The pervasive effects of argument length on inductive reasoning

Evan Heit¹ and Caren M. Rotello²

¹School of Social Sciences, Humanities and Arts, University of California, Merced, Merced, CA, USA
²Department of Psychology, University of Massachusetts Amherst, Amherst, MA, USA

Three experiments examined the influence of argument length on plausibility judgements, in a category-based induction task. The general results were that when arguments were logically invalid they were considered stronger when they were longer, but for logically valid arguments longer arguments were considered weaker. In Experiments 1a and 1b when participants were forewarned to avoid using length as a cue to judging plausibility, they still did so. Indeed, participants given the opposite instructions did not follow those instructions either. In Experiment 2 arguments came from a reliable or unreliable speaker. This manipulation affected accuracy as well as response bias, but the effects of argument length for both reliable and unreliable speakers replicated Experiments 1a and 1b. The results were analysed using receiver operating characteristic (ROC) curves and modelled using multi-dimensional signal detection theory (SDT). Implications for models of category-based inductive reasoning, and theories of reasoning more generally, are discussed.

Keywords: Inductive reasoning; Argumentation; Mathematical modelling.
The expression “weighing the evidence” suggests that making a judgement depends on assessing a quantity or mass. Whether or not an argument seems plausible depends on the amount of evidence presented in its favour. Presumably arguments with more evidence in the premises are stronger arguments. Research has shown that people are highly influenced by the length of an argument, but the situation is more complicated. One argument might be very long but still invalid, and another argument might be very short and perfectly valid. It’s also possible to imagine that, in some cases, short parsimonious arguments would be more convincing than long rambling arguments.

In this paper we address the pervasive effects of argument length on judgements of plausibility. First we review previous research on social cognition, argumentation, category-based inductive reasoning, and causal reasoning on the effects of argument length, as well as our own research (Rotello & Heit, 2009). Then we review theoretical accounts of reasoning that make predictions about the effects of argument length on plausibility judgements. Finally we present three new experiments that examine whether people can avoid using argument length when judging plausibility, or if argument length is so compelling that people cannot help being influenced by it.

PREVIOUS RESEARCH ON ARGUMENT LENGTH

Social cognition

Classic research on attitude change and persuasion has identified number of arguments as a key variable affecting belief in an overall conclusion. For example, Cook (1969) found that people were more likely to agree with counter-attitudinal statements like “cleaning the teeth more than once per week is harmful to health” when more supporting arguments were presented in its favour. Burnstein, Vinokur, and Trope (1973) examined the phenomenon of group polarisation in a risky choice situation. They found that the number of arguments presented in favour of a risky shift, rather than the number of people making arguments, predicted attitude change. In a mock jury study Calder, Insko, and Yandell (1974) varied the number of prosecution and defence arguments, and found that jurors were more likely to convict when there were more prosecution arguments and when there were fewer defence arguments. Petty and Cacioppo (1984) asked college students whether they supported a somewhat unappealing proposal to introduce comprehensive exams for graduating seniors. This proposal was presented along with three or nine arguments, which themselves were either of high or low quality. When this was a low-involvement issue (the exams would affect students 10 years in the future) students were not influenced by
the quality of the arguments, and responded simply based on the number of arguments. However, when this was a high-involvement issue (current students would be affected) argument quality affected attitudes. Moreover there was an interaction between number of arguments and quality: Additional strong arguments increased support for the proposal, but additional weak arguments reduced support. Thus the Petty and Cacioppo study provided important boundary conditions on the notion that more evidence is better.

Argumentation

Closely related to the social cognition literature is research on argumentation, which has sought to identify structural characteristics that make arguments stronger or weaker. O'Keefe (1997, 1998) conducted meta-analyses to examine the effects of making arguments (or standpoints) more explicit (and hence longer). One could imagine that more explicit arguments would be more convincing because they provide more evidence, but on the other hand more explicit arguments may be less engaging, more open to scrutiny, and more likely to provoke disagreement. O'Keefe found that spelling out conclusions in detail, identifying sources, making implicit premises explicit, and providing precise quantitative information all made arguments seem stronger. Relatedly, O'Keefe (1999) conducted a meta-analysis to compare one-sided arguments, in which a positive argument is made for a position, to two-sided arguments, in which opposing views are acknowledge or even refuted. Two-sided arguments will naturally tend to be longer than one-sided arguments, but they run the risk of making counter-arguments more salient. O'Keefe found that two-sided arguments were seen as more convincing than one-sided arguments, and refutational two-sided arguments were seen as more convincing than nonrefutational two-sided arguments. Hence the general trend in the argumentation literature points to longer arguments being stronger (although note that merely adding anecdotal evidence to increase length does not necessarily increase strength, e.g., Hornikx & Hoeken, 2007).

Category-based induction

Research on inductive reasoning has examined the effects of varying the number of premises within a single argument (for a review see Heit, 2000). Nisbett, Krantz, Jepson, and Kunda (1983) systematically varied the given number of observations in an estimation task. For example, participants were told that 1, 3, or 20 obese members of the Barratos group had been observed, and were asked what proportion of all Barratos are obese. In general, inferences were stronger with increased sample size. (Strictly
speaking this is a sample size manipulation rather than an argument length manipulation, but our aim here is to be as inclusive as possible.)

Osherson, Smith, Wilkie, Lopez, and Shafir (1990) identified a phenomenon called *monotonicity* in which arguments with more premises are considered more plausible, for example, (2) is considered a stronger argument than (1).

$\begin{align*}
\text{Sparrows have sesamoid bones} & \quad \text{Eagles have sesamoid bones} \\
\text{All birds have sesamoid bones} \\
\text{Hawks have sesamoid bones} & \quad \text{Sparrows have sesamoid bones} & \quad \text{Eagles have sesamoid bones} \\
\text{All birds have sesamoid bones}
\end{align*}$

This monotonicity effect appears to be very robust: McDonald, Samuels, and Rispoli (1996) reported that, over a large set of arguments, there was a positive correlation between number of premises and judged argument strength. It also appears that the monotonicity effect is reasonably general: Lopez, Gelman, Gutheil, and Smith (1992) and Gutheil and Gelman (1997) reported some evidence for monotonicity effects in 9-year-old children. Interestingly, in a study of individual differences in adults, Feeney (2007) found that adult participants with higher intelligence showed greater monotonicity effects.

Note, however, that having more premises does not always lead to stronger arguments. Osherson et al. (1990) documented exceptions to the monotonicity effect, called *non-monotonicity* effects, in which longer arguments seem weaker, as in (3) and (4).

$\begin{align*}
\text{Crows secrete uric acid crystals} & \quad \text{Peacocks secrete uric acid crystals} \\
\text{All birds secrete uric acid crystals} \\
\text{Crows secrete uric acid crystals} & \quad \text{Peacocks secrete uric acid crystals} & \quad \text{Rabbits secrete uric acid crystals} \\
\text{All birds secrete uric acid crystals}
\end{align*}$

Here, adding premise information about rabbits that is not obviously relevant to the conclusion weakens the argument. In this case rabbits fall into a different superordinate category than the birds that are the focus of the other premises in the argument. However, Sloman (1993) showed
non-monotonicity effects even within a single superordinate category, for example, (6) was considered weaker than (5).

\[
\begin{align*}
\text{All crocodiles have acidic saliva} \quad &\quad (5) \\
\hline
\text{All alligators have acidic saliva} \\
\text{All crocodiles have acidic saliva} \quad &\quad (6) \\
\text{All king snakes have acidic saliva} \\
\hline
\text{All alligators have acidic saliva}
\end{align*}
\]

Here all of the animals are reptiles, but king snakes are so dissimilar to alligators that adding this second premise seems to weaken the argument.

