Journal of Experimental Psychology:
Learning, Memory, and Cognition
2009, Vol. 35, No. 5, 1317–1330

0278-7393/09/$12.00 DOI: 10.1037/a0016648

Modeling the Effects of Argument Length and Validity on Inductive and Deductive Reasoning

Caren M. Rotello
University of Massachusetts

Evan Heit
University of California, Merced

In an effort to assess models of inductive reasoning and deductive reasoning, the authors, in 3 experiments, examined the effects of argument length and logical validity on evaluation of arguments. In Experiments 1a and 1b, participants were given either induction or deduction instructions for a common set of stimuli. Two distinct effects were observed: Induction judgments were more affected by argument length, and deduction judgments were more affected by validity. In Experiment 2, fluency was manipulated by displaying the materials in a low-contrast font, leading to increased sensitivity to logical validity. Several variants of 1-process and 2-process models of reasoning were assessed against the results. A 1-process model that assumed the same scale of argument strength underlies induction and deduction was not successful. A 2-process model that assumed separate, continuous informational dimensions of apparent deductive validity and associative strength gave the more successful account.

Keywords: reasoning, mathematical modeling

How are inductive reasoning and deductive reasoning related? 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 about the arguments themselves. In contrast, according to the *process* view, the question of interest is what cognitive processes characterize induction and deduction and what the similarities and differences between inductive reasoning and deductive reasoning are in processing (see Heit, 2007, for a review). A particularly promising approach is to address the process view as an empirical question, asking people to make either induction judgments or deduction judgments for a common set of arguments (Rips, 2001; see also Osherson et al., 1998). In general, this technique could highlight either similarities or differences between induction and deduction that are not confounded by the use of different problems.

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 is referred to as the *one-process* view. For example, Harman (1999) argued that people do not reason deductively but instead use the same reasoning abilities for both deductive and inductive problems. Several influential research programs embody the one-process view, by applying a common modeling framework to both inductive and deductive problems, assuming a single scale of evidence for argument strength. For example, Oaksford and Chater (2007) showed that a Bayesian reasoning model, probabilistic in nature, can be successfully applied to problems of deduction. Note that Oaksford and Chater (2002) themselves have not denied that people can explicitly perform deductive tasks under limited circumstances, and even probabilistic inferences can sometimes be justified on logical grounds (e.g., Pfeifer & Kleiter, 2009). The key point is that this line of work offers a unifying framework for problems of deduction and induction, based on a single scale of probability. Likewise, Johnson-Laird (1994) explained how mental model theory, typically applied to problems of deduction, can be applied to problems of induction, for example, stating, “The same explanatory framework accommodates deduction and induction” (p. 189). Osherson, Smith, Wilkie, Lopez, and Shafir (1990) and Sloman (1993) presented models of inductive reasoning that, without additional assumptions, account for some deductive reasoning phenomena (e.g., that arguments based on identity matches between a premise and a conclusion are perfectly strong).

According to *two-process* accounts (Evans & Over, 1996; Sloman, 1996; Stanovich, 1999), both heuristic and analytic processes contribute to reasoning, with each process potentially assessing an argument as strong or weak. In effect, there is more than one scale of evidentiary strength. Both induction and deduction could be influenced by these two processes but in different proportions. Induction judgments would be particularly influenced by quick heuristic processes that tap into associative information about context and similarity that do not necessarily make an argument logically valid. In contrast, deduction judgments would be more heavily influenced by slower analytic processes that encompass more deliberative, and typically more accurate, reasoning. Two-

Caren M. Rotello, Department of Psychology, University of Massachusetts; Evan Heit, School of Social Sciences, Humanities, and Arts, University of California, Merced.

This work was supported by collaborative National Science Foundation Grant BCS-0616979. A preliminary report of Experiment 1a appeared in Heit and Rotello (2008). We thank Charles Brainerd, Jonathan Evans, and David Over for comments and Lissette Alvarez, Brooklynn Edwards, Efferman Ezell, Chanita Intawan, Nic Raboy, Anamika Sethi, and Haruka Swendsen for assistance with this research.

Correspondence concerning this article should be addressed to Caren M. Rotello, Department of Psychology, University of Massachusetts, Box 37710, Amherst, MA 01003. E-mail: caren@psych.umass.edu

process accounts have provided an explanatory framework for many results (e.g., content effects, individual differences, effects of time pressure). Although the distinction between one-process and two-process accounts is conventional in reasoning research, the boundary might seem to be in need of further sharpening. To be as explicit as possible, the one-process accounts of reasoning cited above do not have separate processes for induction versus deduction, do not identify separate mechanisms for heuristic and analytic processing, do not rely on separate sources of associative and logical information, and more generally do not have two scales of evidentiary strength.

The one-process view and the two-process view are each embodied by several highly productive and successful research programs. However, relatively little research has directly pitted the one-process view and the two-process view against each other. One important exception is a study by Rips (2001) that compared two types of arguments in two experimental conditions: Participants were instructed to make either deduction judgments or induction judgments. 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 making deduction or induction judgments. 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. Although he did not endorse a particular two-process account, Rips concluded that the results were evidence against the one-process account, which predicts the same ordering of arguments in both conditions, with only a potential change in response bias to distinguish them (see also Oaksford & Hahn, 2007, for an analysis).

Heit and Rotello (2005, 2008) pointed out that distinguishing between one-process and two-process views is also an important enterprise in memory research. In particular, there is a debate about whether recognition memory can be accounted for by a single familiarity process or whether there are two processes, a fast familiarity process that is 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 participants make a recognition judgment, then state whether they just know that 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 (Donaldson, 1996; Dougal & Rotello, 2007).

Memory researchers have developed several standards for examining whether a set of results points to one or two processes. 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 (Dunn & Kirsner, 1988). 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 responses 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 require only a different response criterion than knows, then d' for remembers should equal d' for old decisions (see Macmillan, Rotello, & Verde, 2005). 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. ROC curves that fall higher in space (toward the upper left corner of the graph) reflect greater sensitivity because the hit rate is greater for a given false alarm rate (see Figure 1). In addition, points that fall to the upper right along a given ROC reflect more liberal response biases because both 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.

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 had 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, e.g., 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 several aims. First, in all three experiments, we varied the number of premises in an argument. Although increasing the number of premises does not in itself make an argument valid, research on inductive reasoning has shown that providing more evidence can make a conclusion seem more plausible (cf. Heit, 2000; Osherson et al., 1990). In particular, we predicted that for invalid arguments, increasing the number

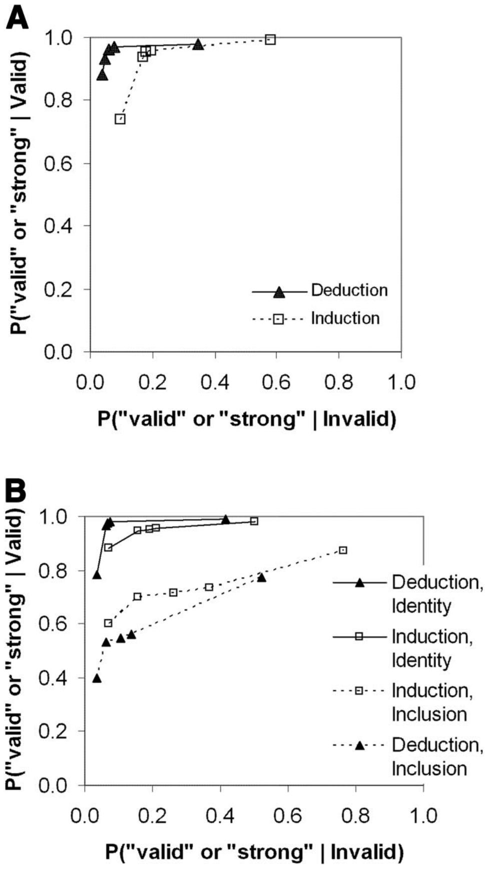

Figure 1. Induction and deduction receiver operating characteristics from (A) Experiment 1a and (B) Experiment 1b.

of premises would affect induction more than deduction, which would be more sensitive to actual validity. There is a parallel in social cognition research, that longer communications lead to greater attitude change than shorter communications, under conditions expected to promote automatic or heuristic processing (e.g., Petty & Cacioppo, 1984).

