are, in the sense of preserving expectations about physical cause and effect. When objects hit brick walls, they are obviously more likely to stop than speed up, and both arguments a and c (but neither b nor d) accord with this relation. For this reason, Table 1 labels arguments a and c *causally consistent*, and arguments b and d *causally inconsistent*. The key prediction concerns argument b (deductively correct and causally inconsistent) and argument c (deductively incorrect but causally consistent). If participants decide that argument b is deductively correct and argument c is not, then b must have more strength than c, according to unitary theories. Such theories therefore predict that participants should also find argument b inductively better than argument c. If participants' judgments do not follow this prediction, this would provide evidence against the unitary view and in support of the idea that more than one type of argument evaluation is in play.

Deductive correctness and inductive strength supply different theoretical criteria for arguments: As noted, an argument is deductively correct if its conclusion is true whenever its premises are true, and an argument is inductively strong if its premises support its conclusion (but is not necessarily deductively correct). It remains to be seen whether a single psychological dimension underlies people's judgment of both correctness and inductive strength, as the unitary view holds, or whether different dimensions are required. If experimental results call for more than one psychological dimension, then we need to consider what the nature of the additional factor or factors might be, but for the time being, we can remain neutral, returning to this issue after we have examined the evidence.

## METHOD

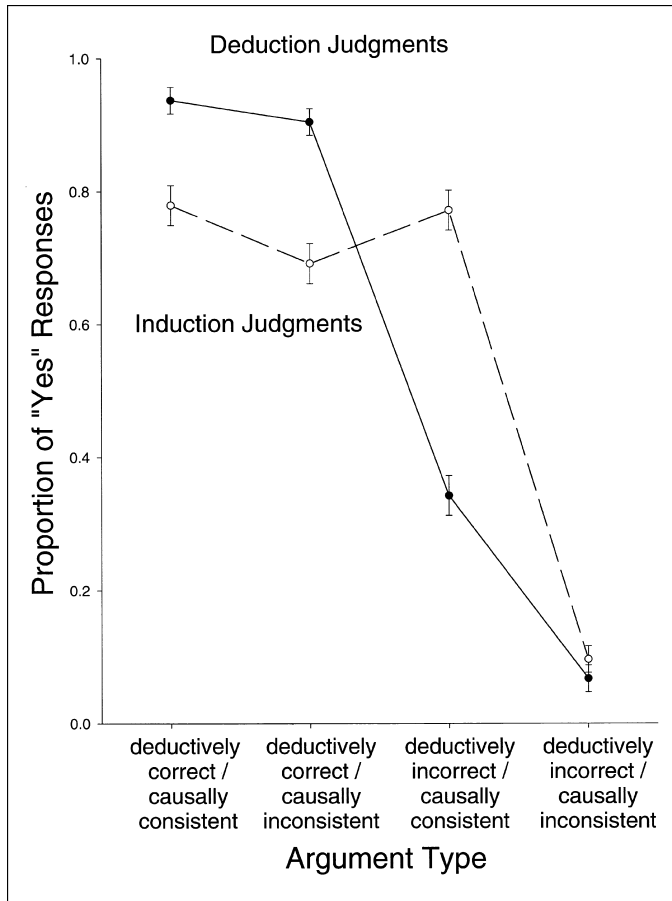
Each participant in this experiment received a booklet containing a page of instructions, followed by 16 arguments (1 per page). The experimenter read the instructions to the participant, while the participant followed along. The participant then continued at his or her own pace to evaluate the arguments, either for deductive correctness or for inductive strength (depending on the condition).

In the *deduction conditions*, participants judged whether each argument was "valid" or "not valid" (by circling one of these responses) and then rated their confidence in their decision. The deduction conditions included three subconditions that differed in the wording of the instructions. The wording was varied to check whether the results would depend on exactly how the instructions explained deductive correctness. The first asked participants, "Assuming the sentences above the line are true, **does this necessarily** make the sentence below the line true?"; the second asked, "Assuming the sentences above the line are true, **can you be certain** that the sentence below the line is true?"; and the third asked, "**Considering just the form of the sentences** (and not their specific content), does this form ensure that if the sentences above the line are true so is the sentence below the line?" (emphasis in the original).