Causal reasoning

In category-based induction tasks participants evaluate a conclusion based on a premise or set of premises. In causal reasoning tasks the task is often reversed: participants are given a conclusion and asked to evaluate a premise or set of premises in terms of how well they explain that conclusion. Borrowing an example from Read and Marcus-Newhall (1993), imagine that Cheryl is tired, is frequently nauseous, and is gaining weight—collectively, these observations are a conclusion to be explained. The shorter explanation that Cheryl is pregnant seems more compelling than the longer explanation that Cheryl has mononucleosis, has a stomach virus, and has stopped exercising. The idea of preferring shorter or simpler explanations follows from the principle of Occam’s razor that “entities should not be multiplied unnecessarily” (for a review see Lombrozo, 2007) and it also follows from Thagard’s (1989) connectionist network model of explanatory coherence. Both Read and Marcus-Newhall, and Lombrozo, provided experimental evidence that people favour shorter and simpler explanations over longer and more complicated explanations, with Lombrozo showing particularly strong results by controlling for prior probability of various explanations.

Whereas most findings on social cognition, argumentation, and category-based induction indicate that longer arguments are more convincing than shorter arguments, in research on causal reasoning it has been found that shorter explanations are more convincing than longer explanations. However, even in the social cognition and category-based induction literatures, there are some exceptions to the generalisation that longer arguments are more convincing. So the picture is somewhat unclear on whether longer arguments are better or worse than short arguments. Of course all of these studies varied in numerous ways, making it difficult to
draw a general conclusion about argument length. We next turn to our own research on argument length, which has found both positive and negative effects of argument length within a single experimental paradigm.

Rotello and Heit (2009)

In a recent study of category-based induction we manipulated the length of arguments while also manipulating whether they are logically valid. (Although an argument encompasses both premises and a conclusion, we varied length just in terms of number of premises.) The arguments had one, three, or five premises, and participants judged the plausibility of the conclusion. An example invalid argument, with three premises, is shown in (7).

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

An example valid argument, with five premises, is shown in (8).

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

For invalid arguments we found the usual monotonicity effect in category-based induction. Five-premise arguments were stronger than three-premise arguments, which in turn were stronger than one-premise arguments. However, the reverse pattern was found for valid arguments. One-premise arguments were the strongest and five-premise arguments were the weakest. This reverse pattern for valid arguments resembles the results from research on causal explanation. If the premises in an argument are conceived of an explanation for the conclusion, then once there is a sufficient explanation for the conclusion, adding additional premises seems to weaken the argument overall.

Although Rotello and Heit (2009) found consistent effects of argument length on judgements of plausibility, we also identified a related task in which participants showed little influence of argument length. When participants were asked to judge logical validity rather than plausibility they showed a greater influence of validity itself and reduced influence of
length. (Although length did not have a statistically significant influence on validity judgements, the trends were the same for validity as for plausibility, so it is impossible to rule out a small effect of length on validity judgements.) We explained this finding in terms of a two-dimensional account of reasoning that was implemented as a multidimensional signal detection model and fitted successfully to the data. One dimension corresponded to sensitivity to the rules of logic, and the other dimension used associative information such as argument length. The difference between the plausibility and validity judgements was accounted for in terms of different relative impacts of the two dimensions on the two kinds of judgement: Plausibility judgements were influenced about equally by the two dimensions, whereas validity judgements were influenced more by the logic-based process. The model is illustrated in Figure 1, where the key difference between plausibility and validity judgements is the slope of the decision bound.

THEORETICAL ACCOUNTS

The two-dimensional nature of the Rotello and Heit (2009) model of reasoning (see also Heit & Rotello, 2010; Heit, Rotello, & Hayes, 2012) is in accord with Petty and Cacioppo’s (1984) own two-process explanation, which referred to central and peripheral processing of information, and

![Figure 1. Schematic of Rotello and Heit’s (2009) model of inductive and deductive reasoning. The same evidence distributions (ellipses) are used for both types of judgement, but different decision bounds (shown as dashed lines). Three distributions are shown for invalid arguments; those to the right reflect invalid arguments with more premises. Two valid distributions are shown; those to the left reflect valid arguments with more premises. Reprinted with permission.](image-url)
itself is representative of theoretical work in social psychology. In addition, two-process accounts of reasoning have been very influential in cognitive psychology (Evans, 2008; Sloman, 1996; Stanovich, 2009). What all these accounts have in common is the notion that some processing, for example noticing the length of an argument, is automatic. Potentially people could avoid using this information—this is what Evans calls an “intervention” and Stanovich calls an “override”. However, what is then required is some other process to substitute for the automatic process. In the Rotello and Heit study some participants were explicitly instructed to pay attention to logical validity, and these participants were indeed able to substitute logical processing for processing based on more superficial information such as argument length. In the present experiments we focused on plausibility judgements, and whether people could avoid the use of argument length when judging plausibility. Under the assumption that noticing the length of an argument is automatic, and without explicit instructions to make judgements on some other basis, we predicted that it would be very difficult to avoid using argument length in judgements of plausibility. Put another way, argument length is intrinsic to judgements of plausibility.

Indeed, theoretical accounts of inductive reasoning are necessarily affected by length of argument (for reviews see Hayes, Heit, & Swendsen, 2010; Heit, 2008). Osherson et al. (1990) presented a model that accounted for plausibility judgements in terms of the maximum similarity between categories in the premises of an argument and the category in the conclusion of an argument (as well as members of the superordinate category which includes all of the other categories). Adding premises will generally increase the maximum level of similarity. For example, in going from (1) to (2), adding the premise about hawks will increase the similarity to hawk-like members of the bird category, in effect increasing the coverage of this category. The main exception is for non-monotonicity effects as in (4)—when the premise about rabbits is introduced, the superordinate category under consideration is all animals rather than birds, as in (3). The larger category is harder to cover. Hence coverage decreases from (3) to (4). However, in our own experiments the superordinate category is always mammal, so this exception does not apply.

Sloman’s (1993) model of inductive reasoning also leans on the concept of coverage, which is implemented via feature overlap in a connectionist network. The network is trained on each successive premise, and additional premises can only increase the level of activation in the network and the potential overlap with the conclusion. This model predicts that making an argument longer will always make its conclusion more plausible.