Second, we included a fluency manipulation that was expected to enhance the role of analytic processing for inductive reasoning. In Experiment 2, participants made induction judgments with materials presented in either a good, fluent or harder to read, disfluent font. Alter, Oppenheimer, Epley, and Eyre (2007) used a similar manipulation for deduction judgments on syllogisms, finding greater accuracy with a bad font. Here our aim was to use the fluency manipulation in a way similar to the comparison between induction and deduction instructions; we kept the arguments the same between the good and bad fonts but expected a different mixture of analytic versus heuristic processing for the two font conditions.

Third, we generated predictions 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. Moreover, we assessed alternative versions of one-process and two-process models. For example, the modeling included an evaluation of the possibility that there is only one process but with variable response criteria (Wixted & Stretch, 2004). A way to think about this is that if either induction or deduction judgments are poorly understood by the participants, this could be reflected in trial-to-trial or person-to-person variability in how the rating scales are applied. With regard to two-process models, we evaluated both the conventional notion that information about logical validity is an all-or-none variable and the possibility that it is on a continuous scale.

Experiment 1

Experiment 1 was run in two variants. In both variants, the main comparison was an instructional manipulation: Participants were asked to make either induction judgments or deduction judgments. Participants saw both valid and invalid arguments, with varying argument lengths (one, three, or five premises). In Experiment 1a, all the valid arguments were identity matches; for example, a statement such as "Horses have Property X" appeared as one of the premises as well as the conclusion. Experiment 1b included a wider variety of valid arguments. In addition to identity matches, there were arguments based on superordinate category inclusion; for example, "Mammals have Property X" appeared as a premise, and "Horses have Property X" appeared as the conclusion. Technically, such an argument is an enthymeme, as it relies on a hidden premise that all horses are mammals (Calvillo & Revlin, 2005). Sloman (1998) found that inclusion arguments were judged to be weaker than identity arguments, although that study did not compare deduction and induction instructions.

We hypothesized that the participants would be more sensitive to argument length in the induction condition and more sensitive to validity in the deduction condition. We expected that this pattern of results would prove difficult for a one-process account of reasoning, but could be fit by a two-process account.

Method

Participants. In Experiment 1a, 60 University of California, Merced, students were paid to participate. Participants were randomly assigned to one of two conditions: induction ($n = 31$) or deduction ($n = 29$). In Experiment 1b, there were 63 participants: 32 in the induction condition and 31 in the deduction condition.

Stimuli. In Experiment 1a, 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." Participants 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: 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.
Rabbits have Property X.
Cats have Property X.

Rabbits have Property X.

The materials in Experiment 1b were similar. The only change was that of the 60 valid arguments, 45 involved an identity relation (as in Experiment 1a) and 15 involved an inclusion relation. The 3:1 ratio of identity versus inclusion relations was maintained for valid arguments having one, three, or five premises. Following is an example of a valid argument with an inclusion relation:

Mammals have Property X.

Horses have Property X.

Procedure. Each experiment was run with a program on a computer; each participant participated individually. At the beginning of the experiment, participants were given instructions on the definition of strong or valid arguments. Specifically, as in Rips (2001), participants in the induction condition were told that strong arguments were those for which "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."

The 120 arguments were presented one at a time, in a different random order for each participant. In the induction condition, participants were again told to assume that the information above the line is true and to assess whether the sentence below the line was plausible. They pressed one of two keys to indicate "strong" or "not strong." In the deduction condition, participants were again told to assume that the information above the line is true and to assess whether the sentence below the line was necessarily true. They indicated "valid" or "not valid" with a key press. Each binary decision was followed with a confidence rating on a 1–5 scale; higher numbers indicated greater confidence. The confidence ratings were used to generate empirical ROCs; although inclusion of a rating task tends to increase the binary decision time (Baranski & Petrusic, 2001), it does not affect the accuracy of those judgments (Baranski & Petrusic, 2001; Egan, Schulman, & Greenberg, 1964).

Results

In Experiment 1a, 3 participants from the induction condition were excluded from the analyses because they either gave the same response for virtually every question or showed little difference in responses to valid versus invalid arguments (d' comparing response rates with the two item types was less than 0.5). In Experiment 1b, 4 participants from the induction condition and 3 participants from the deduction condition were excluded, according to the same criteria.

To provide an overview of the data, we first assessed the proportion of positive ("strong" or "valid") responses to valid and invalid arguments (see Table 1). For the deduction condition, the average proportions were .96 and .06, respectively, in Experiment 1a and .87 and .07 in Experiment 1b. For the induction condition, the average proportions were .95 and .18, respectively, in Experiment 1a and .89 and .19 in Experiment 1b. As in previous experiments (Heit & Rotello, 2005, 2009), d' was greater for deduction (3.31 in Experiment 1a and 2.60 in Experiment 1b) than induction (2.56 in Experiment 1a and 2.10 in Experiment 1b),

Table 1
Proportions of Positive Responses From Experiments 1a, 1b, and 2

Number of premises	Experiment 1a		Experiment 1b		Experiment 2	
	Induction	Deduction	Induction	Deduction	Induction good font	Induction bad font
Not valid						
1	.11	.05	.11	.06	.17	.12
3	.17	.06	.19	.06	.19	.18
5	.26	.07	.28	.09	.23	.32
Average	.18	.06	.19	.07	.20	.21
Valid identity						
1	.98	.98	1.00	1.00	.98	.99
3	.94	.95	.93	.97	.94	.98
5	.94	.96	.92	.96	.95	.95
Average	.95	.96	.95	.97	.95	.97
Valid inclusion						
1			.86	.73	.72	.90
3			.62	.44	.48	.70
5			.66	.46	.41	.63
Average			.71	.55	.54	.74
Valid all						
1	.98	.98	.96	.93	.91	.97
3	.94	.95	.85	.84	.83	.91
5	.94	.96	.86	.84	.82	.87
Average	.95	.96	.89	.87	.85	.92

indicating greater sensitivity to argument validity for deduction than for induction.

Because d' depends on participants' willingness to say "valid" or "strong" (i.e., their response bias) unless the underlying evidence distributions are equal-variance and Gaussian in form (Macmillan & Creelman, 2005; Rotello, Masson, & Verde, 2008), we also evaluated the area under the ROC curve in each condition. ROCs plot participants' hits ("valid" or "strong" responses to valid problems) against their false alarms ("valid" or "strong" responses to invalid problems) as a function of their confidence or response bias. The observed induction and deduction ROCs from Experiment 1 are shown in Figure 1. The left-most point on the deduction ROC indicates the hit and false alarm rate for problems that participants said were "sure valid"; the next point on the ROC includes the hits and false alarms at a slightly lower level of confidence, and so on. Points that fall on a common ROC curve reflect the same level of performance but different response biases. Accuracy, or the ability to discriminate valid problems from invalid ones, is higher when the hit rate is higher for a given false alarm rate, so ROC curves that fall higher in the space reflect better performance. The area under the ROC, A_z, which ranges from 0.5 (chance performance) to 1.0 (perfect discrimination), is a good summary of overall accuracy (Swets, 1986); specialized software (ROCKIT; Metz, 1998) can be used to compare A_z between independent conditions. In Experiment 1a, A_z was significantly higher in the deduction condition than in the induction condition (z-score-based test statistic = 9.45, $p < .001$). Likewise, in Experiment 1b, for identity problems, A_z was significantly higher in the deduction condition than in the induction condition ($z = 4.80$, $p < .001$). For the inclusion problems introduced in Experiment 1b, the area under the curve was somewhat greater for induction than deduction, but this difference did not reach the level of statistical significance ($z = 1.58$, $p > .10$). We return to this point in the *Discussion*.

