STANFORD UNIVERSITY

FIRST YEAR PROJECT

Syllogistic Reasoning as Communication

Author:
M. H. TESSLER

Advisor:
Noah GOODMAN

June 23, 2014

Contents

List of Figures

Proof is reasoning that causes us to know. **Aristotle**, *4th century BC*

The syllogism is like a text: fixed, boxed-off, isolated... The riddle belongs in the oral world. To solve a riddle, canniness is needed: one draws on knowledge, often deeply subconscious, beyond the words themselves in the riddle. **Walter J. Ong**, *Orality and Literacy*

The Jesuit priest, philosopher and cultural historian, Walter Ong, wrote about how the transition from orality to literacy altered the human experience. He says that the invention of written language—and hence, literacy—allowed a way of thinking which was untenable under a primary oral culture (Ong, 1982). For evidence, Ong turns to the the fieldwork of Russian psychologist Aleksandr Romanovich Luria among illiterate people in remote Uzbekistan and Kyrgyzstan in the 1930s. Luria would pose to the locals a question like the following:

> In the Far North, where there is snow, all bears are white.
> Novaya Zembla is in the Far North and there is always snow there.
> What color are the bears?

A typical response among illiterate persons: "I don't know. I've seen a black bear. I've never seen any others... each locality has its own animals." A response from a man who had recently learned to read and write: "To go by your words, they should all by white."

For man with no written language, the world exists as a stream of experiences: "an oral narrative proceeds by accretion, the words passing by in a line of parade past the viewing stand, briefly present and then gone, interacting with one another via memory and associations" (Gleick, 2011). The written word, by contrast, persists into future occasions, permitting analysis. From this, logical necessity—formal reasoning—is born. Formal reasoning, it seems, was a cognitive invention, and the invention takes practice to internalize. Then, the study of formal reasoning in lay people — who often do not have much practice in formal reasoning — may instead address some blend of informal and formal reasoning, the riddle-solver and the rational-reasoner.

The outline for the paper is as follows. First, I will introduce the first formal system of reasoning—the syllogism—and its relation to psychology. Then, I will describe the basic structure of a probabilistic model that is aimed to account for reasoning behavior with classical syllogisms and an experiment (Experiment 1) aimed to test this. I will then extend the probabilistic model to account for syllogistic reasoning using the generalized quantifiers: *most* and *few*, and compare it to a published study. A second extension of the probabilistic model will introduce communicative mechanisms to account for qualitative effects in syllogistic reasoning and quantitatively predict responses for "nothing follows". Finally, I will demonstrate how prior background knowledge can be naturally incorporate into the model to account for effects of "belief bias".

1 Syllogisms as a formal system

Aristotle tried to tame people's intuitive, informal reasoning behavior with a formal system, which he called syllogisms (roughly meaning: "form of thought"). Syllogisms are forms because they have no regard for content, just as variables in mathematical expressions can be instantiated by a multitude of numbers. Indeed, by introducing this form of reasoning, Aristotle actually introduces the concept of a variable (Lukasiewiecz, 1951).

The syllogistic form is made of 2 two-term sentences or propositions. The sentences together connect two "end" terms or variables (below: A,C) by a third, "middle" term (B). There are four unique forms, shown in Table 1.

The other component of a syllogism is a relation between the terms. Aristotle called upon quantifiers to handle this role. Though the meaning of quantifiers is open to debate (e.g. Horn, 1989), Aristotle used a particular meaning of the quantifiers to get the system off the ground. The four quantifiers used in syllogisms are shown below; I will use the bolded words as shorthand going forward.

A – B	**B – A**	**A – B**	**B – A**
B – C	**C – B**	**C – B**	**B – C**
—	—	—	—
A – C	**A – C**	**A – C**	**A – C**

Table 1: The 4 unique term-orderings of syllogisms

All All A are C

Some Some A are C

Not-all Some A are not C (or, Not all A are C)

None No A are C

The syllogistic space is defined by taking all possible combinations of premise term orderings (Table 1) and quantifiers. The space consists of 64 syllogisms. For each syllogism, there are 8 possible conclusions: 4 quantifiers crossed with A–C / C–A orderings. In the experiments described below, the conclusions is always restricted to a particular term ordering, for a total of 4 possible conclusions and 256 total possible responses.

Syllogistic reasoning has been replaced by more modern formalisms (e.g. the predicate calculus), yet persists in the experiments of cognitive psychologists. Syllogisms are an intriguing testing ground for human reasoning because they (a) are a formal system of logic[1] and (b) use natural language in their construction. The first part ensures there is a normatively correct solution, while the second part allows people with no training in formal logic to analyze these forms. What will we observe when we have people without training with a formal system use the system?

It should be noted that the system produces different results (i.e. different valid conclusions) under different interpretations of the quantifiers; it's likely that people need substantial training with these particular meanings to engage in accurate formal reasoning. Only a small fraction of the literature on syllogistic reasoning is concerned with learning (however see (e.g. Barwise & Etchemendy, 1994)). For the most part, psychologists study how people without much training with syllogisms.