Finally, Bayesian models of inductive reasoning (Heit, 1998, 2000; Kemp & Tenenbaum, 2009; Tenenbaum & Griffiths, 2001) generally predict that
longer arguments will have their conclusions judged more plausible. These models operate in terms of a space of competing hypotheses. For example, in (1) and (2) the hypothesis that all birds have sesamoid bones is competing with other hypotheses, for example the hypothesis that only sparrows and eagles have sesamoid bones. Adding the premise that hawks have sesamoid bones rules out the latter hypothesis, so there is less competition for the hypothesis that all birds have sesamoid bones. In general, adding premises will rule out some prior hypotheses and strengthen the remaining ones. In Kemp and Tenenbaum’s terms, when adding premises some hypotheses from the prior distribution are no longer possible; in the posterior distribution for the hypothesis space probabilities are renormalised and the remaining hypotheses increase in probability. For Bayesian models the main exception would be if what Tenenbaum and Griffiths called a size principle is incorporated. Under this principle there is a bias to favour more restricted conclusions over broader conclusions, and this bias is increasingly manifest as more observations are made. The size principle could account for the non-monotonicity effects reported by Osherson et al. (1990) and Sloman (1993).

Finally, we note that all of these models (Heit, 1998, 2000; Kemp & Tenenbaum, 2009; Osherson et al., 2000; Sloman, 1993; Tenenbaum & Griffiths, 2001) are challenged by Rotello and Heit’s (2009) finding that logically valid arguments are judged to have less-plausible conclusions when the arguments are longer. With the exception of Sloman’s model, each model predicts that logically valid arguments are maximally strong. None of these models predicts that adding premises to a logically valid argument will weaken it. As noted by Lombrozo (2007) it is possible to develop a Bayesian account of simplicity, but even the Bayesian models of inductive reasoning predict that adding premises to a perfectly valid argument will not reduce the probability of the conclusion.

OVERVIEW

In three experiments we investigated the effects of argument length on judgements of plausibility. The baseline condition in Experiments 1a and 1b served as replications of Rotello and Heit (2009). Both experiments also included a new condition, forewarning, in which participants were instructed to try to avoid using argument length in making plausibility judgements. Experiment 1b used a somewhat stronger forewarning manipulation than Experiment 1a, and also included an anti-forewarning condition in which participants were encouraged to use argument length. In general, because noticing argument length is automatic, because using argument length is an intrinsic part of judging plausibility, and because participants were not given any alternative means of making the
judgements, we predicted that it would be very difficult for participants to control their own use of argument length. Experiment 2 added a within-participant manipulation of speaker reliability to the basic manipulation of argument length. We expected that, overall, arguments from an unreliable speaker would be rejected more than arguments from a reliable speaker. In addition we examined whether use of logical information and superficial information would vary for reliable versus unreliable speakers. One possibility is that participants would attend to arguments from one type of speaker more than the other, showing greater sensitivity to validity as well as length. Another possibility, by analogy to Petty and Cacioppo (1984) is that validity would matter more for one speaker and length would matter more for the other speaker.

Although our focus was on whether the effects of argument length changed as a function of instructions, to get a better understanding of the results we applied the Rotello and Heit (2009) model to the data. Because participants in each group were asked to make induction judgements we expected that the same decision rule would apply, and therefore that the slope of the decision bound would not differ greatly across conditions. However, fitting the model could help explain other changes in results across conditions. For example, potential differences in encoding arguments could be manifested as different locations of the distributions of the arguments.

EXPERIMENT 1A

This experiment, like Rotello and Heit (2009), examined the effects of length and validity on judgements of inductive plausibility. The length manipulation involved varying the number of premises in the argument. The validity manipulation involved presenting either invalid arguments or valid arguments that maintained either an identity or class inclusion relation between premise and conclusion categories. (See Osherson et al., 1990, for a discussion of the theoretical importance of identity arguments; and see Sloman, 1993, 1998, for further discussion of relations between identity and inclusion arguments and the theoretical implications of studying both kinds of arguments.) The key difference from Rotello and Heit was the introduction of a forewarning condition.

Method

Participants. A total of 88 University of California, Merced undergraduate students, from a variety of majors, were paid to participate. They were randomly assigned to one of two conditions: control \((n = 43)\) or forewarning \((n = 45)\).
Stimuli. There were 120 questions,\(^1\) comprising arguments about the following kinds of mammals: bears, cats, cows, dogs, goats, horses, lions, mice, rabbits, and sheep. An example invalid argument is:

\[
\begin{align*}
\text{Horses have Property X} \\
\text{Mice have Property X} \\
\text{Sheep have Property X} \\
\hline
\text{Cows have Property X}
\end{align*}
\]

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 60 remaining arguments were deductively valid. Of these, 45 were identity arguments in which the conclusion category was identical to one of the premise categories. An example valid, identity argument is:

\[
\begin{align*}
\text{Horses have Property X} \\
\text{Mice have Property X} \\
\text{Sheep have Property X} \\
\text{Rabbits have Property X} \\
\text{Cats have Property X} \\
\hline
\text{Rabbits have Property X}
\end{align*}
\]

The remaining 15 valid arguments were inclusion arguments in which the conclusion category was included in one of the premise categories.\(^2\) 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.

\[
\begin{align*}
\text{Mammals have Property X} \\
\hline
\text{Horses have Property X}
\end{align*}
\]

\(^1\)We acknowledge that 120 questions is more than average for experiments on syllogisms. Collecting so much data per participant has methodological advantages (e.g., providing sufficient data for a complex design without the need for an unfeasible number of participants) and potential disadvantages (e.g., diminishing quality of data over the course of an experimental session). Although we did not address this latter issue systematically, performance appeared to be similar earlier and later within each experiment. For example, in Experiment 1a, \(d^\prime\) (sensitivity to valid versus invalid arguments) was 1.7 for the first 60 trials and 1.8 for the last 60 trials.

\(^2\)Strictly speaking, inclusion arguments are enthymemes, because they rely on a hidden premise, such as that all cows are mammals (Calvillo & Revlin, 2005). For simplicity we refer to both the identity and inclusion arguments as valid.
Procedure. Each experiment was run using a program on a computer; each participant took part individually. At the beginning of the experiment the computer screen displayed instructions on the definition of strong arguments. Specifically, following Rips (2001), participants 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”. In the forewarning condition the following additional instructions were displayed:

You will see questions that vary in length. Sometimes you will only get one sentence of information, but other times you will get several sentences of information. Note that the length of argument is irrelevant to whether it is a strong argument. Sometimes a short argument, with little information, is very convincing, and sometimes a long argument, with a lot of information, is still not very convincing at all. So in making your judgements about whether the following arguments are strong, please try to IGNORE the length of the question.

Following the instructions, 120 arguments were presented one at a time, in a different random order for each participant. That is, different kinds of arguments were intermixed, and participants were given the same instructions for all of the arguments. Participants were 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”. Each binary decision was followed with a confidence rating on a 1–5 scale; higher numbers indicated greater confidence.

Results
Nine participants were excluded from the analyses because they gave the same response for virtually every question, made more “strong” responses to invalid than valid arguments, or had very low performance ($d’ \leq .50$ discriminating valid from invalid arguments). This eliminated three participants from the forewarning condition and six from the control condition.