The design of Experiment 1b allows comparison of two types of valid arguments: those based on identity relations and those based on inclusion relations. The results in the deduction condition replicate Sloman (1998) in showing a higher proportion of positive responses to identity relations (97% of the time) than to inclusion relations (55%). The induction condition showed the same pattern, with more positive responses to identity relations (95%) than inclusion relations (71%).

Our second main prediction was that argument length, or number of premises, would have a greater influence on induction judgments than deduction responses. The results in Table 1 support this prediction. For invalid arguments, there appears to be greater sensitivity to number of premises for induction compared with deduction in both experiments. For example, in the induction condition of Experiment 1a, the response rate to invalid arguments increased from .11 for one-premise arguments to .26 for five-premise arguments, whereas for deduction the corresponding response rate increased only from .05 to .07. An analysis of variance (ANOVA) on the false alarm rates in Experiment 1a as a function of condition (induction or deduction) and number of premises (one, three, or five) supports this observation: The false alarm rates were higher in the induction condition than the deduction condition, $F(1, 55) = 7.644$, $MSE = .083$, $p < .01$, $\eta^2 = .122$, and they increased with number of premises, $F(2, 110) = 8.406$, $MSE = .011$, $p < .001$, $\eta^2 = .133$. However, condition and number of premises interacted, $F(2, 110) = 5.347$, $MSE = .011$, $p < .01$, $\eta^2 = .089$; the effect of number of premises on false alarm rate was reliable only in the induction condition, $F(2, 54) = 8.690$, $MSE = .017$, $p < .01$, $\eta^2 = .243$, not in the deduction condition, $F(2, 56) < 1$, $\eta^2 = .016$. Similarly, in Experiment 1b, the ANOVA revealed that the false alarm rates were significantly higher in the induction condition, $F(1, 54) = 6.463$, $MSE = .057$, $p < .02$, $\eta^2 = .107$, and that they increased with number of premises, $F(2, 108) = 13.087$, $MSE = .013$, $p < .001$, $\eta^2 = .195$. These variables interacted, $F(2, 108) = 5.810$, $MSE = .013$, $p < .01$, $\eta^2 = .097$: The effect of number of premises on false alarm rate was greater in the induction condition, $F(2, 54) = 10.065$, $MSE = .024$, $p < .01$, $\eta^2 = .271$, than in the deduction condition, $F(2, 54) = 3.792$, $MSE = .003$, $p < .05$, $\eta^2 = .125$.

Our main prediction was for the effect of argument length on invalid arguments. Although we did not have strong predictions for the effect of argument length on valid arguments, the data in Table 1 suggest that longer arguments are somewhat weaker. This trend runs through all the valid argument data in both experiments but is most noticeable for valid inclusion relations in Experiment 1b. Indeed, an ANOVA on the hit rates in Experiment 1b revealed a main effect of argument length, $F(2, 108) = 22.962$, $MSE = .034$, $p < .001$, $\eta^2 = 0.32$, and a main effect of inclusion–identity status, $F(1, 54) = 93.309$, $MSE = .100$, $p < .001$, $\eta^2 = 0.63$, that was qualified by an interaction between these two effects: $F(2, 108) = 15.455$, $MSE = .023$, $p < .001$, $\eta^2 = 0.22$. The decrease in hit rate for longer arguments was larger for the inclusion problems but was also significant for the identity problems considered alone, $F(2, 108) = 8.679$, $MSE = .006$, $p < .01$, $\eta^2 = .14$. In Experiment 1a, longer valid arguments also elicited fewer positive responses than shorter arguments: $F(2, 110) = 4.790$, $MSE = .006$, $p < .05$, $\eta^2 = 0.08$. The interaction of argument length with condition was not reliable in either experiment (both Fs < 1).

Discussion

In general, the results pointed to two distinct effects: Validity had a greater effect on deduction judgments, and argument length had a greater effect on induction judgments. As expected, there was a higher proportion of positive responses to identity arguments than to inclusion arguments. One interesting but unanticipated finding was that there was a clear difference between the induction and deduction conditions for identity arguments but not for inclusion arguments (e.g., the difference in area under the ROC curve only reached the level of statistical significance for identity arguments). There are several possible reasons for this finding (which also appeared in another experiment, in Heit & Rotello, 2009). One reason is simply that were fewer inclusion questions than identity questions, so it may be a matter of low power. Relatedly, because the response proportions to inclusion questions were closer to the chance level of 50%, there could be more variability in these data. However, there could also be theoretical rather than methodological reasons for the apparent difference between identity and inclusion arguments. Because the inclusion arguments are enthymemes, it could be said that they are not deductively valid arguments taken in isolation (without assuming hidden premises). Whatever cognitive processes led participants in the deduction condition to a high level of sensitivity to the distinction between identity arguments and invalid arguments may simply be less effective for inclusion arguments.

Another interesting but unanticipated finding was the differing effect of argument length on invalid versus valid arguments. Whereas invalid arguments became stronger as they got longer, valid arguments became weaker as they got longer. Note that unlike the logical validity manipulation, argument length does not have an objective effect on the strength of an argument. Just as an invalid argument may seem stronger because more plausible evidence is brought to bear, valid arguments may seem more compelling, elegant, or parsimonious, and hence stronger, when they are simpler or briefer (cf. Lombrozo, 2007). The finding of weaker inferences on valid arguments with more premises also resembles a result from Stevenson and Over (1995), who found that people were less willing to make a *modus ponens* inference when an argument had a premise added. However, in that study, the extra premise was designed to cast doubt on another premise. It is unclear whether adding premises in our own experiments led to some doubting of premises (e.g., whether informing participants that horses have X would make them doubt another premise such as that mice have X). And it is notable that adding premises strengthened, rather than weakened, invalid arguments.

For present purposes, all the results in Experiments 1a and 1b provide a useful test bed for assessing models of reasoning.

Modeling

Our general approach was to implement one-process and two-process models and apply them to the results from Experiments 1a and 1b. We operationalized the one- versus two-process distinction in terms of dimensions in SDT. One-dimensional SDT makes standard assumptions that there is a single dimension of stimulus strength, in this case distinguishing weak arguments from strong arguments. Whatever differences arise between deduction and induction must take the single dimension of strength as a starting point. To anticipate, we did not find a successful one-dimensional model for the whole pattern of results. We then turned to a two-dimensional SDT model, allowing that arguments can differ in two dimensions of strength. We first fit such a model to the results of Experiment 1b, which had more conditions and hence more data points, then observed the match between those fitted predictions and the results of Experiment 1a.