It should come as little surprise, then, that people do not produce responses consistent with Aristotelian logic. What is surprising, however, is that for many syllogisms with no *necessary* conclusion, people consistently produce *some* conclusion—for some syllogisms as often as 88% of the time (Khemlani & Johnson-Laird, 2012). This suggests that logical necessity may not be the metric by which consequence (i.e. the notion that something follows from something prior) is measured.

In this paper, I propose a theory of syllogistic reasoning that treats the metric of logical consequence as probabilistic support and the understanding of natural language sentences as influenced by processes in social cognition. This latter component is important for all reasoning tasks that use natural language because language understanding can heavily influenced by the psychologically available alternative utterances and the particular roles "in the conversation" understood by the participants. In particular, this work extends a recently proposed formalism for pragmatic language understanding to apply to natural language arguments that invite natural language conclusions. The model incorporates background knowledge by sampling from prior distributions shaped by experience. I present three new experiments exploring aspects of probabilistic support, natural language understanding, and background knowledge in syllogistic reasoning.

2 Probabilistic Model

Our model begins with the intuition that people reason about concrete situations that can be idealized as a collection of objects with properties. To represent this type of richly structured model, we go beyond propositional logic and its probabilistic counterpart, Bayesian networks. We instead build our model using the probabilistic programing language Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008), a

[1] ...under some assumptions about the meaning of the quantifiers.

kind of higher-order probabilistic logic in which it is natural to describe distributions over objects and their properties. For background and details on this form of model representation, see `http://probmods.org`.

Situations are composed of *n* objects:

```
(define objects (list 'o1 'o2 ... 'on))
```

(Ellipses indicate omissions for brevity, otherwise models are specified via runnable Church code[2].) Syllogisms deal with 3 terms or classes of objects, and so the objects in these situations need only represent 3 properties. Properties `A`, `B`, and `C` of these objects are represented as functions from objects to the property value. We assume properties are Boolean, and so property values can be `true` or `false`. Initially, we assume no *a priori* information about the meaning of the properties; thus, they are determined independently:

```
(define A (mem (lambda (x) (flip br))))
(define B (mem (lambda (x) (flip br))))
(define C (mem (lambda (x) (flip br))))
```

Note that the operator `mem` memoizes these functions, so that a given object has the same value each time it is examined within a given situation, even though it is initially a random variable (via `flip`). Previous probabilistic models (Oaksford & Chater, 1994) have invoked a principle of rarity from the observation that properties are relatively rare of objects in the world[3]. For us, this simply means the base rate, `br`, of properties is small.

We interpret the quantifier sentences of syllogistic reasoning as truth-functional operators, consistent with standard practice in formal semantics. A quantifier (e.g. `all`) is then a function of two properties (e.g. `A` and `B`) which maps to a truth value by consulting the properties of the objects in the current situation. For instance:

```
(define all
  (lambda (A B)
    (all-true (map (lambda (x) (if (A x) (B x) true))
                   objects))))
```

Here the helper function `all-true` simply checks that all elements of a list are true, i.e. that all the *As* are indeed *Bs*. The function `map` applies the given function —`(lambda ...)`— to each element of the list `objects`. Similarly we can define `some`, `none`, `not-all` to have their standard logical meanings. We take *existential import* — that sets are non-empty — as an assumption, i.e. *all As are Bs* cannot be true if there are no As.

The key observation to connect these truth-functional meanings of quantifier expressions to probability distributions over situations is that an expression which assigns a Boolean value to each situation can be used for probabilistic conditioning. That is, the quantifier sentences can be used to update a prior belief distribution over situations into a posterior belief distribution. For syllogistic reasoning we are interested not in the posterior distribution over situations *per se*, but the distribution on true conclusions that these situations imply. In Church this looks like:

Listing 1: Full probabilistic model

```
(query
 (define objects (list 'o1 'o2 ... 'on))
 . . . define A, B, C . . .
 . . . define all, some, not-all, none . . .
 (define conclusion (conclusion-prior))

 conclusion

 (and (conclusion A C)
      (premise-one A B)
      (premise-two B C)))
```

The first arguments to a query function are a generative model: definitions or the background knowledge with which a reasoning agent is endowed. Definitions for which a prior is stipulated (e.g. `conclusion`) denote aspects of the world over which the agent has uncertainty. The second argument, called the *query expression*, is the aspect of the computation about which we are interested; it is what we want to know. The final argument, called the *conditioner*, is the information with which we update our beliefs; it is what we know.

[2] A fully-specified version of this model can be accessed at: `http://forestdb.org/models/syllogisms-cogsci14.html`

[3] This article is an article and it's about reasoning, but it's not a cat, and it's not a car, nor an elephant nor the color red. In fact, there's a very large number of things which this article is not.

We assume that the prior distribution over conclusions is uniform—the reasoner believes each conclusion is equally likely *a priori*. As the number of situations sampled grows large, the distribution over conclusions converges to the $P(\text{conclusion} \mid \text{premises})$.