Response rates. For an overview we first considered the probability that participants endorsed the problem conclusion, as a function of validity, number of premises, and condition. These data are shown in Table 1. For both the control and forewarning conditions, increasing the length of invalid arguments led to a higher rate of endorsement. It did not appear that the forewarning instructions were effective in preventing the effect of argument length. Looking at invalid arguments overall, participants in the forewarning condition were somewhat less likely to respond positively compared to the control condition. Next, looking at valid arguments, there was a
tendency for longer arguments to be rejected more often than shorter arguments, in both the control and forewarning conditions.

These results were subjected to separate ANOVAs for invalid and valid arguments. For invalid arguments positive responses increased with the number of premises, $F(2, 154) = 19.04, p < .001, \text{MSE} = .03, \eta^2 = .19$. The effect of condition did not reach the level of statistical significance, $F(1, 77) = 2.52, p = .12, \text{MSE} = .22, \eta^2 = .03$, but condition interacted with the number of premises, $F(2, 154) = 3.56, p < .05, \text{MSE} = .03, \eta^2 = .04$: The effect of argument length was slightly stronger in the forewarning condition. Each condition taken alone showed a significant effect of number of premises on the responses to invalid arguments: control, $F(2, 72) = 6.02, p < .01, \text{MSE} = .03$, Cohen’s $f = .30$; forewarned, $F(2, 82) = 17.18, p < .001, \text{MSE} = .03$, Cohen’s $f = .51$.

For valid arguments positive responses decreased slightly with the number of premises, $F(2,154) = 10.84, p < .001, \text{MSE} = .00, \eta^2 = .12$. There was no effect of condition, $F(1,77) = 1.21, p = .27, \text{MSE} = .02, \eta^2 = .02$, and no interaction of condition with number of premises, $F(2,154) = 1.43, p = .24, \text{MSE} = .00, \eta^2 = .02$.

**Signal detection analyses.** Following the techniques used in previous studies (Dube, Rotello, & Heit, 2010, 2011; Heit & Rotello, 2005, 2008, 2010; Rotello & Heit, 2009), we treated the reasoning task as a signal detection task in which the goal was to discriminate strong arguments from weak arguments. Simple accuracy measures such as the difference between correct and error response rates, or the signal detection based $d’$, are often confounded with participants’ overall tendency to accept a conclusion (see

<table>
<thead>
<tr>
<th>Expt. 1a</th>
<th>Invalid problems</th>
<th></th>
<th></th>
<th></th>
<th>Valid problems</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 premise</td>
<td>3 premises</td>
<td>5 premises</td>
<td>1 premise</td>
<td>3 premises</td>
<td>5 premises</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>.35 (.06)</td>
<td>.36 (.05)</td>
<td>.48 (.05)</td>
<td>.98 (.00)</td>
<td>.96 (.01)</td>
<td>.95 (.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forewarning</td>
<td>.18 (.04)</td>
<td>.32 (.04)</td>
<td>.39 (.05)</td>
<td>.97 (.01)</td>
<td>.97 (.02)</td>
<td>.93 (.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expt. 1b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>.09 (.02)</td>
<td>.14 (.03)</td>
<td>.20 (.04)</td>
<td>.93 (.01)</td>
<td>.87 (.02)</td>
<td>.86 (.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forewarning</td>
<td>.07 (.02)</td>
<td>.14 (.04)</td>
<td>.15 (.04)</td>
<td>.89 (.03)</td>
<td>.85 (.02)</td>
<td>.83 (.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anti-forewarning</td>
<td>.07 (.02)</td>
<td>.11 (.02)</td>
<td>.24 (.05)</td>
<td>.93 (.01)</td>
<td>.84 (.02)</td>
<td>.85 (.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expt. 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliable</td>
<td>.24 (.04)</td>
<td>.38 (.05)</td>
<td>.54 (.05)</td>
<td>.95 (.01)</td>
<td>.85 (.03)</td>
<td>.89 (.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unreliable</td>
<td>.09 (.03)</td>
<td>.17 (.03)</td>
<td>.26 (.05)</td>
<td>.80 (.04)</td>
<td>.75 (.04)</td>
<td>.77 (.03)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard errors are shown in parentheses.
Dube et al., 2010; Rotello, Masson, & Verde, 2008). To allow clear evaluation of accuracy difference across conditions that are not confounded with response bias differences we plotted receiver operating characteristic (ROC) curves using the strong/not-strong responses as well as the 1–5 confidence ratings (for details see Macmillan & Creelman, 2005).

Figure 2 shows ROC curves for arguments of length 1, 3, and 5 in the control condition (left panel) and forewarning condition (right panel). The x-axis refers to the probability of responding “strong” on invalid items and the y-axis refers to the probability of responding “strong” on valid items. Each curve plots these probabilities for various levels of confidence (response bias), with the left-most point on each curve reflecting the “sure valid” response rates. Thus the points along each curve reflect the same accuracy level but different response biases; more liberal response tendencies increase both the correct and incorrect response rates yielding points towards the upper-right in the space. Curves that fall higher in the space (towards the upper-left corner) represent a greater degree of accuracy in discriminating valid from invalid arguments because in that region the correct response rate is high relative to the error rate. Accordingly, an appropriate measure of performance is the area under the ROC, $A_z$ (Swets, 1986), which ranges from 0.5 (for chance performance) to 1.0 (perfect accuracy).

It is evident that, in both conditions, longer arguments are associated with a lower degree of accuracy. For five-premise arguments, participants were most likely to respond positively to invalid arguments and most likely to respond negatively to valid arguments. In comparison, for one-premise arguments...
arguments, participants were less likely to respond positively to invalid arguments and less likely to respond negatively to valid arguments. Hence these ROC curves illustrate the effect of argument length. Indeed, the effect of argument length appears stronger in the forewarning condition (right panel), as there is more separation between the curve for length 1 and the curves for length 3 and 5. For both the control and the forewarning conditions there is greater area under the length 1 curve ($A_z = .92$ and .96, respectively) than the length 5 curve ($A_z = .86$ and .87), reflecting greater ability to discriminate strong from weak arguments for arguments with one premise. These accuracy differences were significant in both conditions: For the control group, $z = 7.10$, $p < .001$, using the methods described by Hanley and McNeil (1983) for comparing dependent ROCs; in the forewarning condition, $z = 12.57$, $p < .001$.

**Modelling.** We adopted Rotello and Heit’s (2009) approach to thinking about these data, in terms of viewing argument strength as varying along two underlying dimensions, associative strength (instantiated as argument length) and apparent validity. Longer invalid arguments were allowed to have greater strength on the associative dimension, but valid arguments with more premises were allowed to have lower strength on that dimension. Valid arguments have more strength on the dimension of validity. Because all participants made induction judgements we assumed that the decision axis was the same across conditions, reflecting the same relative weighting of the two types of information. However, we did consider the possibility that participants in the forewarning condition would pay more attention to argument length during the encoding of the stimulus, which would imply that the means of the distributions of argument strengths might differ across conditions.