One-dimensional model. One possible, and simple, version of a one-dimensional model is shown in Figure 2: Valid problems have greater average strength than invalid problems, as do invalid problems with greater argument length. In this model, the only difference between deduction and induction responses is the response criterion: Deduction responses are assumed to require a higher level of evidence at each level of confidence, as indicated by the rightward shift of the deduction criteria compared with the induction criteria. This model fails to account for the results of Experiments 1a and 1b. These data showed greater sensitivity to validity for deduction than induction, but the model predicts the same sensitivity for both. Because this model assumes that only response bias differs between induction and deduction, it predicts that the resulting data will fall on a common ROC curve on which the deduction points are each shifted to the left of the corresponding induction points, reflecting a more conservative response bias at each confidence level but having the same sensitivity in terms of A_z. The data also showed greater sensitivity to argument length for induction than deduction, but the model predicts the same level of sensitivity to both. This model assumes that the invalid distribution has a mean that varies with argument length for both deduction and induction. In combination with the assumption that deduction decision criteria are more conservative at each confidence level, this implies that the resulting pattern of false alarms as a function of argument length must be the same for deduction and induction. In other words, this model can predict either that argument length increases the false alarm rate (as seen in the induction condition) or, if the invalid distributions are assumed to be the same regardless of argument length, that argument length does not affect the false alarm rate (as seen in the deduction condition). It cannot predict both patterns simultaneously. For these reasons, we can eliminate this simplest one-dimensional model from consideration.

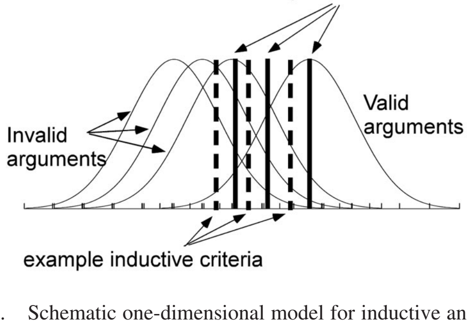

Figure 2. Schematic one-dimensional model for inductive and deductive decisions.

The one-dimensional model could be modified by assuming that, for some reason, the locations of participants' confidence criteria were more variable across trials in the induction condition; criterion variability decreases the apparent accuracy in a condition (e.g., Mueller & Weidemann, 2008; Wickelgren, 1968) and could therefore account for the lower sensitivity in the induction task. (Anecdotally, our participants 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 predict the performance of a one-dimensional model like this over a wide range of parameter values and with the possibility of different valid evidence distributions for each argument length (i.e., like Figure 2 but with three valid distributions). The modeling was intended to be illustrative. For each simulation, we defined the invalid one-premise distribution to be a Gaussian distribution with a mean of 0 and a standard deviation of 0.6. Then we allowed the means of the invalid three- and five-premise problems to be shifted upward by different amounts ranging from 0 (identical to the one-premise invalid distribution) to 0.6 (a 1 standard deviation increment). The mean of the valid one-premise problem distribution was allowed to vary from 0.8 to 1.6, with a fixed standard deviation of 1, and the means of the three- and five-premise valid distributions were allowed to shift from 0 to 0.6 units below the valid one-premise problems. We also simulated the mean locations of five decision criteria, reflecting differing levels of response confidence, for the induction condition and another five for the deduction condition; these latter were shifted upward by 0.1 units relative to the induction condition. Finally, we allowed the locations of these decision criteria to vary from trial to trial: The standard deviation of the criterion location over trials was varied from 0 (no variability) to 0.6.

For each combination of parameter values, we sampled 1,000 strengths from each distribution, reflecting simulated problems, and compared those strengths with the decision criteria to determine the response. For example, sampled strengths that fell above the highest confidence induction criterion were assumed to yield "highest confidence strong" responses but would lead only to "highest confidence valid" responses if they also exceeded the most stringent deduction criterion. From these simulated trials, we generated predicted response rates for each set of parameter values; the best fitting parameters were selected from those evaluated by minimizing the mean square error of prediction. We did not seek to find the best possible fits overall, but we are confident that these simulations capture the essential predictions of the model, because we sampled a large number of parameter combinations and because the consequences of, say, shifting decision criteria or increasing their variability in a one-dimensional model are well understood within SDT (see Macmillan & Creelman, 2005).

For the data in Experiment 1a, the best fitting parameter values indicated that there were effectively two distributions for invalid problems (one for one- and three-premise problems and another for five-premise problems) and two distributions for valid problems (one for one-premise problems and another for three- and five-premise problems). In the best fitting parameterization of the model, there was no criterion variability, meaning that this model predicts, in contrast with the results, the same level of accuracy for both deduction and induction judgments ($d' = 2.4$, 2.3, and 1.8 for one-, three-, and five-premise problems). Because the invalid distributions vary with argument length, the false alarm rate is also predicted to vary with argument length, for both deduction (ranging from .11 to .21), in contrast with the results, and induction (ranging from .14 to .27).

We also fit the data from Experiment 1b, allowing separate distributions to represent the valid identity and inclusion problems. Specifically, we assumed that the number of premises and the inclusion–identity characteristic of the valid problems had independent effects on the mean value of a valid distribution. Here the model fit implied that all the invalid distributions had the same mean, so that the predicted false alarm rate did not vary with argument length, in contrast with the results in the induction condition. Four valid distributions were inferred, one each for one-premise inclusion and identity problems, one for three- and five-premise inclusion problems, and one for three- and five-premise identity problems; in each case, the inclusion distribution had a lower mean strength than the identity distribution. In the best fitting parameterization of the model, criterion variability was again zero, so that the same level of accuracy was predicted for both deduction and induction, in contrast with the empirical data.

In summary, a one-dimensional model that assumed that the same distributions, but potentially different response criteria, are used for both deduction and induction failed to capture the basic patterns in the data. One reason for this failure is clear in Table 1: The false alarm rates are higher in the induction condition than in the deduction condition in both Experiments 1a and 1b, yet the hit rates are quite similar across tasks. Given the assumption that the evidence distributions are the same for both conditions, the observed hit rates for induction and deduction imply that the same criterion is used for both conditions. However, that implication is in conflict with the observed differences in the false alarm rates across conditions: If the same criterion is used, thus generating similar hit rates, then the false alarm rates should also be similar. Allowing for variability in criterion locations cannot resolve this fundamental inconsistency. In one-dimensional models like this one, the presence of criterion variability will increase the estimated false alarm rate and simultaneously decrease the estimated hit rate. Our data are not consistent with that predicted outcome, and thus our modeling efforts led to the conclusion that there was no evidence for criterion variability.

Two-dimensional model. Having ruled out one-dimensional representations in which the same information is used for both deduction and induction, we next considered the possibility that two (orthogonal) dimensions of information were used in the judgments (see Figure 3). The dimensions can be thought of as "apparent logical correctness" and "consistency with associative knowledge," on the principle that these would be the outputs of analytic and heuristic processing, respectively, although other labels are possible (see the General Discussion). Our starting assumptions were that valid arguments would differ more from invalid arguments along the logic axis than along the associative axis and that the number of premises would influence the strength of evidence along the associative axis. Invalid arguments would be generally (but not uniformly) low on apparent logical correctness but would vary in their consistency with associative knowledge; a greater number of premises would result in greater total similarity between premise and conclusion categories. Both valid and invalid arguments were assumed to have bivariate Gaussian distributions; the covariance was allowed to be either zero (implying independence of the information on the two dimensions) or positive (implying some dependence). To make either an induction or deduction judgment in this model, a criterion is required. We

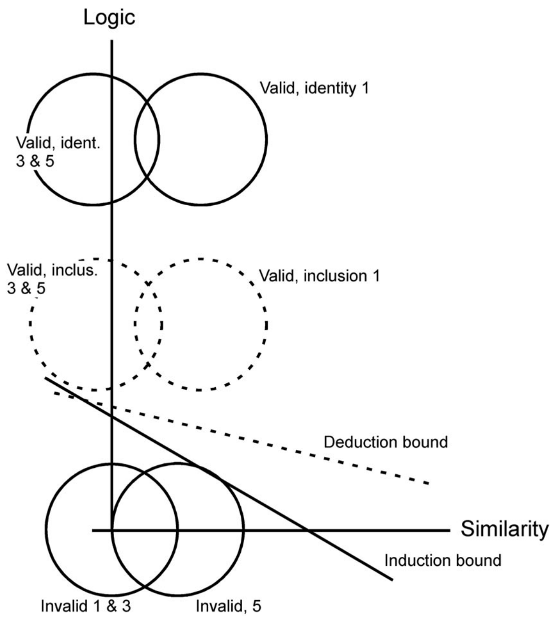

Figure 3. Schematic two-dimensional model for inductive and deductive decisions.

