Modeling Two Kinds of Reasoning

Evan Heit (eheit@ucmerced.edu)
Cognitive Science, University of California, Merced, CA, USA

Caren M. Rotello (caren@psych.umass.edu)
Department of Psychology, University of Massachusetts, Amherst, MA, USA

Abstract

This research addressed how inductive reasoning and deductive reasoning are related. We directly compared the ideas that the same cognitive mechanisms are applied to both types of judgments and that different processes lead induction and deduction to use different information. In an experiment, subjects either judged inductive strength or deductive validity, for a common set of arguments. A finding from Heit and Rotello (2005), greater sensitivity to validity in the deduction condition, was replicated. A new finding was greater sensitivity to number of premises for induction. The results were analyzed using signal detection theory (SDT) and receiver operating characteristic (ROC) curves, and modeled using both one-dimensional and multidimensional versions of SDT. The one-dimensional model could not capture differential sensitivity to number of premises. A successful model for the entire pattern of results assumed two underlying dimensions, deductive correctness and associative strength, with different proportions comprising either induction or deduction judgments.

Keywords: inductive reasoning, deductive reasoning, memory, signal detection theory

Introduction

What is the relation between inductive reasoning and deductive reasoning? According to the *problem* view, induction and deduction refer to different types of reasoning problems, or different types of arguments. An aim of the problem view is to distinguish deductively valid arguments from invalid arguments. The problem view does not make claims about cognitive processes, only the arguments themselves. In contrast, according to the *process* view, the question of interest is what cognitive processes characterize induction and deduction, and whether inductive reasoning and deductive reasoning really differ in terms of processing (see Heit, 2007, for a review).

Explicitly or implicitly, researchers concerned with the process view have taken different positions on whether there are different kinds of reasoning. Some researchers have asserted that induction and deduction depend on the same cognitive mechanisms. This will be referred to as the *one-process view*. For example, Harman (1999) argued that people do not reason deductively, but use the same reasoning abilities for both deductive and inductive problems. Indeed, Oaksford and Chater (2007) have applied a non-deductive, probabilistic account of reasoning to human performance on many deductive problems, and likewise mental model theory (e.g., Johnson-Laird, 1994) is a unified account applying to performance on both deductive and inductive problems. Finally, models of inductive reasoning (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993) have been applied to some deductive problems, in which a probability judgment of 100% corresponds to a judgment of deductive validity.

Other researchers have emphasized that reasoning depends on two kinds of processing, or on two systems (e.g., Evans & Over, 1996; Sloman, 1996; Stanovich, 1999). We refer to this as the *two-process view*. In these accounts there is one system that is fast but heavily influenced by context and similarity, and another system that is more deliberative and accurate. Although these two systems do not necessarily correspond directly to induction and deduction, the idea is that induction would depend more on the first system whereas deduction would depend more on the second system.

Considered separately, the one-process view and the two-process view are each embodied by several highly productive and successful research programs. However, little research has directly pitted the one-process view and the two-process view against each other, thereby directly addressing whether there are two kinds of reasoning. One important exception is a study by Rips (2001), who compared two types of arguments in two experimental conditions, in which subjects were instructed to judge either deductive correctness or inductive strength. Rips noted that if induction and deduction use the same information along a common scale of argument strength, then the relative ordering of two arguments should be the same whether people are judging deductive correctness or inductive strength. One type of argument was deductively correct but causally inconsistent, such as "Jill rolls in the mud and Jill gets clean, therefore Jill rolls in the mud," and the other type was deductively incorrect but causally consistent, such as "Jill rolls in the mud, therefore Jill rolls in the mud and Jill gets dirty." Subjects in the deduction condition gave more positive judgments to the correct but inconsistent arguments, whereas participants in the induction condition gave more positive judgments to the incorrect but consistent arguments. Rips concluded that this result was evidence against the one-process account, which predicts the same ordering of arguments in both conditions, with only a potential change in response bias.