To evaluate the model we ran a large number of Monte Carlo simulations in which the parameters were varied systematically. For each set of parameter values we sampled 2000 simulated problem strengths from each distribution; these strengths were compared to the decision bound, and values falling in the “valid” response region were counted as resulting in a “valid” decision. This process yielded a simulated hit and false alarm rate for each problem type for that set of parameter values; the simulated ROC was mapped out by systematically varying the $y$-intercept of the decision bound from high (yielding conservative responding) to low (resulting in more liberal responding). The best-fitting set of parameter values was chosen from those that produced both a relatively small mean-squared error of prediction and a predicted area under the simulated ROC that fell within the 95% confidence bands for the observed ROC. However, these parameters should be considered illustrative only, as the entire parameter space has not been searched exhaustively.
Estimated parameter values shown in Table 2. The only difference between the control and forewarning conditions is the location of argument distributions for one-premise valid problems, for control versus forewarning condition, reflecting the somewhat higher level of accuracy on those items, and the greater effect of argument length in the forewarning condition. There is nothing in the estimated parameter values to indicate that participants in the forewarning condition were less sensitive to argument length than participants in the control condition.

The model predictions are shown in Figure 3, based on the estimated parameters. The key aspects of the observed ROCs are captured in this simulation. Most important, there is an effect of number of premises in both conditions such that accuracy falls as number of premises increases, because there are more positive responses to invalid arguments and fewer positive responses to invalid arguments. The only difference between the two conditions is that in the forewarning condition, the model accounts for the

### Table 2
Parameter values for the two-dimensional model as applied to each experiment

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$1a$</th>
<th>$1b$</th>
<th>$2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_x = \text{mean of valid 1-premise arguments on } x$-axis</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Variance of $d_x$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$d_y = \text{mean of valid 1-premise arguments on } y$-axis</td>
<td>2.1</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Variance of $d_y$</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Location of valid 3-, 5-premise arguments on $x$-axis</td>
<td>$-0.3, -0.3$</td>
<td>$-0.5, -0.5$</td>
<td>$-1.0, -1.0$</td>
</tr>
<tr>
<td>Change in $d_x$ for valid 3-, 5-premise arguments</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>Location of invalid 3-, 5-premise arguments on $x$-axis</td>
<td>0, 0.2</td>
<td>0.1, 0.2</td>
<td>0.2, 0.4</td>
</tr>
<tr>
<td>Control or reliable condition slope</td>
<td>$-0.3$</td>
<td>$-0.3$</td>
<td>$-0.3$</td>
</tr>
<tr>
<td>Forewarning or unreliable condition slope</td>
<td>$-0.3$</td>
<td>$-0.3$</td>
<td>$-0.5$</td>
</tr>
<tr>
<td>Anti-forewarning condition slope</td>
<td>NA</td>
<td>$-0.3$</td>
<td>NA</td>
</tr>
<tr>
<td>Change in $d_x$ for 1-premise valid arguments in forewarning or unreliable condition</td>
<td>0.2</td>
<td>0</td>
<td>$-0.2$</td>
</tr>
<tr>
<td>Change in $d_y$ for 1-premise valid arguments in forewarning unreliable condition</td>
<td>0.7</td>
<td>0</td>
<td>$-0.3$</td>
</tr>
<tr>
<td>Covariance of $x$ and $y$ for valid arguments</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>Covariance of $x$ and $y$ for invalid arguments</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

The distribution of invalid low-similarity arguments was located at $(0, 0)$ with a variance of 1 on each dimension and a small covariance.
even higher level of accuracy on 1 premise items, which is part of the greater effect of number of premises in the forewarning condition.

Discussion

The control condition of Experiment 1a replicated Rotello and Heit (2009) in terms of showing that invalid arguments are considered stronger when they are longer, and valid arguments are considered weaker when they are longer. These results were shown using conventional analyses in terms of response proportions as well as analyses in terms of area under ROC curves plotted using confidence ratings.

The results of the forewarning condition were similar to the control condition. If anything, there was a slightly stronger effect of argument length in the forewarning condition. Clearly, warning participants not to use argument length did not discourage them from doing so. It appears that argument length is such a compelling attribute, both in terms of making invalid arguments seem strong and in terms of making valid arguments seem weak, that it is very difficult to ignore, even whether there is an explicit request from the experimenter to do so.

In Experiment 1b we attempted a stronger forewarning manipulation, and we examined the issue of whether the use of argument length can be influenced in another way, by encouraging a group of participants to try to show an argument length effect.
In this experiment we strengthened the forewarning manipulation somewhat by presenting the instructions to ignore argument length twice, on paper before the experiment as well as on the computer during the experiment. The instructions were provided with the experimenter present. We also included an anti-forewarning condition, in which participants were told that argument length is important and were encouraged to consider argument length. This experiment was patterned after a study by Heit, Brockdorff, and Lamberts (2004), who found that a memory illusion (false alarms to semantic associates of studied words) could not be reduced by a warning, but that participants could be encouraged to exaggerate the illusion (make more false alarms to semantic associates). In the anti-forewarning condition, if participants are responding based on their inferences about the experimenter’s intent, they should respond even more strongly to argument length.

Method

The method was the same as Experiment 1a except for the following: 135 individuals participated; control (n = 45), forewarning (n = 44), anti-forewarning (n = 45).

The control condition was the same as in Experiment 1a. Participants in the forewarning condition received the additional instructions, to ignore argument length, twice. They first received these on a sheet of paper, before the computer-based part of the experiment began. The experimenter watched the participant read these instructions and asked whether he or she had any questions. Once the computer-based experiment itself began, the forewarning instructions were displayed on the computer screen as in Experiment 1a. Participants in the anti-forewarning conditions also received additional instructions, twice. However, they were instructed to consider argument length when evaluating arguments. In particular, they were told:

You will see questions that vary in length. Sometimes you will only get one sentence of information, but other times you will get several sentences of information. Note that the length of argument is important to whether it is a strong argument. Often a short argument, with little information, is not very convincing at all, and often a long argument, with a lot of information, is very convincing. So in making your judgements about whether the following arguments are strong, please try to CONSIDER the length of the question.

Results

A total of 14 participants were excluded from the data analyses (4, 3, and 7 in the control, forewarning, and anti-forewarning conditions, respectively), according to the same criteria as Experiment 1a.
Response rates. For an overview we first considered the probability that participants endorsed the problem conclusion, as a function of validity, number of premises, and condition. These data are shown in Table 1. For the control, forewarning, and anti-forewarning conditions, increasing the length of invalid arguments led to a higher rate of endorsement. It did not appear that the forewarning instructions were effective in preventing the effect of argument length—the results are close to the control condition. Likewise, the anti-forewarning instructions did not seem to have much effect. Next, looking at valid arguments, there was a tendency for longer arguments to be rejected more often than shorter arguments, in all three conditions. Overall the pattern of results is similar to Experiment 1a, except that there is even less difference between the control and forewarning conditions, and the overall rate of responding positively is somewhat lower. For this latter result we do not offer an interpretation, but note that the average rate of positive responding is not crucial to the issues investigated here.

The results were subjected to ANOVA. On invalid arguments there was a significant effect of number of premises, $F(2, 234) = 26.88$, $p < .001$, $MSE = .02$, $\eta^2 = .19$, with more positive responses to longer arguments. The effect of condition did not reach the level of statistical significance, $F(2, 117) < 1$. The interaction of number of premises with condition did not quite reach the level of significance, $F(3, 184) = 2.44$, $p = .06$, $MSE = .03$, $\eta^2 = .04$.