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. The relative weight given to the two dimensions is reflected in the angle of the decision bound that divides positive from negative decisions in each task: Deduction places a relatively greater emphasis on logic, and therefore the slope of that decision bound is shallower, as is shown schematically in Figure 3. Because the slope of the decision bound is assumed to differ for induction and deduction, and because the mean value of a distribution may be greater on one dimension than the other, this model naturally allows that accuracy for inductive and deductive decisions will differ (without the need to consider criterion variability). Differences in the pattern of false alarms with argument length are also possible, depending on the angle of the decision bounds.

We simulated this two-dimensional model using Monte Carlo sampling over a wide range of parameter values. 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 simulated with this two-dimensional model would fall within the 95% confidence intervals of the observed ROCs for both induction and deduction, and for one-, three-, and five-premise problems, assuming that the only difference between the tasks was the slope of the decision criterion. To provide additional constraints on the model, we assumed that the valid identity and inclusion problems from Experiment 1b were represented by different distributions, with inclusion differing from identity only on the logic dimension, and we simultaneously generated the ROCs for both problem types. Figures 4 (identity problems) and 5 (inclusion problems) show that we were reasonably successful: The dashed functions are the upper and lower bounds of the 95% confidence intervals for each observed ROC, and the solid functions are the model-generated ROCs. In each case, the predicted ROCs generally fall within the confidence limits and are slightly higher for deduction than for induction, reflecting higher predicted accuracy in the deduction condition. As argument length increases, the predicted ROCs tend to shift rightward along the *x*-axis, more so for induction that deduction, reflecting greater tendency to respond positively to invalid arguments when they are longer. Also, as argument length increases, the predicted ROCs tend to shift downward along the *y*-axis, reflecting a lower tendency to respond positively to valid arguments when they are longer. Table 2 shows the parameter values used to generate the ROCs in Figures 4 and 5; the schematic two-dimensional representation in Figure 3 approximates these values. With these parameters, the predicted false alarm rates are .09, .12, and .15 for one, three, and five premises in the deduction condition and .11, .12, and .22 in the induction condition, showing a larger influence of argument length on induction responses to invalid problems (as in Table 1).

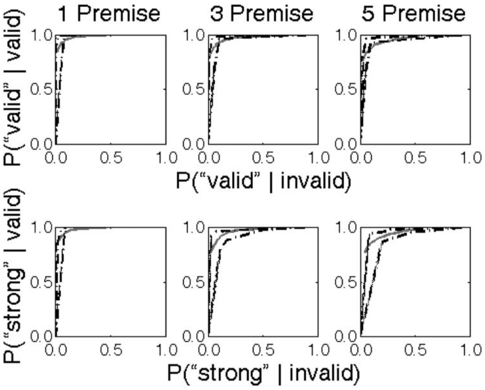

Figure 4. Simulated receiver operating characteristics (ROCs) from the two-dimensional model (solid function) and 95% confidence intervals for the observed ROCs in Experiment 1b, for the valid identity problems. The upper row shows the deduction condition; the lower row shows the induction condition.

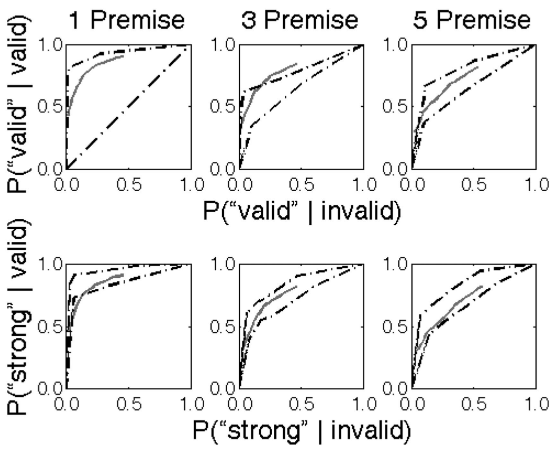

Figure 5. Simulated receiver operating characteristics (ROCs) from the two-dimensional model (solid function) and 95% confidence intervals for the observed ROCs in Experiment 1b, for the valid inclusion problems. The upper row shows the deduction condition; the lower row shows the induction condition. The model's parameter values are the same as in Figure 4 (see Table 2).

Averaging over argument length, the model predicts d' to be 2.73 in the deduction condition and 2.51 in the induction condition. This d' difference is more modest than the data show, but there are several reasons to be cautious about using d' as a measure of accuracy in the model. First, the model does not assume that the distributions of valid and invalid problems have the same variance; second, response bias may differ across conditions. Under those conditions, d' is a poor estimator of accuracy (Macmillan & Creelman, 2005; Rotello et al., 2008), which is why we have focused on the ROCs. The predicted ROCs, shown in Figures 4 and 5, are reasonably consistent with the empirical ROCs. Note that our model-fitting procedure targeted fit to the ROCs rather than fit to d' or, for that matter, to the hit rates and false alarm rates.

We also applied this model to the data from Experiment 1a, using the same parameter values as for Experiment 1b. The only change we made to the model was to eliminate the inclusion distributions, which are not relevant for the design of Experiment 1a. The results are shown in Figure 6. As for the fit to Experiment 1b's data, the deduction ROC curves are slightly higher in the space, reflecting greater sensitivity to validity in that condition than in the induction condition, and the curves shift slightly to the

Table 2
Parameter Values for the Two-Dimensional Model as Applied to Each Experiment

Parameter	Experiment 1a	Experiment 1b	Experiment 2
d_x = mean of valid one-premise problems on x-axis	0.5	0.5	0.5
Variance of d_x	0.4	0.4	0.4
d_y = mean of valid one-premise problems on y-axis	3.6	3.6	3.6
Variance of d_y	2.0	2.0	2.0
Induction (or good font) slope	−0.5	−0.5	−0.5
Deduction (or bad font) slope	−0.3	−0.3	−0.3
Change in d_y for inclusion problems		−1.8	−2.0
Change in d_x for valid three-premise problems	−1.1	−1.1	−1.5
Change in d_x for valid five-premise problems	−1.1	−1.1	−1.7
Change in d_x for invalid three-premise problems	0.0	0.0	0.0
Change in d_x for invalid five-premise problems	0.6	0.6	0.6
Covariance of x and y for valid problems	0.0	0.0	0.0
Covariance of x and y for invalid problems	0.0	0.0	0.0

right, reflecting higher false alarm rates, as the number of premises increases. Although better quantitative fits could likely be found, the model captures the basic trends in the data from Experiment 1a without any modification.

We are very 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 the number of premises increases in the induction condition.