3 Experiment 1

Experiment 1 sought out to test to the extent to which syllogistic reasoning patterns among lay people can be explained as a computation of $P(\text{conclusion} \mid \text{premises})$ — the probabilistic model.

3.1 Methods

110 participants located in the United States were recruited using Amazon's Mechanical Turk (MTurk) platform. All participants had over a 95% approval rate for MTurk submissions. Participants were compensated for their participation.

3.2 Materials

Participants were shown all 16 syllogisms (i.e. pairs of quantifiers) of a particular term ordering. The term ordering used in this experiment was:

A – B
B – C
—
A – C

45 participants saw the exact ordering above, while the other 65 saw the same form with the ordering of the premises flipped (i.e. B–C, A–B).

The terms were instantiated by common objects with common properties. Each syllogism was about a particular class of objects (e.g. microwaves) and the terms were common properties of that object (e.g. white, new, powerful). For example:

All white microwaves are new
No new microwaves are powerful

16 different object-properties tuples were used and randomly associated with a particular syllogism.

3.3 Procedure

Participants were told to assume that only the statements of the arguments were true and that each conclusions either did or did not follow from the argument. They were asked to rate their confidence in each of 4 possible quantifier conclusions (i.e. {*all*, *some*, *not-all*, *none*}) using a vertically-oriented slider bar. The slider scale ranged from "Definitely follows" to "Definitely does not follow". The ordering of the 4 quantifier conclusion sliders was randomized between subjects and remained constant within a subject.

A fifth horizontally-oriented slider bar was presented below the four and ranged from "None of the above conclusions definitely follows" to "At least one of the above conclusions definitely follows" (see Figure 1). Participants were required to touch each of the slider bars before advancing to the next problem[4].

3.4 Results

3.4.1 Raw data

The duration of the experiment was approximately normally distributed (see Figure 2(a)) with mean 15.1 and standard deviation of 6.0 minutes. 4 subjects were excluded for completing the study in under 5 minutes, leaving a total of 106 subjects.

[4]The experiment in its entirety may be viewed at: `http://stanford.edu/~mtessler/syllogism02/syllogism02_4.html`

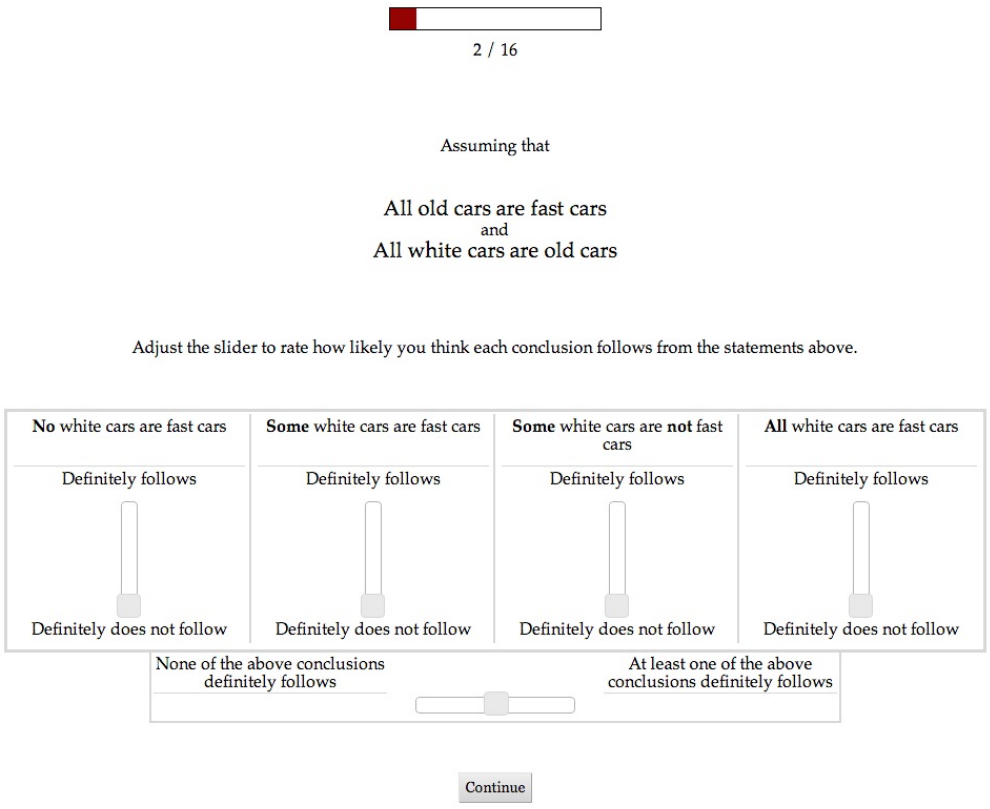


Figure 1: A sample syllogism from Experiment 1. Experiments 2 & 3 have a very similar layout.