On valid arguments there was a significant effect of number of premises, $F(2, 234) = 28.20$, $p < .001$, $MSE = .18$, $\eta^2 = .19$, with more positive responses to shorter arguments. There was no effect of condition, $F(2, 117) < 1$, and no interaction of condition with number of premises, $F(2, 234) < 1$.

Signal detection analyses. Figure 4 shows ROC curves for arguments of length 1, 3, and 5 in the control condition (left panel), the forewarning condition (right panel), and the anti-forewarning condition (bottom panel). As in Experiment 1a, in all conditions, longer arguments are associated with a lower degree of accuracy: Those ROCs fall lower in the space. There are more positive responses to longer invalid arguments, and fewer positive responses to longer valid arguments. Given the visual similarity of the ROCs produced in each of our three conditions, and the ANOVA results, we did not expect to find differences in reasoning accuracy across conditions.

As in Experiment 1a we compared the length 1 versus length 5 curve in each condition. For the control, forewarning, and anti-forewarning conditions there was greater area under the length 1 curve than the length 5 curve, reflecting greater ability to discriminate strong from weak arguments for arguments with one premise. Hence there were robust effects of argument length in each condition. Using Hanley and McNeil’s (1983)
method, the highly-significant $z$-statistics were 11.31, 9.43, and 11.38 for the control, forewarning, and anti-forewarning conditions.

**Modelling.** Using the same strategy as in Experiment 1a we fitted Rotello and Heit’s (2009) two-dimensional model to these data. The resulting parameter values are shown in Table 2, and the simulated ROCs are shown in Figure 5. Because the overall level of positive response was lower in Experiment 1b, and there was less difference between the control and forewarning conditions, some parameter values differed slightly from
Experiment 1a. Most important, exactly the same parameter values were used to model the three conditions in Experiment 1b. Therefore the modelling provides no evidence that participants’ encoding strategies differed across conditions (the distributions have the same locations in all conditions), and also no evidence that participants’ response strategies differed across conditions (the same decision bound was used in all cases). This modelling work suggests that participants in Experiment 1b were unable or unwilling to avoid or exaggerate the influence of argument length, even when asked.

Discussion
As in Experiment 1a there was an overall effect of argument length, with invalid arguments seeming stronger as they get longer, and valid arguments...
seeming weaker as they get longer. Despite the stronger, in-person, request in this experiment, again the forewarning instructions were ineffective in reducing the impact of argument length. Moreover, in the anti-forewarning condition encouraging participants to use argument length also had little effect, as if participants were already making maximum use of argument length in the control condition. Unlike the Rotello and Heit (2009) study, which showed robust effects of instructions (to judge validity or plausibility), here participants’ judgements were not influenced by instructions. We conclude that argument length has intrinsic effects on inductive reasoning.

Other than a difference in the average rate of positive responding, there were minimal differences in the results of the two experiments and accordingly in the estimated parameter values for the fitted model. None of these differences suggests that experimenter instructions affected participants’ use of argument length as a basis for judging plausibility. For both experiments the model from Rotello and Heit (2009) gave a satisfactory account of the detail results in terms of ROC curves.

Experiments 1a and 1b may have implications for another account of inductive reasoning, relevance theory (Medin, Coley, Storms, & Hayes, 2003), which conceives of experiments on reasoning within a social context, in which the experimenter is expected to follow Gricean principles of informativeness and relevance (Grice, 1975). Simply put, making an argument longer might reveal what the experimenter intends to convey about the conclusion. Although Medin et al. did vary the length of arguments, they did not explicitly address the possibility that length itself is a cue to participants. If responding to the length of an argument reflects an inference about the experimenter’s communicative intent, then an explicit, even in-person, instruction from the experimenter to ignore length should at least reduce the influence of length. Instead we found no influence of this instruction. We would not minimise the importance of the social context of the experiment, or Gricean inferences, when participants make judge inductive plausibility, and we find Medin et al.’s (2003) relevance theory of induction to be appealing in general (but see Heit & Feeney, 2005; Heit, Hahn, & Feeney, 2005). However, in the present case the effects of argument length seem to be due to the core nature of plausibility judgement, rather than the participant’s perception of the experimenter’s communicative intent.

**EXPERIMENT 2**

In this experiment we manipulated speaker characteristics: Each argument was cued as coming from a reliable or an unreliable speaker. Overall we
expected that arguments coming from a reliable speaker would be considered more plausible than arguments coming from an unreliable speaker. This prediction simply assumes that participants were influenced by the reliability cue in the direction of the cue. Another way of stating this prediction is that reliability information is a kind of *ad hominem* comment that is treated as a reasonable part of an argument (cf., Van Eemeren, Garssen, & Meuffels, 2012 this issue).

In addition, we examined how this manipulation interacted with sensitivity to validity. It is possible that arguments from unreliable speakers would receive greater scrutiny, leading to greater sensitivity to validity. Such a result would be consistent with what is commonly reported as a belief bias effect, namely that there is greater sensitivity to validity for unbelievable conclusions than for believable conclusions (e.g., Thompson & Evans, 2012 this issue). However, Dube et al. (2010, 2011) showed that this commonly reported result is a measurement artefact, and reported no true difference in sensitivity. Another possibility is that arguments from unreliable speakers are more likely to be rejected without consideration of their content. By this account, participants would be less sensitive to validity for unreliable speakers. Indeed, Hahn, Harris, and Corner (2009) found greater sensitivity to evidence strength for reliable sources than for unreliable sources, in an experiment on argument from ignorance. (For a review of other studies with similar results see Pornpitakpan, 2004.) Another reason to make this same prediction is in terms of assimilation and contrast effects (Bohner, Ruder, & Erb, 2002). Valid arguments coming from reliable speakers may seem particularly strong, due to an assimilation effect, but invalid arguments from reliable speakers may seem particularly weak due to a contrast effect. To the extent that any such effects would be weaker for unreliable speakers, we would expect greater sensitivity to validity for arguments coming from reliable speakers.

Finally, with regard to sensitivity to argument length, a number of predictions are possible. Given the findings from Experiments 1a and 1b that participants have great difficulty not attending to argument length, one would expect that sensitivity to validity and sensitivity to length would go hand in hand. If arguments from unreliable speakers get more scrutiny, then participants should be more sensitive to both validity and length for these arguments. On the other hand, if arguments from unreliable speakers tend to be dismissed without consideration of their content, then participants should be more sensitive to both validity and length for arguments for reliable speakers. Another possibility is that sensitivity to length is affected differently from sensitivity to validity by speaker reliability. For example, paying attention to relatively superficial cues such as argument length may be more congruent with processing messages from unreliable speakers than from reliable speakers. Relatedly, Petty and Cacioppo (1984) found that
superficial features had a greater effect for low-involvement issues than high-involvement issues.

**Method**

The method was the same as the control condition of Experiments 1a and 1b except for the following: 50 individuals participated. After the main set of instructions, participants were given additional information:

Note that each argument comes from one of two sources, Jane or Mary. Jane is a reliable source. 65% of her conclusions are strong. Mary is an unreliable source. 35% of her conclusions are strong. When you judge whether you think each conclusion is strong or weak, you should consider the reliability of the source.

