Are There Two Kinds of Reasoning?

Evan Heit (E.Heit@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry CV4 7AL, UK

Caren M. Rotello (caren@psych.umass.edu)
Department of Psychology, University of Massachusetts
Amherst MA 01003-7710, USA

Abstract

Two experiments addressed the issue of how deductive reasoning and inductive reasoning are related. According to the criterion-shift account, these two kinds of reasoning assess arguments along a common scale of strength, however there is a stricter criterion for saying an argument is deductively correct as opposed to just inductively strong. The method, adapted from Rips (2001), was to give two groups of participants the same set of written arguments but with either deduction or induction instructions. Signal detection and receiver operating characteristic analyses showed that the difference between conditions could not be explained in terms of a criterion shift. Instead, the deduction condition showed greater sensitivity to argument strength than did the induction condition. Implications for two-process and one-process accounts of reasoning, and relations to memory research, are discussed.

Introduction

How do convincing arguments differ from non-convincing arguments? Rips (2001) has referred to the intuitive case for a single psychological dimension of argument strength, in which arguments can range from utterly worthless to completely compelling. Hence, the convincingness of an argument could be judged by assessing its position on the scale, in a similar manner to how judgments of loudness or brightness would use a psychophysical scale.

This intuition of a unitary scale needs to be reconciled with the notion that there are different kinds of reasoning. In particular there is the textbook distinction between deduction and induction, with deduction being concerned with drawing logically valid conclusions as opposed to induction which involves drawing plausible inferences. Strictly speaking, there are different kinds of arguments, such as deductively correct arguments, with respect to a well-defined logic, and inductively strong arguments (Skyrms, 2000). It is still an open question whether there are different kinds of reasoning, such as deductive reasoning and inductive reasoning.

Some researchers have suggested that rather than having specialized cognitive processes for each kind of reasoning, people use a common set of reasoning processes for both deductive and inductive arguments. For example, Chater and Oaksford (2000) have applied an account of probabilistic reasoning, explicitly non-deductive in nature, to a range of deductive problems. Likewise, Harman (1999) has argued that people reason in an essentially non-deductive way, and bring these same reasoning processes to bear on both inductive and deductive reasoning problems. Taking a related approach, Johnson-Laird (1994) has extended the mental models account, more frequently applied to deductive problems, to a range of inductive problems as well. Finally, some researchers have proposed accounts that focus mainly on reasoning about inductive arguments, and have treated deductively correct arguments as special cases that would be covered by the same accounts (Heit, 2000; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993).

In contrast, other researchers have emphasized a distinction between two kinds of reasoning (e.g., Evans & Over, 1996; Sloman, 1996; Stanovich, 1999). In these two-process accounts there is one system that is relatively fast but heavily influenced by context and associations, and another system that is more deliberative and analytic or rule-based. Although these two systems do not necessarily correspond directly to induction and deduction, it is plausible that induction would depend more on the first system whereas deduction would depend more on the second system. In addition there is some neuropsychological evidence, based on brain imaging, for two anatomically separate systems of reasoning (Goel, Gold, Kapur, & Houle, 1997; Parsons & Osherson, 2001).

These one- and two-process proposals are mainly aimed at accounting for a range of phenomena rather than drawing a sharp line between deduction and induction. In contrast, the proposal by Rips (2001) does not aim for a detailed description of reasoning processes but instead focuses on a key commonality and a key difference between deduction and induction. In his account, there is a single scale for evaluating arguments. This account will be referred to as the criterion-shift account, and it is illustrated in Figure 1. Here, the unitary scale of argument strength is shown, with different points on the scale corresponding to arguments of different strengths. Criterion 1 indicates the dividing line between arguments that are inductively weak, or implausible, and arguments that are inductively strong, or plausible. In order to make an assessment of deductive correctness, the criterion would be shifted rightwards, to Criterion 2. Some arguments might be strong enough to be judged plausible but not strong enough to be judged deductively correct, whereas other arguments might be so strong that they are also judged to be deductively correct.