Heit and Rotello (2005) pointed out that distinguishing between one-process and two-process views is also an important enterprise in memory research. In particular, there is a lively debate on whether recognition memory can be accounted for by a single familiarity process, or if there

are two processes, a fast familiarity process influenced by similarity, and a slower, recollective process that is more accurate. This issue is often examined in the remember-know paradigm (Tulving, 1985), in which subjects make a recognition judgment then state whether they just know they have seen the item before or actually remember it. Although these two judgments may not correspond directly to familiarity and recollection, under the two-process view "know" judgments depend more on familiarity whereas "remember" judgments depend more on recollection. Under the one-process view, "remember" judgments reflect a stricter response criterion than "know."

Memory researchers have developed several standards for examining whether a set of results points to one or two processes (Dunn & Kirsner, 1988). One such standard is monotonicity, namely that across a common set of stimuli, the response rates for two types of memory judgments should be highly correlated. To the extent that monotonicity holds, the one-process view is supported, and to the extent that monotonicity is violated, there is evidence for the two-process view.

Essentially, Rips (2001) applied the monotonicity standard to reasoning. The different ordering of argument strengths under induction and deduction instructions was a violation of monotonicity and thus evidence for two processes. Heit and Rotello (2005) focused on another standard. In memory research, it has been argued that if "remember" judgments measure recollection these should show greater sensitivity than "old" responses that reflect a mixture of recollection and familiarity. That is, the difference between the hit rate and the false alarm rate, measured in d' units, should be greater for "remember" responses than for "old" judgments. In contrast, if remembers just reflect greater strength than knows, then d' for remembers should equal d' for old decisions. In two experiments, Heit and Rotello found that sensitivity was about twice as high for deduction judgments (d' = 1.69 on average) than for induction judgments (d' = 0.86). They took that difference as evidence for two processes of reasoning, with the more accurate, deliberative process contributing more to deduction.

Heit and Rotello (2005) also plotted receiver operating characteristic (ROC) curves for their data. ROC curves are frequently used in memory research, but to our knowledge had never been used in reasoning research. The ROC curves plotted the probability of a positive response to valid arguments (called hits) on the y-axis and to invalid arguments (called false alarms) on the x-axis; the points indicated varying response biases, obtained from confidence ratings (see Macmillan & Creelman, 2005). ROC curves go well beyond *d'* measures; they are useful for checking that the assumptions of signal detection theory (SDT) are met and fostering further inferences about the underlying processes. ROCs that fall higher in space reflect greater sensitivity because the hit rate is greater for a given false alarm rate (see Figure 1 for illustration). In addition, points that fall to the upper-right along a given ROC reflect more liberal response biases, because both the hit and false alarm rates are higher. In Heit and Rotello (2005), the ROC curves did not fall at the same height for deduction and induction, supporting the conclusion that a two-process account was needed.

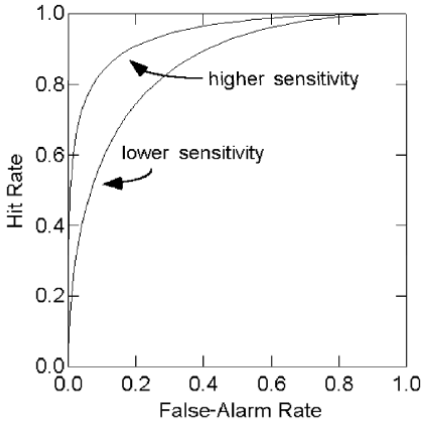

Figure 1 Example ROCs

Although Heit and Rotello (2005) made some progress in amassing evidence for two processes in reasoning, there were some limitations to their work. First, although they showed that argument validity affects deduction more than induction, they did not manipulate any variable that targeted induction. Second, the inferences based on d' (and ROCs) themselves have some limitations. The d' differences were consistent with two-process accounts, but it might be possible to fashion a one-process account to explain the results, for example one in which response criteria are variable, and differentially so for induction and deduction (see Wixted & Stretch, 2004). Third, and most important, Heit and Rotello did not actually implement one- or two-process accounts of reasoning.