As each argument was presented for evaluation, at the top of the computer screen it was noted either “Jane says (reliable, 65% strong)” or “Mary says (unreliable, 35% strong)”.

Because speaker reliability was a within-participant manipulation we increased the total number of arguments from 120 to 144, to increase the amount of data per participant per condition. The arguments were presented in a random order, intermixing arguments from the two speakers. Half of the arguments were said to come from the reliable speaker and half were said to come from the unreliable speaker. Note that this was a completely uninformative cue: For each speaker 50% of arguments were valid. In other respects the arguments maintained the same proportions as in the other experiment. For valid arguments there was a 3:1 ratio of identity to inclusion arguments. Overall, one third of the arguments had one premise, one third had three premises, and one third had five premises.

**Results**

Six participants were excluded from the data analyses according to the same criteria as Experiments 1a and 1b.

*Response rates.* For an overview we first considered the probability that participants endorsed the problem conclusion, as a function of validity, number of premises, and speaker reliability. These data are shown in Table 1. The manipulation clearly had a strong effect: Arguments from the reliable speaker were more likely to be accepted than arguments from the unreliable speaker. For both reliable and unreliable speakers, increasing the length of invalid arguments led to a higher rate of endorsement. Next, looking at valid arguments, there was a similar pattern for reliable and unreliable speakers: Increasing argument length made three-premise arguments weaker than one-premise arguments, but five-premise arguments were slightly stronger.
than three-premise arguments. In general these patterns seem to replicate Experiments 1a and 1b.

The results were subjected to ANOVA. On invalid arguments there was a significant effect of number of premises, $F(2, 84) = 25.61$, $p < .001$, $MSE = .02$, $\eta^2 = .38$, with more positive responses to longer arguments. There was a main effect of speaker reliability, $F(1, 42) = 39.79$, $p < .001$, $MSE = .074$, $\eta^2 = .49$, such that arguments from reliable speakers were accepted more than arguments from unreliable speakers. The interaction between number of premises and speaker reliability was also statistically significant, $F(2, 84) = 6.21$, $p < .01$, $MSE = .014$, $\eta^2 = .13$. There was a greater effect of number of premises for reliable speakers than for unreliable speakers.

On valid arguments there was a significant effect of number of premises, $F(2, 84) = 11.63$, $p < .001$, $MSE = .009$, $\eta^2 = .22$. There was a main effect of speaker reliability, $F(1, 42) = 15.86$, $p < .001$, $MSE = .062$, $\eta^2 = .27$, such that arguments from reliable speakers were accepted more than arguments from unreliable speakers. The interaction between number of premises and speaker reliability did not reach the level of statistical significance, $F(2, 84) = 1.48$, $p = .23$.

**Signal detection analyses.** Figure 6 shows ROC curves for arguments of length 1, 3, and 5 for reliable (left panel), and unreliable (right panel) speakers. As in Experiments 1a and 1b, in both conditions longer arguments are associated with a lower degree of accuracy. There are more positive

![Figure 6](image-url)
responses to longer invalid arguments, and fewer positive responses to longer valid arguments.

When comparing the curves for reliable versus unreliable speakers, three trends are evident. First, it appears that participants were more accurate at distinguishing valid from invalid arguments for reliable speakers than for unreliable speakers, at least for arguments of length 1. The ROC for reliable speakers falls higher in space than that for unreliable speakers. We again used Hanley and McNeil’s (1983) method to compare areas under these dependent curves. For length 1 arguments the difference between reliable and unreliable speakers was significant, $z = 6.47$; for lengths 3 and 5 the differences were not significant ($z = 1.19$ and 1.12, respectively).

Second, participants responded more liberally to the reliable sources; the operating points are shifted to the right along the ROCs. We estimated $c_a$ for each participant, assuming a $z$ROC slope of 0.61 (obtained from the group ROCs), and subjected those bias estimates to ANOVA. Participants set more liberal criteria for longer arguments, $F(2, 84) = 11.54, p < .001$, $MSE = .079, \eta^2 = .22$, and for the reliable source, $F(1, 42) = 46.12, p < .001$, $MSE = .148, \eta^2 = .52$; the two factors did not interact $F(2, 84) = 1.80, p = .17$.

The third trend was that it appears that participants’ responses were more affected by argument length for reliable speakers than for unreliable speakers. This result is evident from the separation of the curves. The spread of curves from length 1 to length 5 is greater for reliable speakers than for unreliable speakers. For both conditions, the difference between length 1 and length 5 is statistically significant (reliable, $z = 13.79$; unreliable, $z = 12.22$).

Modelling. Again we fitted Rotello and Heit’s (2009) two-dimensional model to these data. The resulting parameter values are shown in Table 2, and the simulated ROCs are shown in Figure 7. With regard to Figure 7, again the model predictions fall in a satisfactory way within the 95% confidence intervals of the data. Crucially, parameter values differed by condition (see Table 2). The locations of the valid distributions were shifted down and to the left for the unreliable condition relative to the valid condition. In other words, for the unreliable condition compared to the reliable condition, valid items were closer to invalid items. This finding corresponds to lower sensitivity to both validity (y axis) and length (x axis) for the unreliable condition. In addition, the slope of the decision boundary is somewhat steeper for the unreliable condition than for the reliable condition, suggesting somewhat greater use of length relative to validity, in the unreliable condition compared to the reliable condition.
The analyses for Experiment 2 converge well. In comparison to Experiments 1a and 1b, Experiment 2 showed a more dramatic effect of the instructional manipulation. Compared to reliable arguments, evaluation of arguments from unreliable speakers was more likely to lead to rejection, less sensitive to argument validity, and less sensitive to argument length. It appears that participants were more engaged with arguments from the reliable speaker, taking account of both validity and number of premises. Arguments from the unreliable speaker were more likely to be rejected out of hand, consequently taking less account of both validity and number of premises. This finding is consistent with previous results from the argumentation and social cognition literatures (Bohner et al., 2002; Hahn et al., 2009; Pornpitakpan, 2004). This finding is the opposite of what is commonly, but incorrectly, reported as a belief bias effect, that people are less sensitive to validity for believable conclusions compared to unbelievable conclusions (Dube et al., 2010, 2011).

Figure 7. Predicted ROCs and observed 95% confidence bands for Experiment 2. Upper row: reliable source condition; lower row: unreliable source condition. (The colour version of this figure is available in the online article.)
Our modelling also suggested that, for the unreliable speaker, there was proportionately greater use of more superficial information, related to length, rather than validity. However, we interpret this result with caution because there was less use overall of both validity and length in the unreliable condition.

In other respects the results from Experiment 2 resembled those from Experiments 1a and 1b. Most importantly, the same effects of argument length appeared in both conditions. For invalid arguments longer arguments were considered stronger. For valid arguments longer arguments were generally weaker (comparing length 1 arguments to length 3 and 5 arguments). Hence the pervasive effect of argument length was pervasive even across this manipulation.