In the *induction conditions*, participants judged whether the argument was "strong" or "not strong," and then rated degree of strength. There were also three induction subconditions. In one of these, participants were asked, "Assuming that the sentences above the line are true, **how plausible** does this make the bottom sentence?"; in the second, participants were asked, "Assuming that the sentences above the line are true, **how convincing** does this make the bottom sentence?"; and in the third, they were asked, "Assuming that the sentences above the line are true, **would this causally produce** the situation described by the bottom sentence?" (emphasis in the original).

In the analysis described later, these variations in wording within the two critical conditions produced no main effect and no interactions with other factors. For this reason, in the discussion of the results, decisions in the first three subconditions are grouped as deduction judgments, and decisions in the second three subconditions are grouped as induction judgments. The ratings that participants gave in the deduction condition (degree of confidence) and the induction condition (degree of strength) differ in a way that makes direct comparison difficult. For this reason, the analyses that follow are confined to the probabilities of responding "valid" or "not valid" or of responding "strong" or "not strong." Relying on these judgments also has the advantage that the results are immune to variations in the way the different groups of participants used the rating scale.

Two Kinds of Reasoning



**Fig. 2.** Proportion of trials on which participants judged arguments deductively correct (solid line) or inductively strong (dashed line). Each point represents 240 observations. Error bars show 1 standard deviation of the proportion:

$$\sqrt{\frac{p(1-p)}{240}}$$

The arguments came from quartets of items, such as the one in Table 1, that varied in logical form and content. Within each quartet, two items were deductively correct and two deductively incorrect. The correct items in a quartet were based on one of four simple inference forms: *modus ponens* (If  $p$  then  $q$ ,  $p \vdash q$ ), conjunction elimination ( $p$  and  $q \vdash p$ ), conjunctive syllogism (not( $p$  and  $q$ ),  $p \vdash$  not  $q$ ), and disjunctive syllogism ( $p$  or  $q$ , not  $p \vdash q$ ) (see Rips, 1994, chap. 4). Incorrect items altered this form so that they were no longer deductively correct. Table 1 provides an example of a quartet based on *modus ponens*.

Each logical form was instantiated by four quartets, each with a different type of content. For example, the arguments in Table 1 instantiate the *modus ponens* quartet with a content based on a car crash; the other three instantiations involved painting a room, locking a door, and igniting wood shavings. In each case, 1 deductively correct and 1 incorrect argument preserved typical beliefs about causal relations, as in arguments a and c of Table 1. The other 2 arguments violated these beliefs, as in arguments b and d. We assigned each participant 16 arguments, 4 based on each logical form. Within a form, 1 argument was

deductively correct and causally consistent, 1 was deductively correct and causally inconsistent, 1 was deductively incorrect and causally consistent, and 1 was deductively incorrect and causally inconsistent, as in Table 1. For an individual participant, however, different content instantiated each of these arguments, according to a random Latin square. Thus, for example, no participant received more than 1 argument from the Table 1 quartet. The order of the problems in the test booklets was random.

The participants were 120 Northwestern University undergraduates, 20 in each subcondition. They took part in the study either individually or in small groups of up to 6 individuals. Participants completed the task in 15 to 35 min and received credit toward a course requirement for their time. None had taken a college-level logic course.

**RESULTS AND DISCUSSION**

The unitary view of argument evaluation predicts that judgments of deductive correctness should mimic judgments of inductive strength. Any increase in the proportion of arguments judged to be deductively correct from one type of argument to another should accompany an increase in the proportion judged to be inductively strong. This is because, according to the unitary view, the two judgments reflect a difference in the positions of the arguments on the same underlying scale. The results of this experiment, however, suggest that this view is incorrect. Figure 2 presents the main findings: the proportion of trials on which participants identified an argument as deductively correct (solid line) or inductively strong (dashed line). The  $x$ -axis displays the four types of arguments, corresponding to the examples in Table 1. The proportion of arguments judged to be deductively correct decreased between the deductively-correct-and-causally-inconsistent items and the deductively-incorrect-and-causally-consistent ones. The proportion of arguments judged to be inductively strong increased between these same arguments. So that the deduction and induction conditions could be compared in the analysis, “valid” and “strong” answers were grouped as “positive” responses, and “invalid” and “not strong” answers were grouped as “negative” responses.