Hence, the present work had three aims. First, we varied the number of premises in an argument. Although increasing the number of premises does not in itself make an argument valid, providing more evidence can make a conclusion seem more plausible (cf., Osherson et al., 1990). In particular, we predicted that for invalid arguments, increasing the number of premises would affect induction more than deduction, which itself would be more sensitive to actual validity. Second, the analyses included a direct evaluation of the possibility that there is just one process, but with variable response criteria. A way to think about this is that if either induction or deduction judgments are poorly understood by the subjects, this could be reflected in trial-to-trial or person-to-person variability in how the rating scales are applied. Third, predictions for the experiment were generated from actual implementations of one-dimensional and two-dimensional SDT models. Because fitting these models to data could require results from many experiments, here the models are presented as illustrations and existence proofs rather than as optimally fitted accounts.

Method

Sixty University of California, Merced students were paid to participate. Subjects were randomly assigned to one of two conditions: induction (n=31) or deduction (n=29).

The initial instructions for the induction condition gave a definition of a strong argument, "assuming the information above the line is true, this makes the sentence below the line **plausible.**" Likewise, the deduction instructions gave a brief definition of a valid argument: "assuming the information above the line is true, this **necessarily** makes the sentence below the line true."

There were 120 questions, comprising arguments about the following kinds of mammals: bears, cats, cows, dogs, goats, horses, lions, mice, rabbits, and sheep. An example invalid argument is:

```
Horses have Property X
Mice have Property X
Sheep have Property X
---------------------
Cows have Property X
```

Note that we literally used "Property X." Subjects were instructed to treat this as a novel biological property. One-third of the arguments had a single premise, that is a single category above the line. One-third had three premises (as in the previous example) and one-third had five premises. Half the arguments were not deductively valid. The remaining arguments were deductively valid: Either one of the premises was "Mammals have Property X," implying that the category in the conclusion also had Property X, or the conclusion category was identical to one of the premise categories. An example valid argument is:

```
Horses have Property X
Mice have Property X
Sheep have Property X
Mammals have Property X
Cats have Property X
---------------------
Rabbits have Property X
```

The experiment was run using a program on a computer. After the instructions for either the induction or deduction condition were displayed, the 120 arguments were presented individually, in a different random order for each subject. In the induction condition, subjects pressed one of two keys to indicate "strong" or "not strong." In the deduction condition, subjects pressed one of two keys to indicate "valid" or "not valid." This judgment was followed by a confidence rating on a 1-5 scale, with higher numbers indicating greater confidence.

Empirical Results

Overall, there were more positive responses ("strong" in the induction condition, "valid" in the deduction condition) to valid arguments than invalid arguments, as shown in Table 1. Although the hit rates (P(positive response | valid)) are similar in the two conditions (F<1), the false alarm rates (P(positive response | invalid)) are much lower in the deduction condition ($F(1,58)=10.82$, $p<.01$, MSe=1.38). Thus, subjects' ability to discriminate valid from invalid arguments was greater for deduction than induction (d' = 3.52 v. 2.37, averaging over number of premises). This result replicates Heit and Rotello (2005).

Table 1 Probability of positive response.

	Valid			Invalid		
Number of premises	1	3	5	1	3	5
Induction	0.98	0.92	0.93	0.18	0.23	0.30
Deduction	0.99	0.96	0.96	0.05	0.07	0.07

The novel aspect of this study is that we measured the effect of number of premises. Number of premises had the most obvious effect on the false alarm rates in the induction condition: the proportion of positive responses on invalid arguments increased with the number of premises. Note that because there were fewer positive responses overall to invalid arguments in the deduction condition, there was the most room for an increase in that condition. Condition significantly interacted with number of premises ($F(2,116)=4.42$, $p<.05$, MSe=.048): As expected, the number of premises affected false alarm rates in the induction condition ($F(2,60)=8.03$, $p<.01$, MSe=.128) but not in the deduction condition (F<1).