Finally, we briefly turn to implications of our reliability manipulation for the argumentation literature. As reviewed by Hahn and Oaksford (2012) there have been multiple theoretical approaches to evaluation of argument quality. One classic approach is the epistemic approach, being concerned with matters of truth. Possibly listeners could infer that unreliable speakers have poorer access to the truth. Bayesian approaches to argumentation (e.g., Hahn et al., 2009; Hahn & Oaksford, 2007) treat evaluation as a matter of evidence and belief. Listeners’ prior beliefs could embody the assumption that unreliable speakers’ arguments are less likely to be correct, and these beliefs could be revised upward or downward depending on the content of arguments. An alternative approach is the pragma-dialectical approach (e.g., Rips, 1998; Van Eemeren et al., 2012 this issue; Van Eemeren & Grootendorst, 2004), which is concerned with procedural aspects of argumentation. Listeners could infer that unreliable speakers are less competent at taking the steps necessary to make a sound argument, and this inference could lead to more negative evaluations of arguments from unreliable speakers. All of these accounts could readily explain the main effect of reliability, but each account would need to be expressed in greater detail to address the other results of Experiment 2, such as greater sensitivity for both length and validity for arguments coming from reliable speakers.

GENERAL DISCUSSION

Using traditional analyses, ROC analyses, and modelling techniques, we have converged on a consistent pattern of results. Replicating Rotello and Heit (2009), when an argument is invalid, making it longer makes it seem more plausible, and when an argument is valid, making it longer makes it seem less plausible. The effects of argument length were so pervasive that warning participants not to be influenced by length did not reduce its influence. Participants were unwilling or unable to avoid the influence of argument length, even when asked. Moreover, participants had so little
control over the influence of argument that they did not show greater effects of argument length when encouraged to do so. Likewise, a manipulation of speaker reliability had substantial effects on argument evaluation, but these effects of argument length were robust for both reliable and unreliable speakers.

Looking across several areas of reasoning research there have been results pointing to increased strength with increased length (Burnstein et al., 1973; Calder et al., 1974; Cook, 1969; Feeney, 2007; Gutheil & Gelman, 1997; Lopez et al., 1992; Nisbett et al., 1983; O'Keefe, 1997, 1998, 1999; Osherson et al., 1990; Petty & Cacioppo, 1984; Sloman, 1993) as well as decreased strength with increased length (Lombrozo, 2007; Lopez et al., 1992; Osherson et al., 1990; Petty & Cacioppo, 1984; Read & Marcus-Newhall, 1993; Sloman, 1993). Our own contribution has been to document both effects of argument length, within the same task, while holding most things constant and varying only whether the argument is logically valid.

We do not presume to have developed a complete account of when making arguments longer will make them stronger or weaker for different reasoning tasks. However, in addressing this interesting issue, we do think it will be valuable for future theoretical accounts of reasoning to cross task boundaries, e.g., to explain results from argumentation, attitude change, causal reasoning, and category-based induction paradigms, rather than treating these as separate issues. Perhaps by treating each of these as signal detection tasks, and applying multidimensional models of reasoning, some progress can be made. However, our own signal detection models of reasoning are not fully developed process models; we see these as potentially constraining and informing process models of reasoning (for further discussion see Dube et al., 2010, 2011; Heit & Rotello, 2010; Rotello & Heit, 2009).

More generally, our results are compatible with recent theoretical work on two-process accounts of reasoning (Evans, 2008; Stanovich, 2009): participants are heavily influenced by automatically available information, such as argument length, and can only override this information in limited situations, for example when there is an alternative basis for response. Because our participants were simply instructed to judge plausibility, there was no readily available alternative.

Although our preference is for models that cut across experimental paradigms, we conclude by discussing implications for models of category-based induction, corresponding most closely to the paradigm in our own experiments. As we have noted, most models of category-based induction (Heit, 1998, 2000; Kemp & Tenenbaum, 2009; Osherson et al., 2000; Sloman, 1993; Tenenbaum & Griffiths, 2001) assume that argument length and plausibility judgement are closely connected: Each of these models predicts robust effects of argument length on judgement. In fact, if we had
found that instructions were effective in reducing the effect of argument length separately from the effect of validity, such results would have been challenging for these models, which do not have a means for evaluating plausibility that is not influenced by both length and validity.

The finding that valid arguments are considered weaker when they are longer is somewhat problematic for these models. Although some of these models (Osherson et al., 1990; Tenenbaum & Griffiths, 2001) predict exceptions to the usual effect that longer arguments are stronger, these exceptions involve changes in the superordinate category—for example, from *birds* to *animals* in (3) and (4). That exception does not apply to our experiments, in which the superordinate category was always *mammal*. Thus these models predict that all of our longer arguments should be judged to be stronger, yet participants consistently found the longer valid problems to be weaker (replicating Rotello & Heit, 2009).

We see a few possible routes to augmenting models of category-based induction. One possibility is to build in some notion of parsimony to evaluation of arguments (Lombrozo, 2007). For example, if a model of induction had an added component to give a slight penalty for longer arguments, valid arguments could only lose plausibility with increased length, because valid arguments would start with a very high level of plausibility. On the other hand, invalid arguments might be able to overcome the length penalty if making the argument longer increases its plausibility by other means (e.g., in terms of the coverage component, in the Osherson et al., 1990, model). Hence the model might predict a net increase in plausibility as invalid arguments get longer, overcoming a slight penalty for lack of parsimony. It may even be possible to build a more elaborate Bayesian model taking such constraints into account (for some examples of a related approach see Hahn et al., 2009; Hahn & Oaksford, 2007).

Another possible explanation is in terms of attention. For our valid arguments, validity depended on one crucial premise: Arguments were valid when one of the premises was a statement about mammals, or when there was a match between a premise category and the conclusion category, e.g., they both concerned rabbits. Whereas judging that an invalid argument is longer and hence more plausible does not depend on detailed scrutiny of each premise, judging an argument to be valid does depend on identification of the key premise. In a longer argument the crucial premise may be missed, so longer valid arguments may be judged weaker, on average, than shorter valid arguments. Comparing judgements about arguments from reliable versus unreliable speakers, one key difference may be poorer attention to the details of arguments from unreliable speakers. This attentional explanation highlights some important asymmetries between invalid arguments and valid arguments: Invalid arguments can vary greatly in their plausibility, but valid arguments need to be perfect (Skyrms, 2000). And it is a useful
reminder that higher-level reasoning processes depend on lower-level attentional processes.

Related to the both of the above explanations, the weakening of conclusions when valid arguments are longer can be thought of as a kind of dilution effect (e.g., Nisbett, Zukier, & Lemley, 1981; for the case that this effect is normative see Tetlock, Lerner, & Boettger, 1996). Whatever mechanisms are used to evaluate arguments could have limits on how much information is processed, or could in effect average together all of the evidence that is presented (cf., Anderson, 1965). Hence including non-diagnostic or weak evidence in addition to strong evidence could actually make a conclusion seem weaker, if the weak evidence is replacing or being averaged with the strong evidence.

To conclude, the pervasive effects of argument length on judgements of plausibility, across multiple reasoning tasks, present some interesting empirical puzzles and offer some challenges for future models of reasoning that ideally will address how people weigh evidence in multiple reasoning tasks.

REFERENCES


ARGUMENT LENGTH AND INDUCTIVE REASONING 275