Arguments that were both deductively correct and causally consistent produced positive responses from participants in both the deduction and the induction conditions. Likewise, both groups gave mainly negative responses to arguments that were deductively incorrect and causally inconsistent. However, results for the two remaining argument types show a crossover in Figure 2 and contributed to a significant three-way interaction among the factors of deductive correctness, causal consistency, and condition (deduction vs. induction judgments),  $F(1, 114) = 31.66$ ,  $MSE = 0.113$ ,  $p < .0001$ . A planned comparison confirmed that there were significantly more positive responses in the deduction condition for deductively-correct-and-causally-inconsistent problems than for deductively-incorrect-and-causally-consistent ones; however, there was also a significant difference in the opposite direction for the induction condition; for both tests,  $F(1, 114) > 6.80$ ,  $p < .05$ .

Figure 2 also shows that participants in the deduction condition were reasonably successful at discriminating deductively correct from incorrect arguments. These participants responded positively to 92.1% of the correct arguments but to only 20.4% of the incorrect ones. The argument forms that the experiment employed are among the simplest ones and unlikely to call for extensive mental calculations (Braine, Reiser, & Rumin, 1984), so the participants’ success is not surpris-

ing. However, causal consistency of the premises and conclusion also affected responses in this condition. Participants judged consistent arguments deductively correct on 63.9% of trials and judged inconsistent arguments correct on 48.5% of trials. The obtained interaction between correctness and consistency within the deduction condition echoes earlier findings on content effects (e.g., Evans, Barston, & Pollard, 1983): The size of the consistency effect was larger for deductively incorrect arguments than for deductively correct ones,  $F(1, 57) = 41.01$ ,  $MSE = 0.086$ ,  $p < .0001$ . The results also agree with earlier findings that causal relations sometimes influence deduction judgments (e.g., Cummins, Lubart, Alksnis, & Rist, 1991).

Participants in the induction condition gave mostly positive responses to arguments that were either causally consistent or deductively correct. Only deductively-incorrect-and-causally-inconsistent arguments yielded many “not strong” judgments. The effects of consistency were slightly bigger than the effects of deductive correctness in this condition. Participants responded positively to 77.5% of the consistent arguments and to 39.4% of the inconsistent ones. Similarly, they responded positively to 73.5% of deductively correct items and to 43.3% of deductively incorrect ones.

If participants were making their induction judgments on the basis of conditional probability (or some similar measure), we would expect the pattern of data to be similar to that of Figure 2. Probability theory guarantees that  $P(\text{conclusion} | \text{premises})$  will be high for deductively correct arguments, as noted earlier, and causally consistent items will also have high conditional probability. One open question is why induction responses did not reach the ceiling levels that deduction judgments achieved. Perhaps participants were relying on a method other than calculating  $P(\text{conclusion} | \text{premises})$ , or perhaps they were reducing the number of positive answers in order to achieve about equal numbers of positive and negative responses overall. In the latter case, the three left-hand points on the induction curve in Figure 2 underestimate the true proportions. Correcting for such an underestimate, however, would not eliminate a difference in the shape of the functions in Figure 2.

The results of this study tend to disconfirm the unitary view of Figure 1. If participants base all forms of argument evaluation on the position of the argument on a single psychological dimension, then induction and deduction judgments should increase or decrease together. Figure 2 shows, however, that the two judgment types are not monotonically related. Dissociations are not foolproof evidence against unitary processes, but it is worth noting in the present case that the percentage of positive responses for deduction judgments was not a monotonic function of those for induction judgments. This is a consequence of the crossover in Figure 2 and can be seen by regraphing the data of the figure as a scatter plot with the deduction percentages on the y-axis and the induction percentages on the x-axis for each of the four conditions. Monotonicity of the judgments is a necessary condition for a single process to account for both tasks (Dunn & Kirsner, 1988). Hence, the violation of this condition provides evidence against unitary views.