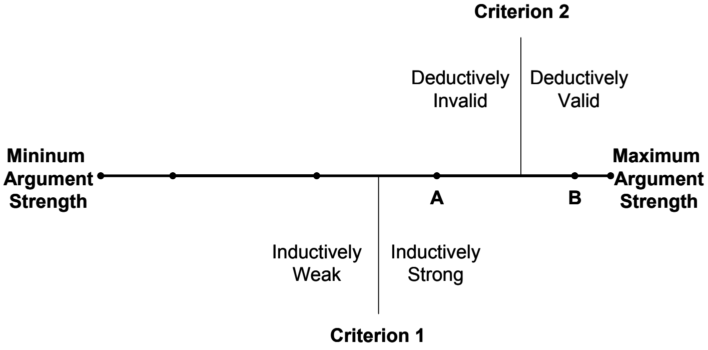

Figure 1: Criterion-shift account of deduction and induction.

An important virtue of the criterion-shift account is that it makes a number of testable predictions regarding the relations between deduction and induction. One prediction is that the relative ordering of two arguments should be the same whether people are judging deductive correctness or inductive strength. If one argument is more likely to be called deductively correct than another, then this argument should also more likely be called inductively strong. Rips (2001) assessed this prediction by comparing two types of arguments in two experimental conditions, in which participants were instructed to judge either deductive correctness or inductive strength (see Parsons & Osherson, 2001, for a related technique). 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". Participants 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 contradicted the criterion-shift account, which predicted a monotonic ordering of arguments in the two conditions.

However, it may be possible to rescue the criterion-shift account by making the reasonable assumption that participants who judged inductive strength were more likely to use background knowledge than participants who judged deductive correctness. In effect, participants in the induction condition considered other premises, such as "Rolling in the mud tends to make people dirty", based on their own knowledge. The uncertain nature of inductive inferences makes it particularly appropriate to take account of other background knowledge (e.g., Heit, Hahn, & Feeney, 2005; Skyrms, 2000). So it is still possible that the induction and deduction conditions did use a common scale of argument strength, but the participants were relying on different information in the two conditions, and therefore reached different conclusions.

The present experiments were aimed at testing another, closely-related, prediction of the criterion-shift account, while avoiding the potential problem of people introducing different background knowledge in the deduction and induction conditions. Referring back to Figure 1, note that any two arguments, such as A and B, will have a fixed distance between them regardless of the response criterion. That is, whether people are making judgments of inductive strength, and applying Criterion 1, or are making judgments of deductive correctness, and applying Criterion 2, the distance should be constant. In terms of signal detection theory (SDT) (e.g., Macmillan & Creelman, 2005), the difference in responses to A and B, expressed in standardized units like *d′*, should be the same in the deduction and induction conditions. On this view, the only change between conditions is the more conservative response criterion in the deduction condition.

Both experiments were closely modeled on the method of Rips (2001), giving either deduction or induction instructions to two groups of participants who otherwise saw the same set of arguments. Because the results could depend on how exactly the participants are instructed to perform deduction or induction, Rips had compared three different versions of instructions for both deduction and induction. However, there were no differences found, so the present experiments only used one version for deduction and one version for induction.

The arguments in Experiment 1 were created by modifying arguments from the Rips (2001) study, in effect stripping out their meaning so that background knowledge would not be useful. For example, "Jill rolls in the mud" was replaced with "Jill does D". There were two types of arguments, deductively correct and deductively incorrect. Hence this experiment allowed an assessment of the criterion-shift account without the problem of possibly different use of background knowledge for deduction versus induction. Experiment 2 had somewhat different stimuli, that allowed participants to use knowledge of category inclusion, which would be relevant in both the deduction and induction conditions. For example, one correct argument was "All birds have property C, therefore all robins have property C", after taking account of category membership. The criterion-shift account would predict for both experiments that the difference in responses to correct arguments and incorrect arguments, expressed in *d′* units, should be the same in the deduction and induction conditions. In contrast, a substantial change in *d′* from induction to deduction would make it difficult to explain deduction and induction as following the same scale of argument strength but varying only in response criterion.