Alternative two-dimensional models. We next considered the possibility that the dimension of apparent logical correctness is discrete rather than continuous. This is equivalent to the conventional idea that logical validity is an all-or-none variable rather than a continuous quantity (e.g., Skyrms, 2000). In 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 dimension and only invalid arguments can have low values. We modeled this possibility 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.

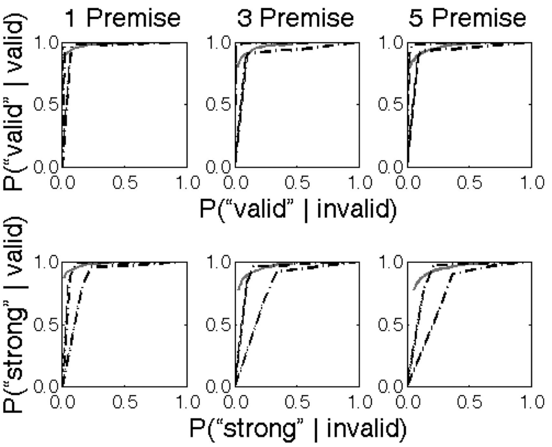

Figure 6. Simulated receiver operating characteristics (ROCs) from the two-dimensional model (solid function) and 95% confidence intervals for the observed overall ROCs in Experiment 1a. The upper row shows the deduction condition; the lower row shows the induction condition. The model's parameter values are the same as for the fit to the Experiment 1b data (see Figures 4 and 5) except that there was no need for inclusion distributions (see Table 2).

Over the large range of parameter values we have assessed, the predicted ROCs from this model are not consistent with the empirical ROCs. Specifically, the resulting ROCs are fairly linear and have shallow slope, whereas the observed ROCs are distinctly nonlinear (see Figures 4–6). Linear ROCs are typical of threshold models. The likelihood that a sampled stimulus strength was drawn from one of the two stimulus classes, rather than the other, is constant over the range of strengths on which the two classes' distributions overlap. (Consider two overlapping rectangular distributions: The ratio of their heights is constant over their region of overlap.) Therefore, shifting the decision criterion within that range affects the hit and false alarm rates to the same degree, which yields a linear ROC. Because our observed ROCs are not linear, we did not pursue this model any further.

Finally, an alternative threshold account, known as a low-threshold model (Macmillan & Creelman, 2005), could treat reasoning as a single process with a binary outcome. Such a model can account for "corners" in the ROC curves (see Figure 1) but would otherwise have no predictive power—the location of the corner would be two free parameters for each argument in each condition.

Summary. The modeling results provided converging evidence for our main conclusions. In light of the differences between the induction and deduction conditions, we were unable to fit a one-dimensional model to these results, although of course we do not rule out the possibility that a future one-dimensional model could be developed. The differences between induction and deduction were readily accommodated by a two-dimensional model assuming separate continuous scales of evidence for apparent logical correctness and associative strength.

Experiment 2

Having successfully applied a two-dimensional model to the results of Experiments 1a and 1b, we set out to assess this model on a similar experiment, also varying validity and argument length but with an additional manipulation. In Experiment 2, we manipulated fluency by displaying the materials in either a good, readable font or a bad, less readable font. It was expected that using a disfluent font would encourage the use of analytic processes, increasing sensitivity to validity of an argument (Alter et al., 2007). According to our two-dimensional model, an increased reliance on analytic processes would be reflected in a shallower decision slope. Only induction instructions were used because participants had shown such a high level of sensitivity to argument validity for deduction instructions in Experiments 1a and 1b that it seemed unlikely that sensitivity could be increased further. Our model-based prediction is that the primary difference between conditions would be in the slope of the decision bound: shallower for the bad font condition and steeper for the good font condition. It is also possible, of course, that the font itself would influence the strength of encoding. The bad font condition might force participants to pay more attention to the problems, thus yielding greater strength of evidence on the logic and associative information dimensions. That greater strength would be reflected in larger values for the parameters that locate the means of the distributions along the axes.

Method

The method of Experiment 2 was the same as in Experiment 1b except for the following: Sixty-nine University of California, Merced, students participated, 34 in the good font condition and 35 in the bad font condition. All participants received induction instructions.

The good font was Courier New, which was the same font as in Experiments 1a and 1b. The bad font was also Courier New but was italicized and had lower contrast: The letters were a silver (gray) color rather than black.

Results

Four participants from the good font condition and 6 participants from the bad font condition were excluded, according to the same criteria as in Experiments 1a and 1b.

Table 1 provides an overview of the results in proportion of positive responses, for valid versus invalid arguments. For the bad font condition, the average proportions were .92 and .21, respectively. For the good font condition, the average proportions were .85 and .20, respectively. On the basis of these averages, d' was higher for the bad font condition (2.21) than for the good font condition (1.88), suggesting greater sensitivity to argument validity for the bad font condition. The ROCs for both conditions are shown in Figure 7. Accuracy, measured by A_z, was reliably higher in the bad font condition than in the good font condition (ROCKIT's z test statistic = 6.21, $p < .001$), as predicted; this was true for both the identity problems ($z = 5.92$, $p < .001$) and the inclusion problems ($z = 3.57$, $p < .01$).

Argument length had an effect in both conditions, not surprisingly, because all participants made induction judgments, so this is essentially a replication of the induction conditions in Experiments 1a and 1b. As in those experiments, increasing argument length made invalid arguments stronger. An ANOVA on the false alarm rates in Experiment 2 as a function of condition (good or bad font) and number of premises (one, three, or five) supports this observation: The false alarm rates did not differ with condition, $F(1, 57) < 1$, $MSE = .083$, $p > .8$, and they increased with number of premises, $F(2, 114) = 17.61$, $MSE = .016$, $p < .001$, $\eta^2 = .236$. Condition and number of premises interacted, $F(2, 114) = 5.06$, $MSE = .016$, $p < .05$, $\eta^2 = .081$; the effect of number of premises on false alarm rate was reliable in the bad font condition, $F(2, 56) = 14.69$, $MSE = .022$, $p < .001$, $\eta^2 = .344$, but was smaller and marginally significant in the good font condition, $F(2, 58) = 3.12$, $MSE = .010$, $p < .07$, $\eta^2 = .097$. Increasing argument length also made valid arguments weaker, as in Experiment 1: The hit rate was higher in the bad font condition, $F(1, 57) = 8.37$, $MSE = .023$, $p < .01$, $\eta^2 = .312$, and decreased with argument length in both conditions, $F(2, 114) = 25.83$, $MSE = .006$, $p < .001$, $\eta^2 = .128$. The interaction of condition and number of premises was not reliable, $F(2, 114) < 1$.

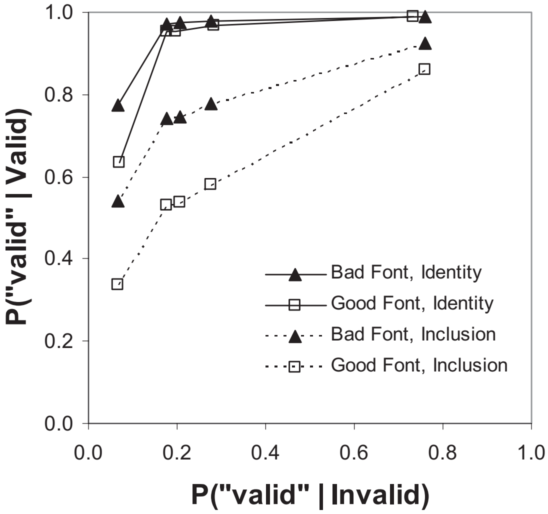

Figure 7. Observed receiver operating characteristics from Experiment 2.