In addition, the dissociation puts some strong restrictions on possible theories (Shoben & Ross, 1986). The results suggest that people are not using, for example, probability as the sole basis for both judgments (as in the theory proposed by Oaksford & Chater, 1998). However, the results also run counter to other possible unitary views. Induction and deduction cannot both be based on the proportion of situations, possible worlds, or mental models of the premises in which the conclusion is true (Johnson-Laird, 1994; see also Johnson-Laird,

Legrenzi, Girotto, Legrenzi, & Caverni, 1999). The present findings agree with previous brain-imaging evidence for distinct cognitive mechanisms in deductive and inductive reasoning (Goel et al., 1997; Osherson et al., 1998), and they help eliminate possible counterexplanations of those results.<sup>3</sup>

The present findings do not mean, however, that there are no connections between induction and deduction. People may sometimes use inductive judgments when they are officially supposed to be deciding on deductive correctness, especially if the problem is difficult. The fact that causal consistency affected deduction in this experiment is a hint that there are procedures common to the two types of evaluation. It is also possible to devise theories in which substantially the same method determines assessments of both deductive correctness and inductive strength, but makes use of two separate sources of information in doing so. For example, one could envision a theory in which people based induction judgments on the proportion of all causally possible situations consistent with the premises in which the conclusion is true and, similarly, based deduction judgments on whether the conclusion is true in all logically possible situations (regardless of causal consistency) in which the premises are true. Such a theory, however, would have to specify how people determine the difference between causally possible situations and logically possible ones. The results are consistent with overlap—but not complete overlap—in people’s reasoning techniques.

Why would people have more than one way to evaluate arguments? Like many other complex objects, arguments have different roles and purposes, and people assess them differently depending on which purpose they have in mind. In some settings, arguments are useful in providing information about what follows from given hypotheses in view of one’s knowledge of how the world works. In other settings, one can take a more abstract approach and ask what follows from given information on the basis of important pieces of logical or mathematical structure, generalizing over specific content. It is unlikely that one of these purposes subsumes the other. Emphasizing one goal of inference at the expense of others leads back to the unitary view and a hobbled theory of reasoning.

**Acknowledgments**—I thank Adelia Falk for assistance with the experiment reported here, and Gilbert Harman, Reid Hastie, Philip Johnson-Laird, Ariela Lazar, Elizabeth Lynch, Douglas Medin, and Daniel Osherson for comments on earlier presentations of these ideas. National Science Foundation Grant SBR-9514491 supported this research.

## REFERENCES

- Bishop, Y.M.M., Feinberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.
- Braine, M.D.S., Reiser, B.J., & Rumin, B. (1984). Some empirical justification for a theory of natural propositional reasoning. *Psychology of Learning and Motivation*, 18, 313–371.

3. A number of investigators (Evans & Over, 1996; Sloman, 1996; Stanovich, 1999) have proposed a distinction between two reasoning processes—one rule-based (analytic, explicit) and the other similarity-based (heuristic, implicit). These authors clearly intend their distinction to apply both to deductive reasoning and to inductive or probabilistic reasoning; if so, it is orthogonal to the difference discussed here. There is also no evidence that participants in the deduction condition were employing mainly explicit strategies, whereas those in the induction condition were employing implicit ones.

## Two Kinds of Reasoning

- Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.
- Cummins, D.D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19*, 274–282.
- Dunn, J.C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review, 95*, 91–101.
- Evans, J.S.B.T., Barston, J.L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295–306.
- Evans, J.S.B.T., & Over, D.E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.
- Goel, V., Gold, B., Kapur, S., & Houle, S. (1997). The seats of reason? An imaging study of deductive and inductive reasoning. *NeuroReport, 8*, 1305–1310.
- Johnson-Laird, P.N. (1994). Mental models and probabilistic thinking. *Cognition, 50*, 189–209.
- Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M.S., & Caverni, J.-P. (1999). Naive probability. *Psychological Review, 106*, 62–88.
- Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, England: Psychology Press.
- Osherson, D., Perani, D., Cappa, S., Schnur, T., Grassi, F., & Fazio, F. (1998). Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia, 36*, 369–376.
- Priest, G. (1999). Validity. In A.C. Varzi (Ed.), *The nature of logic* (pp. 183–206). Stanford, CA: CSLI Publications.
- Rips, L.J. (1994). *Psychology of proof*. Cambridge, MA: MIT Press.
- Shoben, E.J., & Ross, B.H. (1986). The crucial role of dissociations. *Behavioral and Brain Sciences, 9*, 568–571.
- Slooman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.
- Stanovich, K.E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

(RECEIVED 2/8/00; REVISION ACCEPTED 6/19/00)