The distribution of raw slider values is shown in Figures 2(b) and 2(c). The vast majority of responses are 0. The next most common response is 1. This is consistent with the task instructions (“conclusions either do or do not follow from the premises").

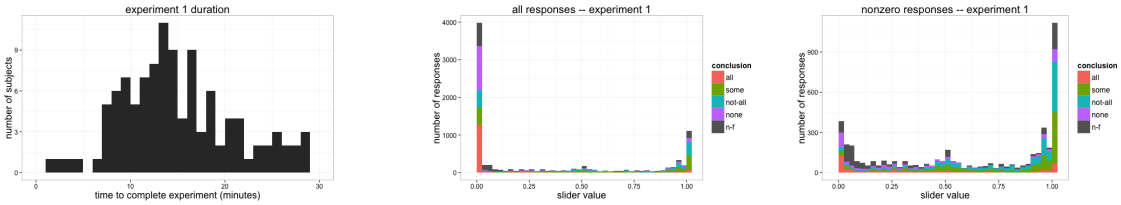
The correlation between the maximum slider value for a conclusion chosen by a participant for a syllogism and the “None of the above" slider value, across all syllogisms and subjects. was $r = -0.86$, suggesting that subjects responses were relatively internally consistent.

3.4.2 Comparison to other data sets

This experiment uses a novel dependent measure. Subjects were asked to rate their confidence in the presented conclusions. Other studies in syllogistic reasoning either ask participants to draw their own conclusion(s) (often specifying the appropriate format)—“production" (Johnson-laird & Steedman, 1978; Johnson-Laird & Bara, 1984)— accept or reject presented conclusions—“evaluation" (Rips, 1994)— or choose among a list of presented conclusions— “forced choice" (Dickstein, 1978; M. J. Roberts, Newstead, & Griggs, 2001). The confidence rating is most akin to the “evaluation" task. The “evaluation" format used by Rips (1994) did not include an explicit “nothing follows" option. As such, we can only compare to this data set by removing the “nothing follows" responses from our data set and renormalizing.

Two meta-analyses (Chater & Oaksford, 1999; Khemlani & Johnson-Laird, 2012)– henceforth CO & JL– have compiled data across multiple studies of syllogistic reasoning. OC included 5 studies and JL included 6; 3 of the constituent studies included in the two meta-analyses were the same — from Johnson-laird and Steedman (1978); Johnson-Laird and Bara (1984).

The sample sizes of the constituent studies tended to be small ($\text{mean}_{CO} = 31.6$, $\text{median}_{CO} = 20$; $\text{mean}_{JL} = 27.2$, $\text{median}_{JL} = 20$). The samples themselves were typically undergraduates — both American and Italian — though one study included in JL was with high school students in Florence, Italy. The data presented for both of the meta-analyses is in similar form: proportion of responses for each of the conclusions.



(a) Time to complete study. (b) All slider values. (c) Non-zero slider values.

Figure 2: Experiment 1 raw data

Our data set is somewhat correlated with each of the meta-analyses ($r_{JL} = 0.68$; $r_{CO} = 0.65$). Qualitatively, the data sets disagree most strongly on the endorsements of the *some* and *not-all* conclusions, with Experiment 1 data tending to have higher endorsements of these conclusions. This is not altogether surprising given the response format. Subjects are required to consider each conclusion (by touching the sliders) and rate their confidence. These two conclusions are the more *conservative* conclusions and tend to elicit an overall higher endorsement than *all* and *none* (see e.g. Figure 2(c), slider values near 1). Our data set without the "nothing follows" response is also highly correlated with data from Rips (1994) ($r_{Rips} = 0.74$). These correlation values are consistent with the mean correlations between data sets reported in the CO meta-analysis (mean=0.72).

3.4.3 Responses by syllogism

Mean responses with bootstrapped 95% confidence intervals for all 16 syllogism are shown in Figure 3. It is immediately apparent that at least half of these syllogisms produce qualitatively similar responses, with predominant responses for *some* and *not-all* (e.g. rows 2 and 3). Only one syllogism gives rise to a modal *all* response, and only one elicits a modal *none* response. The "nothing follows" (n-f) response varies appreciably across syllogisms. Compare for instance, the bottom left and the top right syllogisms (none / all vs. all / none). There is a considerable different pattern of response as logic would predict (*all* is an asymmetric relation so the two syllogisms are not equivalent). This is an example of syllogism where tracking the order of the terms (i.e. the form of the syllogism) is critical — this is something that heuristic model (e.g. the Probability Heuristics Model, see below) neglects.

3.4.4 Model predictions

The basic probabilistic model introduced in Section 2 generates posterior distributions over the 4 conclusions for each syllogism (Figure 4). Critically, this does not include predictions for "nothing follows" (grey bars; Figure 3). Thus, we remove the "nothing follows" responses for the Experiment 1 data and renormalize.

The model has two parameters: the number of objects in a situation `n_objects` and the base rate of properties `br`. I fit these parameters to the data by maximum likelihood to 7 and 0.58, respectively. This relatively high `br` is likely due to the contents of the syllogistic arguments being common properties of common objects.