Experiment 1

Method

There were 40 participants in the deduction condition and 40 in the induction conduction, all University of Warwick undergraduates. The instructions for the deduction condition 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". Likewise for the induction condition, there was a definition of a strong argument, "assuming the information above the line is true, this makes the sentence below the line **plausible**".

Each questionnaire contained 8 questions, presented in one of two random orders. The questions were of the following form in the deduction condition.

```
Jill does D and Jill does R
-------------------------------
Jill does D

Assuming the information above the line
is true, does this necessarily make the
sentence below the line true?

Circle one: VALID or NOT VALID
```

In the induction condition, questions were of the following form.

```
Jill does D and Jill does R
-------------------------------
Jill does D

Assuming the information above the line
is true, does this make the sentence
below the line plausible?

Circle one: STRONG or NOT STRONG
```

Each forced-choice judgment was followed by a confidence rating, on a 1 to 7 scale, with 7 corresponding to maximum confidence.

The 8 arguments themselves were the same for the two conditions. There were 4 deductively correct arguments as in the above example, and 4 deductively incorrect arguments, such as "Robert does not do V, therefore Robert does S". The arguments were adapted from Rips (2001), replacing elements of the arguments with uninformative letters so that participants could not use background knowledge. For example, the corresponding argument for Jill used by Rips was "Jill rolls in the mud and Jill gets clean, therefore Jill rolls in the mud".

Results and Discussion

For the deduction condition, the mean proportion of positive or "valid" responses for correct arguments was .89 and the corresponding proportion for incorrect arguments was .22. Although this proportion of "valid" responses for incorrect arguments may seem somewhat high, Rips (2001) reported a similar value (20%). For the induction condition, the mean proportion of positive or "strong" responses for correct arguments was .93 and the proportion for incorrect arguments was .55. The results were examined using a two-way analysis of variance (ANOVA), with instructional condition and correctness of argument as independent variables. The overall proportion of positive responses was significantly higher in the induction condition than in the deduction condition, F(1,78)=20.28, MSE=.07, p<.001. The overall proportion of positive responses was higher for correct arguments than for incorrect arguments, F(1,78)=122.06, MSE=.09, p<.001. There was also a significant interaction between these two variables, F(1,78)=9.67, MSE=.09, p<.01.

In terms of sensitivity, that is, ability to distinguish between correct and incorrect arguments, the greater difference in the deduction condition suggests a greater level of discrimination. For each participant, a d' measure was calculated. (Comparable conclusions were obtained with alternative measures such as d_a.) The average d' was significantly higher in the deduction condition, 1.68, than in the induction condition, 0.93, t(78)=3.11, p<.01. Response criterion was not calculated because this is difficult to compare between two conditions that differ in d' (Macmillan & Creelman, 2005; see also Heit, Brockdorff, & Lamberts, 2003).

A further analysis used not only choice proportions but also confidence ratings, to allow the plotting of receiver operating characteristic (ROC) curves and estimation of their slopes. In this case, an ROC curve plots the probability of a positive ("valid" or "strong") response to valid arguments on the y-axis and to invalid arguments on the x-axis; the points indicate varying levels of confidence, with higher-confidence positive decisions appearing to the left in the space (see Macmillan & Creelman, 2005). Figure 2 shows the *z*ROC curves (normal-normal transformations of the ROCs) for this experiment. The curves are approximately linear, as they should be when derived from underlying Gaussian distributions of argument strength. It should also be clear that the curve for the deduction condition is more distant from the origin than is the curve for the inductive condition, corresponding to the previous conclusion that sensitivity is greater for deduction. If deduction and induction had equal sensitivity and different response criteria, then the curves for the two conditions would be co-linear. The deduction instructions did also lead to more conservative responding, as can be seen in the leftward and downward translation of the points in that condition. Finally, it should be noted that the slopes in Figure 2 differ. The slope for deduction is .84 and the slope for induction is .60. The slope indicates the ratio of standard deviations of the invalid and valid argument distributions. This result suggests that the range of acceptable items was narrower in the deduction condition than in the induction condition.