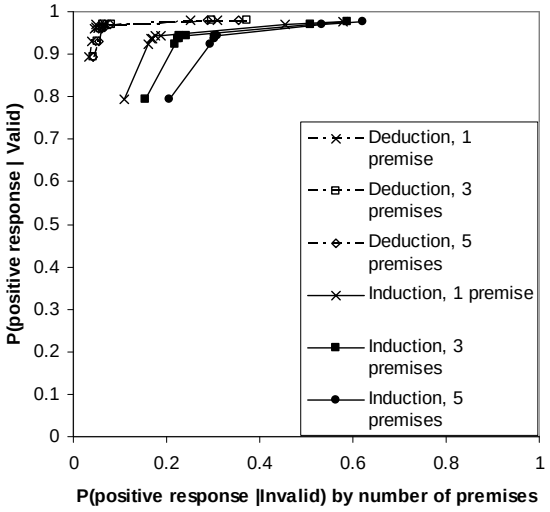

Figure 2 Observed ROCs

ROC curves were plotted to give a more complete look at performance. The greater sensitivity to validity for deduction compared to induction is apparent in the group ROCs (see Figure 2) and does not depend on use of the *d'* statistic itself: The deduction ROCs all fall higher in the space than the induction curves. In addition, the deduction ROCs do not vary with number of premises, but the

induction ROCs show a clear shift to the right (i.e., higher false alarm rates) as number of premises increase.

Modeling

Our modeling approach was illustrative. Although we did assess a large range of parameter combinations for each model, we did not attempt to find the best-fitting parameters for any given model. Instead, our goal was to evaluate the general form of the predicted ROCs, to assess whether a single representation could account for both deductive and inductive judgments to find out, we allowed only the decision bounds to vary between conditions, analogous to the approach of Rips (2001).

We began by asking assessing a one-dimensional model One possible representation would have a single Gaussian distribution with high average strength to represent the valid arguments, and 3 different invalid distributions (for the different numbers of premises) of lower mean strength; the presumed difference between conditions would be a more conservative location of the confidence criteria in deduction compared to induction (see Figure 3). Such a model cannot fit the data, because it predicts that only response bias changes between conditions, not subjects' ability to distinguish valid from invalid arguments: *d′* would be the same for deduction and induction; the ROCs for deduction and induction would fall at the same height in the space. Our data clearly show that deduction led to higher ROCs and higher *d′*.

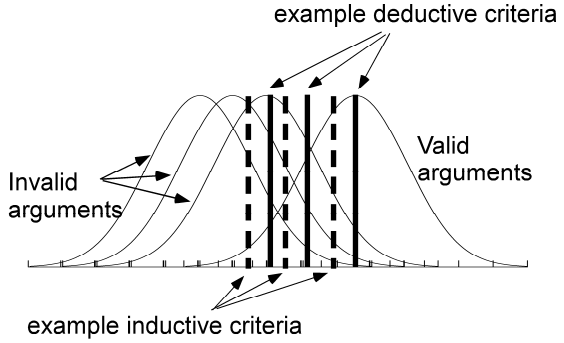

Figure 3 One-dimensional model

The one-dimensional model could be modified by assuming that, for some reason, the locations of subjects' confidence criteria were more variable across trials in the induction condition; criterion variability decreases the apparent accuracy in a condition and could therefore account for the lower sensitivity in the induction task. (Anecdotally, our subjects expressed less familiarity with validity judgments than with plausibility judgments, so just the opposite effect might be expected, namely more criterion variability in the deduction condition.)