The model shows good overall correspondence with the data (Figure 5(b), $r = 0.91$). It is evident, however, that by removing the "nothing follows" responses, the variability of responses between syllogisms is somewhat diminished. Most syllogisms elicit relatively strong *some* and *not-all* responses. The model is able to capture this as well as the syllogisms that elicit the modal *all* and *none* responses (Figure 4, top left and top right). Thus, it seems that the $P(\text{conclusion} \mid \text{premises})$ is a good model for syllogistic reasoning.

One concern may be that much of the data is clustered around high *some* and *not-all* responses. Indeed, a model that computes just $P(\text{conclusion})$ — without conditioning on the truth of the premises — provides a pretty good fit to the data (Figure 5(a); $r = 0.72$). This suggests that many syllogisms are bad tests for reasoning models in general, since the same conclusions can be reached by incorporating or ignoring the premises. This is likely a result of how the syllogistic space is derived: by taking all possible combinations of term orderings and quantifiers. There is no guarantee that this combinatorial technique would give rise to

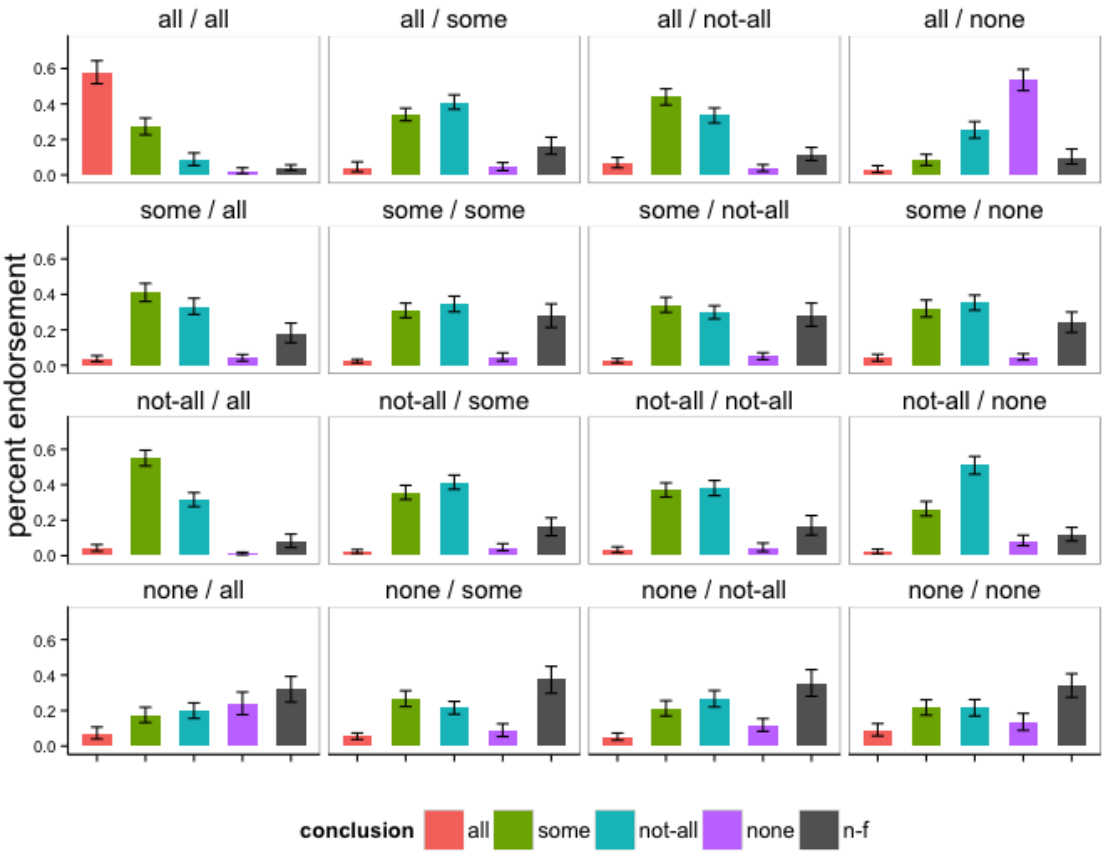


Figure 3: Experiment 1 – responses to syllogisms. Syllogistic premises are of the form: A–B / B–C. Conclusions are: A–C.

meaningful content. Indeed, most syllogisms in the space are invalid. As well, because a conclusion-only model can account for many patterns of responses, many syllogisms appear to be uninformative.

4 Model extension 1: Generalized quantifiers

The probabilistic model introduced in Section 2 is based on a truth-functional semantics. Sentences are used to update prior distributions over worlds. As such, the model in its most basic form is able to accommodate any sentence with a truth-functional meaning. A very simple extension of this model is then to use other quantifiers than those specified in the classical syllogism. Chater and Oaksford (1999) tested this very same idea in their Probability Heuristics Model by giving participants syllogisms using the generalized quantifiers: *most* and *few*.