In sum, the results were not consistent with the criterion-shift account, which would represent differences between deduction and induction solely as a change in response criterion. Instead, there were also changes in sensitivity, and in slopes of zROC curves, that would not be predicted by the criterion-shift account. Hence, the results fit those of Rips (2001) who also found differences between deduction and induction that could not be explained by a change in criterion.

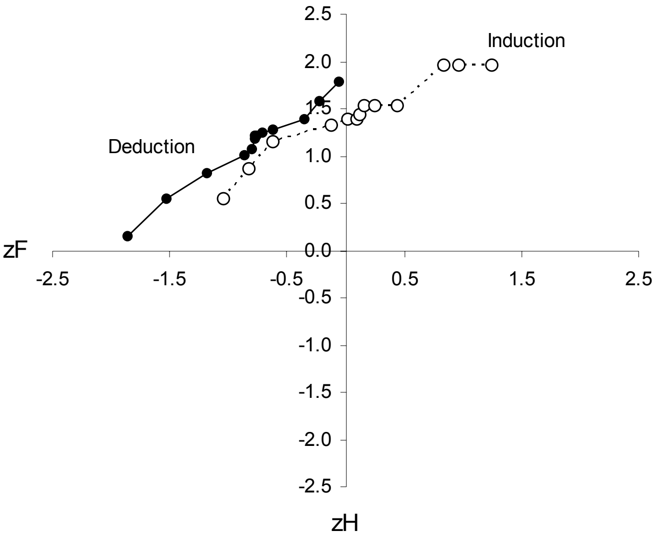

Figure 2: zROC curves for Experiment 1, comparing deduction and induction instructions

Next, we present a second experiment, which applied a similar method to different materials, adapted from previous studies of inductive reasoning (Osherson et al., 1990; Sloman, 1993, 1998). Osherson et al. (1990) had suggested that correct arguments such as "All birds have property C, therefore all robins have property C" would be considered perfectly strong, but Sloman found that such arguments were not considered perfectly strong when participants were asked to judge convincingness or conditional probability. Furthermore, Sloman reported a typicality effect (which was referred to as inclusion similarity) in which other correct arguments like "All birds have property D, therefore all penguins have property D" were considered even weaker, in terms of inductive strength, because they involved atypical rather typical category members. However, Sloman did not ask participants to judge deductive correctness. In Experiment 2, these correct arguments, as well as other incorrect arguments, were compared using deduction and induction instructions.

Experiment 2

Method

The method was like Experiment 1 except for the following. There were 48 participants in the deduction condition and 48 in the induction condition. The same 16 arguments were used in the two conditions. There were 4 types of critical arguments. The first type was correct-typical, such as "All birds have property C, therefore all robins have property C". The second type was correct-atypical, such as "All birds have property D, therefore all penguins have property D". The third type was incorrect-typical, such as "All robins have property E, therefore all birds have property E". The final type was incorrect-atypical, such as "All penguins have property F, therefore all birds have property F". There were 3 versions of each type of critical argument, hence there were 12 critical arguments in total. One version used "birds" as a category; another used "mammals" as a category, with "horses" and "dolphins" as typical and atypical category members; and the third used "fruits" with "apples" and "blueberries". Finally, there were 4 filler arguments that were deductively incorrect and indeed not especially plausible, such as "All bees have property A, therefore all elephants have property A". The filler arguments were intended to give the stimuli a wide range of plausibility.

Results and Discussion

The proportions of positive responses for types of critical arguments are shown in Table 1. The main result was that, as in Experiment 1, the proportion of positive responses on correct arguments was about the same for the deduction and induction conditions, but there was a higher proportion of positive responses on incorrect arguments in the induction condition than in the deduction condition. It appeared that participants more sharply distinguished between correct and incorrect arguments in the deduction condition than in the induction condition.