Discussion

Overall, the results were consistent with those of Experiments 1a and 1b in showing that making an argument longer strengthened invalid arguments and weakened valid arguments. In addition, the main prediction was supported, namely that introducing a disfluent font increased sensitivity to validity. Note that unlike Experiment 1b, Experiment 2 did not use deduction instructions, so this experiment did not directly investigate the issue of how inclusion arguments compare under deduction versus induction instructions.

Given the overall similarity to Experiments 1a and 1b, we approached modeling Experiment 2 by looking to make the minimal change necessary, in terms of predicted changes in the slope of the decision bound.

Modeling

We applied the two-dimensional model of Experiment 1b to these data, varying only three parameters of those earlier simulations (see Table 2). The shallower decision bound from the deduction condition of Experiment 1b was associated with the bad font condition in Experiment 2. The results are shown in Figure 8 for the identity problems and in Figure 9 for the inclusion problems. Although better fits might be found with a thorough consideration of parameter space, the basic patterns in the data are reflected in this simulation (lower sensitivity to validity for the bad font condition, similar effects of argument length as in Experiment 1b), which is sufficient for current purposes.

General Discussion

The experimental results imply that induction judgments and deduction judgments draw on different cognitive resources, even when people are judging the same arguments. Put another way, there is not just a single scale of evidence for evaluating the strength of arguments. Experiments 1a and 1b highlighted differences between induction and deduction. People are more influenced by argument length for induction, and they are more influenced by validity for deduction. Implementations of one-process models of reasoning were not able to accommodate these results, but an implemented two-process model gave a successful account, by assuming that apparent logical validity and associative strength contribute to both kinds of judgments but in different proportions. This model was tested further in Experiment 2, which included a fluency manipulation, with a bad font intended to increase the influence of analytic processes on induction judgments. Participants showed a greater influence of validity in the bad font condition, and the results of Experiment 2 were successfully accommodated by the same two-process model, by assuming that the bad font led to a change in the slope of the decision boundary, representing greater sensitivity to apparent logical validity.

In a related study, Heit and Rotello (2009) found two analogous effects. One of their experiments varied premise–conclusion similarity (i.e., similarity between the premise and conclusion categories in one-premise arguments) as well as logical validity. Given the same set of arguments, induction judgments were affected more by similarity, and deduction judgments were affected more by actual validity. These results, too, pointed to two-process accounts and were not readily explained by one-process accounts. In a follow-up experiment, we compared speeded deduction judgments and unspeeded deduction judgments. The speeded deduction judgments were more like induction judgments in showing a lesser impact of validity and a greater impact of validity. This result, too, is readily explained by a two-process account on the assumption that underlying processing needed to assess logical validity has a slower time course. Hence, speeding deduction judgments will lead to a relatively lower influence of validity on responses.

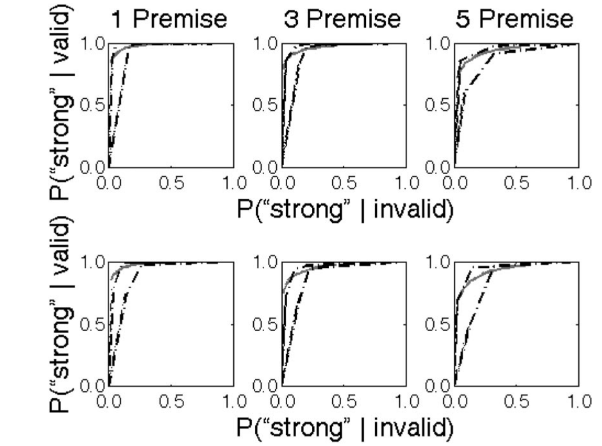

Figure 8. Simulated receiver operating characteristics (ROCs) from the two-dimensional model (solid function) and 95% confidence intervals for the observed ROCs to identity problems in Experiment 2. The upper row shows the bad font condition; the lower row shows the good font condition. The model's parameter values are nearly all the same as for the fit to the Experiment 1b data (see Table 2).

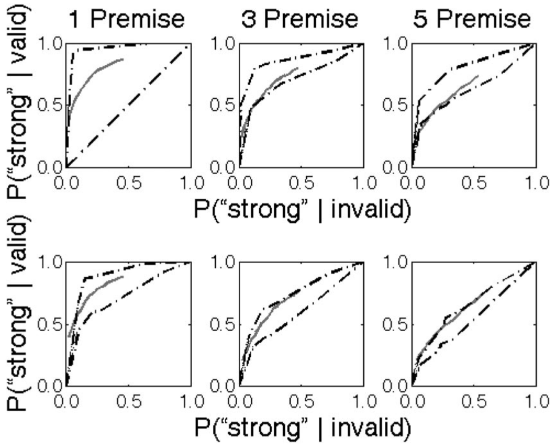

Figure 9. Simulated receiver operating characteristics (ROCs) from the two-dimensional model (solid function) and 95% confidence intervals for the observed ROCs to inclusion problems in Experiment 2. The upper row shows the bad font condition; the lower row shows the good font condition. The model's parameter values are nearly all the same as for the fit to the Experiment 1b data (see Table 2).

In the remainder of the General Discussion, we consider implications of these results for reasoning research and highlight some more general issues in model selection.

Implications for Reasoning Research

As noted in the introduction, one-process and two-process accounts of reasoning are each represented by several very successful research programs. Although one-process and two-process accounts stand in opposition, they are seldom pitted against each other directly. In one-process accounts, it is more typical to compare different accounts of deductive reasoning against one another (e.g., mental model theory vs. mental logical theory or probabilistic models) and different accounts of inductive reasoning against one another. Although two-process accounts have been used to explain many results, that work has not impeded research on one-process models. Hence, we see great value in implementing both one- and two-process models and comparing them in fit to common data sets. Our own work has implications for research on both one-process and two-process models.

For current one-process models in the literature (e.g., Johnson-Laird, 1994; Oaksford and Chater, 2007; Osherson et al., 1990; Sloman, 1993), we make the following points. These models do not make explicit predictions about reasoning under different instructional conditions, such as deduction versus induction instructions. To the extent that there are differences, and indeed there are, these would need to be accommodated by additional assumptions for these models. One possibility, suggested by Rips (2001), is that deduction judgments have a stricter response criterion, but such an account was not successful for either Rips's data or our own. We examined another possibility, that deduction and induction judgments could differ in variability of criteria, but that account was not successful for our data either. A third possibility, that two completely separate one-process accounts are used, one for deduction and another for induction, would abandon any assumption that the two types of judgments should be systematically related.

Still, by no means do we rule out one-process models based on our results, and we hope that our results will spur the development of future one-process models. However, we suspect that a successful one-process model would have to become more like a two-process model. It is noteworthy that none of the one-process models of reasoning that we have examined maintain separate sources of information about validity and argument length (or similarity). For example, Bayesian models of reasoning (Oaksford & Chater, 2007; see also Heit, 1998; Tenenbaum & Griffiths, 2001) measure the strength of an argument in the probability of its conclusion; within these models there are no component probabilities for validity and argument length that could be recombined to perform induction or deduction. Perhaps some existing one-process model could be modified so that it can tap into information sources that we are calling apparent logical validity and associative strength and use this information differentially for induction and deduction, but, again, such a model would be getting very close to two-process models in having two ways of evaluating an argument and two scales of argument strength.