As with the classical quantifiers in the probabilistic model, I have to specify some definition of the word. Unlike the classical quantifiers, however, there is considerable debate about the precise meaning of words like *most* and *few*. To a first approximation, the words can be defined as thresholded functions. That is, *most* A are B is true if $\frac{n_{A\&B}}{n_A} > \theta_{most}$ for some $\theta_{most} \in (0,1)$ and *few* A are B is true if $\frac{n_{A\&B}}{n_A} < \theta_{few}$ for some $\theta_{few} \in (0,1)$.

In Church, this looks like:

```
(define most (lambda (A B theta)
   (>
    (/ (num-true (map (lambda (x) (if (A x) (B x) true)) objects))
       (length (map A objects)))
    theta)))
```

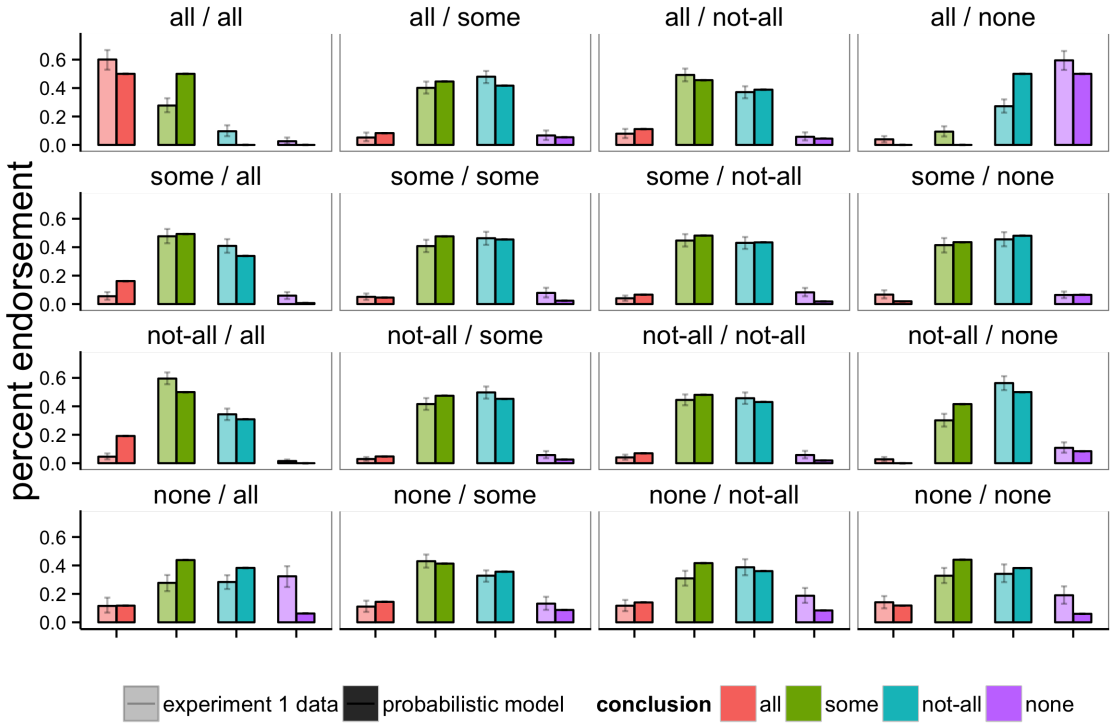


Figure 4: Experiment 1 – data and model predictions. “Nothing follow" responses have been removed and the data renormalized. Syllogistic premises are of the form: A–B / B–C. Conclusions are: A–C.

As a first test of the model, I take $\theta_{most} = \theta_{few} = \frac{1}{2}$. That is, I take *most* to mean literally *more than half*.

4.1 Studies with generalized quantifiers

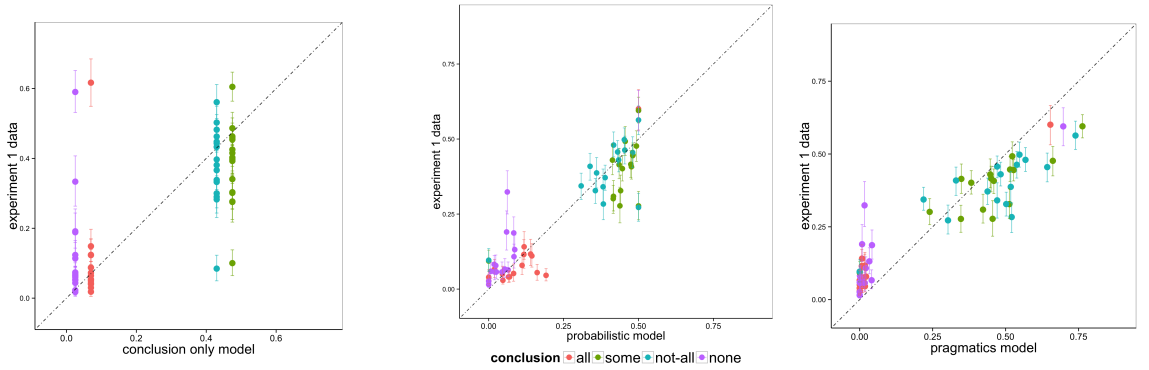
Chater and Oaksford (1999) extended the set of possible syllogistic reasoning problems by including the quantifiers *most* and *few* with the set of classical quantifiers. With 6 quantifier types, the space of syllogisms includes 144 problems. Pilot work suggested it was too taxing for participants to complete all 144 syllogisms so the authors split the quantifiers into 2 sets of 4, and ran two experiments using different sets of quantifiers:

1. *all*, *most*, *few*, and *not-all*
2. *most*, *few*, *some*, *none*

Each study included twenty participants. Participants received a booklet with 64 syllogisms in random order. All syllogisms involved lexical categories used in previous syllogistic reasoning studies e.g. *Most artists are beekeepers; Few chemists are beekeepers*. Participants were presented with all 4 possible conclusions in the C–A ordering (i.e. the first term of the conclusion was always the one from the second premise) and were told to mark any and all conclusions they believed followed from the premises. If they believed that no conclusion followed, they were instructed to leave the entry blank.

4.2 Results and model fit

For each experiment, there are 64 syllogisms and 256 total responses (not including the “nothing follows" responses). Again, I fit `n_objects` and `br` to the data by maximum likelihood to 6 and 0.30, respectively, for both experiments. Notice how `br` is relatively smaller for these data than for Experiment 1, again likely due

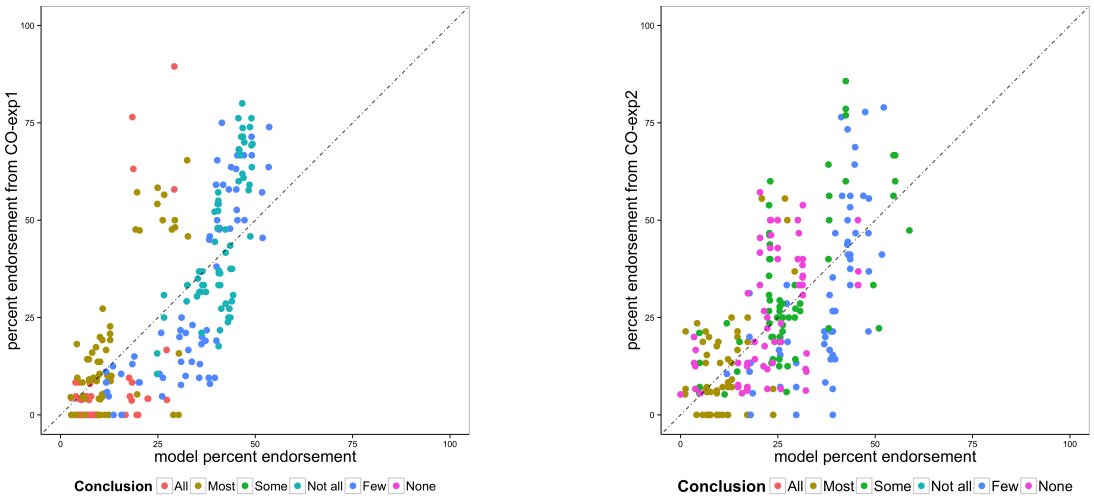


(a) conclusion-only model $r = 0.72$ (b) probabilistic model $r = 0.91$ (c) pragmatics model (Section 5) $r = 0.94$

Figure 5: Model fits – experiment 1

to the contents of the syllogistic arguments; in these experiments, the contents were professions and hobbies (e.g. artists, bakers, chemists).

There is good correspondence between the experimental data and the probabilistic model (Figure 6). In CO–Experiment 1, the correlation between model predictions and experimental data is $r = 0.79$; in CO–Experiment 2, $r = 0.65$. CO–Experiment 2 had appreciably more *nothing follows* responses than Experiment 1. This may be the source of the relatively poorer fit.



(a) CO – Experiment 1 $r = 0.79$ (b) CO – Experiment 2 $r = 0.65$

Figure 6: Experiments using generalized quantifiers. Experiment CO-1 used the quantifiers {*all*, *most*, *few*, *not-all*}. Experiment CO-2 used the quantifiers {*most*, *few*, *some*, *none*}.

The probabilistic model naturally incorporates other quantifiers to make reasonable predictions for a range of two-sentence arguments. This is even so using a coarse model of semantics (i.e. a thresholded function). It is likely that a more sophisticated model of the semantics of words would better account for the data.

This concludes the section on generalized quantifiers. Next, I will discuss how communicative principles might affect syllogistic reasoning.

5 Model extension 2: Communication and “nothing follows"

Consider the following multiple choice problem that one might find in a standard reading comprehension test.