Table 1
Proportion of Positive Responses for Experiment 2.

Argument Type	Deduction Condition	Induction Condition
Correct-Typical	.94	.90
Correct-Atypical	.89	.85
All Correct	.92	.88
Incorrect-Typical	.25	.60
Incorrect-Atypical	.24	.53
All Incorrect	.24	.57

In addition, there appeared to be a small effect of typicality, that is, there was a higher proportion of positive responses for arguments with typical category members than for arguments with atypical category members. For the correct arguments in the induction condition, this finding corresponds to the inclusion similarity phenomenon reported by Sloman (1993, 1998). For the incorrect arguments, the finding corresponds to the premise typicality phenomenon previously reported in studies of inductive reasoning (Osherson et al., 1990; see Heit, 2000, for a review).

Responses to the critical items were analyzed with a three-way ANOVA, with instructional condition, correctness of argument, and typicality as independent variables. The overall proportion of positive responses was significantly higher in the induction condition than in the deduction condition, F(1,94)=10.25, MSE=.19, p<.01. The overall proportion of positive responses was higher for correct arguments than for incorrect arguments, F(1,94)=143.29, MSE=.16, p<.001, and there was a significant interaction between these two variables, F(1,94)=19.22, MSE=.16, p<.001. Finally, the main effect of typicality was significant, F(1,94)=6.91, MSE=.03, p<.01. None of the remaining interactions approached statistical significance, all F's < 1.

For each participant, a *d'* measure was calculated, comparing the overall proportion of positive responses on correct arguments versus incorrect arguments. The average *d'* was significantly higher in the deduction condition, 1.69, than in the induction condition, 0.78, t(94)=4.39, p<.001. For the deduction condition, *d'* was almost the same as in Experiment 1; for the induction condition *d'* was again relatively poor.

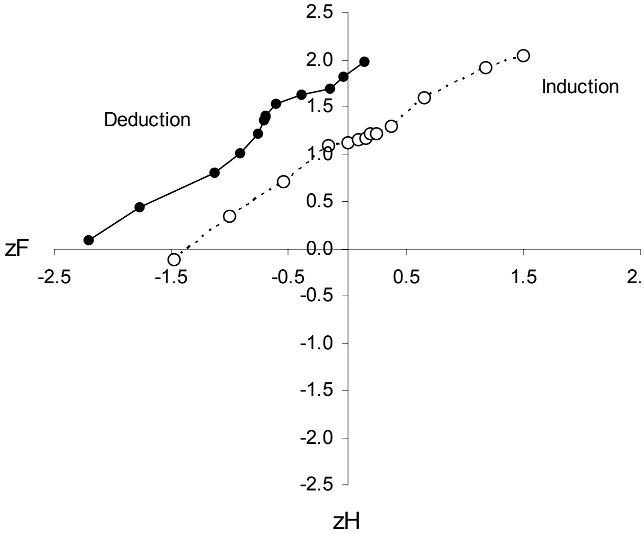

Figure 3: zROC curves for Experiment 2, comparing deduction and induction instructions

Next, the confidence ratings were used to plot zROC curves, as shown in Figure 3. The pattern is similar to Experiment 1. The deduction curve is more distant from the origin, corresponding to greater sensitivity compared to induction. Again, the slope is steeper for deduction, .82, compared to .71 for induction.

In sum, the main result again was greater discrimination between correct and incorrect arguments in the deduction condition compared to the induction condition. It is not possible to explain this result in terms of the criterion-shift account