We used Monte Carlo simulations to generate predicted ROCs for this model, sampling 10,000 values from each of the assumed strength distributions. The predicted ROCs were then generated by choosing locations for decision criteria, ranging from conservative positions that yield low hit and false alarm rates to liberal positions that yield higher hit and false alarm rates. For each confidence level (i.e., rating 5, say), we assumed that the deduction criterion was more conservative than the induction criterion, consistent with the argument made by Rips (2001). To allow for greater criterion variability in the induction condition, we sampled the locations of the inductive criteria from Gaussian distributions; the conservative criteria (ratings 5 and 4) were sampled from distributions that had more variability than the liberal criteria. More specifically, we assumed that the valid arguments were normally distributed with mean of 2.3 and standard deviation of 1. Invalid arguments varied in mean strength, depending on the number of premises: 1, 3, and 5 premise problems had means of 0, 0.1, and 0.2; the corresponding standard deviations were all 0.5. Finally, the induction criteria had standard deviations that ranged from .7 (for the most conservative criterion) to .1 (for the most liberal criterion).

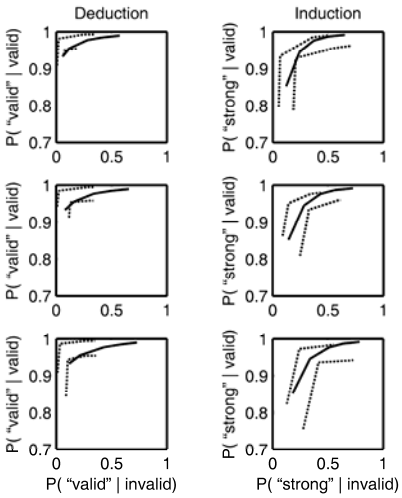

Figure 4. ROCs simulated with one-dimensional model assuming variable criteria for induction, by condition and number of premises (1=top row, 3=middle, 5=bottom row). Dashed functions are 95% CIs based on empirical data; Solid functions are model-generated predictions.

Figure 4 shows this model's predictions—the solid lines—plotted against the data: The dashed functions are the upper and lower bounds of the 95% CIs for each observed ROC (obtained by bootstrapping the individual subjects' data 2000 times and selecting the resulting group ROCs at the 2.5 and 97.5 percentiles as the confidence bounds). This model captures the finding that sensitivity is greater for deduction than induction (curves shifted upwards for deduction) as well as the approximate shape of the observed ROCs. This model still fails, though, because it predicts the same pattern of false alarm rates with number of premises for both conditions: either an increase with number of premises or no difference, but not the observed interaction with condition. For the parameter values used to generate the ROCs in Figure 4, the predicted false alarm rates are .06,

.08, and .12 for 1, 3, or 5 premises in the Deduction task and .12, .14, and .18 in the Induction task, showing small increases in both conditions (in contrast to Table 1).

Having ruled out this one-dimensional representation in which the same information is used for both deduction and induction, we considered the possibility that two different (orthogonal) dimensions of information were used in the judgments. These dimensions can be thought of as "apparent deductive correctness" and "consistency with associative knowledge," although other labels are possible. Our starting assumptions were that valid arguments would differ more from invalid arguments along the logic axis, and that the number of premises would influence the strength of evidence along the associative axis. Invalid arguments are generally (but not uniformly) low on apparent correctness, but vary in their consistency with associative knowledge; a greater number of premises would increase the effect of similarity between premise and conclusion categories. Both valid and invalid arguments were assumed to have bivariate Gaussian distributions with small but non-zero covariance, which means that an argument's strength on one dimension is slightly correlated with its strength on the other.