With regard to two-process models, one contribution we have made is implementing specific models and providing a framework for describing various models (e.g., by showing that differences between induction and deduction can be explained in the slope of a decision boundary and by comparing continuous vs. discrete dimensional representation). Although our models represent explicit hypotheses about how two sources of evidence are brought to bear on making judgments, we do not claim to have developed a process model of reasoning. Two-process theories of reasoning come in different varieties (see Evans, 2008, for an overview). For example, analytic processes might follow heuristic processes sequentially, or these processes might run in parallel. The multidimensional models we have developed could constrain process models, but they are not process models themselves and thus make no claims about the sequence of the underlying processes. Indeed, our two-dimensional model does not make strong claims about the precise nature of the underlying dimensions; our choice of labels (x = similarity; y = logic) is arbitrary. A preference for alternative labels for the dimensions (such as gist and verbatim representations as in fuzzy-trace theory; Reyna & Brainerd, 2008) would not detract from our main point that deductive and inductive judgments are based on different weighted combinations of at least two underlying sources of information (for a similar point on modeling recognition memory judgments, see Macmillan & Rotello, 2006). One interesting possibility is that the outputs of existing process models could serve as inputs to our own multidimensional models. For example, a current model of deductive reasoning could provide information for the y-axis, serving as a proxy for apparent logical correctness, and a current model of inductive reasoning could provide information for the x-axis, serving as associative strength.

We hope that our results and analyses will encourage future modeling efforts, encompassing both traditional deduction problems and traditional induction problems, made under either deduction or induction instructions.

Issues in Model Selection

We selected a particular two-dimensional model to describe our data, after consideration of several alternative models. One general concern that arises in any model selection situation is the relative flexibility of the models under consideration: Models with greater flexibility are better able to mimic the data produced by other models (Pitt, Myung, & Zhang, 2002). Model flexibility generally increases with the number of free parameters but also varies with its functional form. For example, a two-parameter linear model ($y = a + bx$) has less flexibility than a two-parameter sinusoidal model ($y = a \sin bx$) because the parameters of the latter can be adjusted to fit either linear or curved functions.

One way of putting nonnested models with different numbers of parameters on a more equal playing field is to adjust their goodness-of-fit statistics for the number of parameters. Measures like Akaike's Information Criterion (Akaike, 1973) and the Bayesian Information Criterion (Schwarz, 1978) are commonly used for this purpose. Simply adjusting for the number of parameters fails to account for differences in the functional forms of competing models, as our linear–sinusoidal example demonstrates; more complicated techniques are available (e.g., Pitt et al., 2002; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004) that have sometimes been applied successfully to memory models (Cohen, Rotello, & Macmillan, 2008).

An altogether different strategy is to consider the type of data outcome that cannot be predicted by a particular model and then to evaluate whether such data have ever been observed empirically. If such data have been reported, support for that model is reduced; if those data do not exist (despite appropriate conditions having been run), support for that model is somewhat increased. This approach was recently adopted by Dunn (2008) to evaluate models of the remember–know task in memory research, and it has been argued to be the most powerful strategy for model selection (Myung, Pitt, & Navarro, 2007). It is exactly this strategy that allowed us to rule out the one-process model of reasoning sketched in Figure 2: It cannot simultaneously predict no change in false alarm rate with number premises (to fit the deduction data) and some change (to fit the induction data). Similarly, that one-process model cannot predict different values of sensitivity for deduction and induction without assuming that there is criterion variability, and the presence of criterion variability would increase the false alarm rate while simultaneously decreasing the hit rate. This predicted pattern of data is simply not consistent with our results, and thus support for the one-dimensional model is reduced.

The two-process models that we considered have more parameters and greater flexibility than the one-process model; they can fit a larger variety of data patterns. However, our conclusions about which model best captured the data were based on qualitative evaluations of what the models can and cannot predict rather than on quantitative comparisons of the goodness-of-fit measures. For example, we rejected a two-process model that assumed the dimension of apparent logical correctness is discrete rather than continuous because the predicted ROCs were always more linear than the empirical ROCs. Moreover, we developed the successful two-process model in a principled fashion (e.g., assuming that inclusion and identity problems differ only on the *y*-axis); the model is not completely flexible. Additional data sets that include confidence ratings and allow the construction of ROCs will be important for further progress; these will provide a more extensive test bed for model comparison and selection. The existing data (see also Heit & Rotello, 2009) are consistent in pointing to two continuous dimensions of argument strength that are differentially weighted depending on the reasoning task.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kiadó.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136,* 569–576.

Baranski, J. V., & Petrusic, W. M. (2001). Testing architectures of the decision–confidence relation. *Canadian Journal of Experimental Psychology, 55,* 195–206.

Calvillo, D. P., & Revlin, R. (2005). The role of similarity in deductive categorical inference. *Psychonomic Bulletin & Review, 12,* 938–944.

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember–know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review, 15,* 906–926.

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

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14,* 423–429.

Dunn, J. C. (2008). The dimensionality of the remember–know task: A state-trace analysis. *Psychological Review, 115,* 426–446.

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

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1964). Operating characteristics determined by binary decisions and by ratings. In J. A. Swets (Ed.), *Signal detection and recognition by human observers: Contemporary readings* (pp. 172–186). New York: Wiley.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.

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

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

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford, England: Oxford University Press.

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

Heit, E. (2007). What is induction and why study it? In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Cognitive, mathematical, and neuroscientific approaches* (pp. 1–24). Cambridge, England: Cambridge University Press.

Heit, E., & Rotello, C. M. (2005). Are there two kinds of reasoning? In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Meeting of the Cognitive Science Society* (pp. 923–928). Mahwah, NJ: Erlbaum.

Heit, E., & Rotello, C. M. (2008). Modeling two kinds of reasoning. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the Thirtieth Annual Meeting of the Cognitive Science Society* (pp. 1831–1836). Austin, TX: Cognitive Science Society.

Heit, E., & Rotello, C. M. (2009). *Relations between inductive reasoning and deductive reasoning.* Manuscript submitted for publication.

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

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232–257.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Macmillan, N. A., & Rotello, C. M. (2006). Deciding about decision models of remember and know judgments: A reply to Murdock. *Psychological Review, 113,* 657–665.

Macmillan, N. A., Rotello, C. M., & Verde, M. F. (2005). On the importance of models in interpreting remember–know experiments: Comments on Gardiner et al.'s (2002) meta-analysis. *Memory, 13,* 607–621.

Metz, C. E. (1998). ROCKIT [Computer software]. Retrieved January 4, 2006, from at http://xray.bsd.uchicago.edu/cgi-bin/roc_software.cgi

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494.

Myung, I. J., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review, 14,* 1043–1050.

Oaksford, M., & Chater, N. (2002). Commonsense reasoning, logic, and human rationality. In R. Elio (Ed.), *Common sense, reasoning, and rationality* (pp. 174–214). Oxford, England: Oxford University Press.

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

Oaksford, M., & Hahn, U. (2007). Induction, deduction and argument strength in human reasoning and argumentation. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Cognitive, mathematical, and neuroscientific approaches* (pp. 269–301). Cambridge, England: Cambridge University 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.

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

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology, 46,* 69–81.

Pfeifer, N., & Kleiter, G. D. (2009). Mental probability logic [Commentary on Oaksford & Chater: *Bayesian rationality: The probabilistic approach to human reasoning*]. *Behavioral and Brain Sciences, 32,* 98–99.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences, 18,* 89–107.

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

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70,* 389–401.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Skyrms, B. (2000). *Choice and chance: An introduction to inductive logic* (4th ed.). Belmont, CA: Wadsworth/Thomson.

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.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48*A, 613–643.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–641.

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

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48,* 28–50.

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5,* 102–122.

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

Received February 21, 2009
Revision received May 15, 2009
Accepted May 19, 2009 ■