General Discussion

There is a striking parallel between the issue of whether there are two kinds of reasoning and a central issue in memory research. In memory research there is an important distinction between one- and two-process accounts (Rotello & Heit, 1999; Yonelinas, 2002). According to two-process accounts of recognition memory, recognition judgments depend on a quick, approximate, familiarity-based process and a slower, more deterministic process based on specific item recollection. In effect, there are two different kinds of recognition, because either process could dominate a recognition judgment. In contrast, according to one-process accounts, it is not necessary to assume two processes in order to explain experimental results. This distinction has come up in the context of whether remembering and knowing correspond to different processes. According to some researchers (Donaldson, 1996; Dunn, 2004; Wixted & Stretch, 2004) the distinction between remembering and knowing is simply a matter of a criterion shift, i.e., both judgments are based on a common scale of memory strength, but there is a stricter criterion for saying that something is directly remembered. Hence, in terms of SDT, the difference between remembering and knowing should appear as a change in response criterion rather than sensitivity. However, recent assessments (Gardiner, Ramponi, & Richardson-Klavehn, 2002; Rotello, Macmillan, & Reeder, 2004) have rejected a one-dimensional signal detection model. In particular, in memory research there are standard signs taken as evidence against a single process, such as unequal sensitivity for different types of judgments on the same memory probes, slope differences in ROC curves, and a non-monotonic relationship between the two types of judgments across a set of probes. On this basis, Rotello et al. have proposed a two-dimensional model, incorporating information from familiarity and recollection.

By the standards of memory research, there is already a good case against a single process account of reasoning. Putting together the present two experiments with the experiment reported by Rips (2001), there is already evidence for sensitivity differences, slope differences, and non-monotonicity. Still, we think it is too early to rule out single-process accounts of reasoning. Often with signal detection analyses, it is valuable to examine the pattern over a large set of experiments, e.g., the Dunn (2004) and Rotello et al. (2004) analyses were based on hundreds of previous recognition experiments.

Furthermore, it would be desirable, on the basis of further experimentation, to develop a two-process account of reasoning, using two-dimensional SDT (e.g., Ashby & Gott, 1988; Rotello et al., 2004). It could be assumed that both induction and deduction rely on two sources of information, one derived from quick and context-dependent associations and the other from more controlled, and possibly-rule based, deliberations. The criterion for distinguishing convincing from non-convincing arguments would be a line in this two-dimensional space. For judgments of deductive correctness, the criterion would depend heavily on information from the controlled deliberations and less on contextual information. The criterion for judgments of inductive strength could be less strict in terms of the dimension for controlled deliberations and also take more account of the other dimension, corresponding to contextual associations. Hence, it would be predicted that inductive judgments have less sensitivity in distinguishing deductively correct from deductively incorrect arguments, and are more likely to take account of other background knowledge in evaluating arguments.

It is best to think of these signal detection accounts as analytical tools, allowing some predictions of one- and two-process models to be sharpened and tested, in particular allowing clearer predictions regarding the relations between deduction and induction, and allowing these models to be developed further. In conclusion, the novel technique applied by Rips (2001), of directly comparing deductive and inductive judgments on the same set of arguments, when combined with signal detection analysis, appears to have considerable promise for developing accounts of reasoning that more explicitly address the fundamental distinction between deduction and induction.

Acknowledgments

We are grateful to Lance Rips for providing stimuli from Rips (2001) and to Elisabeth Blagrove and Yana Weinstein for assistance in conducting this research.

References

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 33-53.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese, 122*, 93-131.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523-533.

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review, 111*, 524-542.

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

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory, 10*, 83–98

Goel, V., Gold, B., Kapur, S., Houle, S. (1997). The seats of reason: A localization study of deductive and inductive reasoning using PET (O15) blood flow technique, *NeuroReport, 8,* 1305-1310.

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

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

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review, 10,* 718-723.

Heit, E., Hahn, U., & Feeney, A. (2005). Defending diversity. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff, (Eds.), *Categorization inside and outside of the laboratory: Essays in honor of Douglas L. Medin*, 87-99. Washington, DC: APA.

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.

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

Parsons, L. M., & Osherson, D. (2001). New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex, 11,* 954-965.

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

Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject? *Journal of Memory and Language, 40*, 432-453.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal detection model. *Psychological Review*, *111*, 588–616.

Skyrms, B. (2000). *Choice and chance: An introduction to inductive logic.* (Fourth edition). Belmont, CA: Wadsworth.

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

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

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441-517.