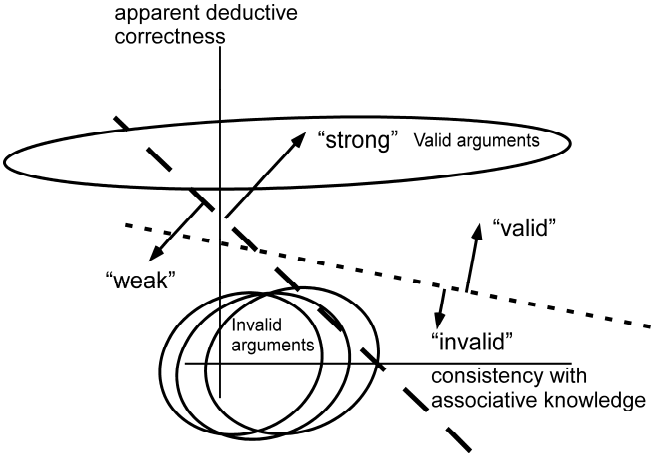

Figure 5. Two-dimensional model. Ellipses are equal-probability contours of bivariate Gaussian distributions of arguments.

To make either an inductive or deductive decision in this model, a criterion is required. We assumed that both types of judgments involve weighted combinations of evidence on the two dimensions, which yields decision criteria that do not parallel either strength axis. An example representation and associated decision bounds are shown in Figure 5. The long-dash line is the decision bound for induction judgments and the short-dash line is the decision bound for deduction judgments. The parameters used to simulate data are approximated in this figure: Valid items are drawn from a bivariate Gaussian distribution with a mean strength of 0.8 and standard deviation of 4 on the associative strength dimension, mean strength of 3 and standard deviation of 0.4 on the logic dimension, and a covariance of 0.3. We can describe this distribution as $Valid_{assoc} \sim N(0.8, 4)$, $Valid_{logic} \sim N(3, 0.4)$, $cov_{Valid}(assoc,logic)=0.3$. Invalid items have strengths that increase with the number of premises. Using analogous notation and subscripts of 1, 3, 5 to note number of premises, $Invalid_{assoc1} \sim N(0.2, 1)$, $Invalid_{logic1} \sim N(0, 0.9)$, $Invalid_{assoc3} \sim N(0.6, 1.1)$, $Invalid_{logic3} \sim N(0, 1.0)$, $Invalid_{assoc5} \sim N(1, 1.2)$, $Invalid_{logic5} \sim N(0, 1.1)$, and $cov_{Invalid}(assoc,logic)=0.4$. We assumed the inductive judgments would depend more on associative knowledge than would deductive judgments; this assumption is reflected in the steeper slope of the decision bound in the induction condition (-0.7, compared to -0.2 for deduction).

As for the one-dimensional model, we simulated the model using Monte Carlo sampling. The predicted ROCs were swept out by systematically varying the intercepts of the induction and deduction decision bounds to calculate the hit and false alarm rates at a range of confidence criteria. Our goal was to demonstrate that ROCs could be generated from this two-dimensional model that would fall within the 95% CIs of the observed ROCs for both induction and deduction, assuming that the only difference between the tasks was the slope of the decision axis. Figure 6 shows that we were successful: the dashed functions are the upper and lower bounds of the 95% CIs for each observed ROC, and the solid functions are the model-generated ROCs. In each case, the predicted ROCs fall within the confidence limits. We are encouraged by these simulations because the parameter values were not optimized for the observed data. The model's predictions capture the key trends in the data, namely greater overall sensitivity for deduction than induction, and a rightward shift of the ROC curve as number of premises increases in the induction condition.

Finally, we considered the possibility that the dimension of apparent logical correctness is discrete rather than continuous. Psychologically speaking, this is equivalent to the conventional idea that logical validity is an all-or-none variable rather than a continuous quantity. In terms of modeling, this is called a double-high threshold process (Macmillan & Creelman, 2005). Such an assumption would mean that only valid arguments can have high values on this axis, and only invalid arguments can have low values. We modeled this by assuming that arguments had a rectangular distribution along the logical correctness dimension and no correlation between strength axes. We retained the assumption that judgments reflect a combination of information on both axes, so that the decision bounds did not parallel the strength dimensions.

Over the range of parameter values we have assessed, the predicted ROCs from this model are reasonably consistent with those observed for deduction, but are unsuccessful in describing the inductive data. Specifically, the simulated inductive ROCs have much higher hit rates than are actually observed at the most conservative criteria (the left-most portion of the predicted ROC falls above and to the left of the 95% CIs). Finally, an alternative threshold account, known as a low-threshold model (Macmillan & Creelman, 2005) would treat reasoning as a single process with a binary outcome. Such a model can account for "corners" in the ROC curves (see Figure 2), but would otherwise have no

predictive power—the location of the corner would be two free parameters for each argument in each condition.

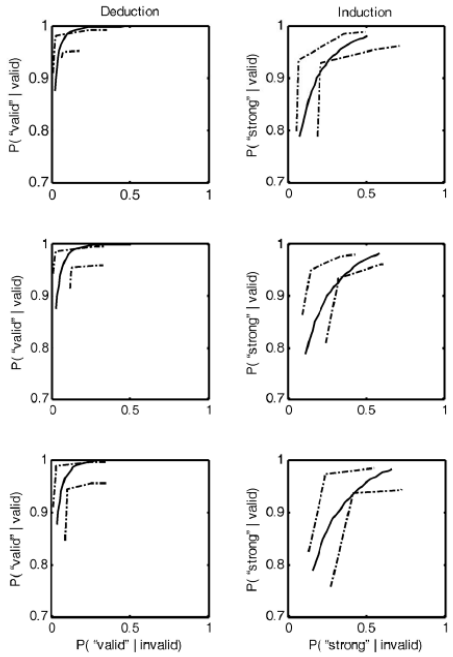

Figure 6 ROCs simulated from continuous two-dimensional model, analogous to Figure 4

General Discussion

The experimental results and modeling lead to the following conclusions. We replicated the result of Heit and Rotello (2005) showing greater sensitivity to validity for deduction judgments than for induction judgments, over a common set of stimuli. To this we have added the result that induction judgments are more sensitive to number of premises than are deduction judgments. Along with Rips (2001), these are some of the few experiments directly pitting one- and two-process accounts of reasoning against each other. All of the results point to two underlying processes of reasoning, contributing differentially to induction and deduction.

We have begun to develop multidimensional models of reasoning to accommodate these results and others. (For example, in other experiments we have found another dissociation between induction and deduction, namely that premise-conclusion similarity has a greater effect on induction.) One-dimensional models of reasoning have been implemented and do not seem able to fit the results, because they predict too great of a likeness between induction and deduction. As an existence proof, a two-dimensional model has been implemented, fitting the main results.

There seems to be great promise in this approach, particularly in terms of future model development. For example, one question to ask is what information goes into the second, associative dimension, and should this dimension itself be split further, reflecting distinct non-deductive subprocesses? In general, what is needed for more progress in model development are additional experiments that systematically influence various aspects of the model. For example, experiments that manipulate subjects' attention to associative or logical strength should influence the slope of the decision criterion, and experiments that manipulate the quality of the evidence should affect the locations of the distributions of arguments. Overall, the two-dimensional model of reasoning makes many testable predictions to be evaluated in future research.

Acknowledgments

This work was supported by a collaborative grant from the National Science Foundation (BCS-0616979). We thank Brooklynn Edwards, Efferman Ezell, Chanita Intawan, Nic Raboy, and Haruka Swendsen for assistance, and Shane Mueller for comments on an earlier draft.

References

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91-101.

Evans, J. St. B. T. & Over, D. E. (1996). *Rationality and reasoning.* Hove: Psychology Press.

Harman, G. (1999). *Reasoning, meaning, and mind*. Oxford: Oxford University Press.

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

Heit, E., & Rotello, C. M. (2005). Are there two kinds of reasoning? *In Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P.N. (1994). Mental models and probabilistic thinking. *Cognition, 50,* 189-209.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Cambridge: Cambridge University Press.

Oaksford, M., & Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. Oxford: Oxford University Press.

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

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science, 12*, 129-134.

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

Sloman, 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.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26,* 1-12.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal-detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*, 616-641.