> It has been suggested that long-term prisoners, on release from jail, be given a reasonable state pension to reduce the likelihood of their resorting to crime. Most people instinctively reject the suggestion as they feel it would be like rewarding criminal activity.
>
> The supporters of the prisoners' pension scheme have criticized those who reject this possibility, by claiming that for the critics...
>
> (Which of the following is the most logical completion of the sentence above?)
>
> A. emotion is more important than justice
>
> B. punishment for criminals is more important than crime prevention
>
> C. crime prevention is not an important issue
>
> D. money has too high a value
>
> E. the law should not be concerned with what happens after jail
>
> (*GRE Reading Comprehension Practice Test 01 majortests.com*, n.d.)

How do you think a test-taker might arrive at the correct answer? The question requires not only logical consistency or probabilistic support but pragmatic inference.

The probabilistic model as it stands now generates situations consistent with the premise sentences, and checks which of the conclusion sentences are also consistent. There are two observations that the model in its current form does not capture: (1) that interpreting and producing sentences involves principles of communication and (2) that participants often respond “nothing follows" to a syllogistic argument. I now introduce the formalism that will attempt to capture both of these.

The intuition that communicative principles may play a role in syllogistic reasoning comes from the following observation: in the syllogism

> All expensive rugs are new
>
> All new rugs are brown

approximately 80% of responses are *All expensive rugs are brown* while only 10% of responses are *Some expensive rugs are brown*, even when multiple responses are allowed and even when participants must explicitly evaluate each. Since *all* logically entails *some*, all situations in which *all* is true are situations in which *some* is true. Indeed, they are both valid conclusions. The consequences of this symmetry can be observed in Figure 4 in the *all / all syllogism*. This is same syllogism quoted above: it is valid with both *all* and *some* as conclusions. This is evident in that each response gets half of the total endorsement: they are both equally good conclusions from a probabilistic standpoint. The observed asymmetry in human responses challenges standard logical accounts of reasoning, even those rooted in probability like the model described above.

Viewing syllogistic reasoning as a special case of communication suggests that reasoning should go beyond the semantics of language. Following the *rational speech-act* (RSA) theory (Goodman & Stuhlmüller, 2013; Frank & Goodman, 2012), I suggest that the participant interprets the sentences of the syllogism as if they were crafted by an informative agent (in this case, the experimenter). The experimenter, then, is believed to convey information about which only he has access. Intuitively, this *private information* for the experimenter is “the right answer". Thus, the syllogisms are interpreted as being informative with respect to a particular conclusion.

5.1 Communication in Church

This intuition is formalized in Church by introducing into the model a function I call the `experimenter`, which takes a conclusion as an argument (“the experimenter has a particular conclusion in mind..." or, equivalently, “this question has a particular answer"). The experimenter then chooses premises he believes will lead the reasoner to the right conclusion `(equal? conclusion (reasoner premises))`. The `reasoner` function is almost exactly our `query` from Listing 1, which computes the $P(\text{conclusion} \mid \text{premises})$. The only difference is the introduction of the final conditioning statement, which says the premises that were given to the reasoner were given by an

experimenter who had a particular conclusion in mind: `(equal? premises (experimenter conclusion))`. This sort of recursive reasoning is interpreted as a type of pragmatic inference: language understanding that comes not only from the meaning of words but from social cognition as well.

Listing 2: Probabilistic Pragmatics model

```
(define (experimenter conclusion depth)
  (query
   (define premises (premise-prior))

   premises

   (equal? conclusion (reasoner premises depth))))

(define (reasoner premises depth)
  (query
   . . . define objects, A,B,C, {quantifiers} . . .
   (define conclusion (conclusion-prior))

   conclusion

   (and (conclusion A C)
        (if (= depth 0)
            (and (first premises) A B)
                 (second premises) B C))
            (equal? premises (experimenter conclusion (- depth 1)))))))
```

The model described above is instantiated in this Church program when `(= depth 1)`. The `depth` parameter is used to determine when the recursive inference should complete, i.e. when to interpret the words literally. When `(= depth 0)`, the model is the same model that we derived in Listing 1: $P(\text{conclusion} \mid \text{premises})$. In theory, increasing `depth` could qualitatively change the inference; in these models, however, only quantitative changes are observed, tantamount to an "inverse temperature" parameter. As such, I will treat `(= depth 0)` and `(= depth 1)` as two different models (Probabilistic and Pragmatics) and introduce an "inverse temperature" or "optimality" parameter, α, separately.

5.1.1 Experiment 1 revisited

I know revisit the data from Experiment 1 using the predictions of the Probabilistic Pragmatics model with parameters `br`, `n_objects`, and α fit by maximum likelihood to values 0.47, 7, and 3.5, respectively (Figure 7). The most striking effect is that the pragmatics model has a preference among equally probable conclusions (i.e. valid conclusions which are, by definition, 100% probable; row 1, columns 1 & 4), just as people do. As well, a slight preference in probability can be enhanced to a more dramatic preference for some syllogisms (Figure 7, light bars vs. lightest bars; row 3, columns 1 & 4), but not for others (row 3, columns 2 & 3). Still, there are some responses the pragmatics model is not able to capture (e.g. row 4, column 1, *none* response). This may be due to an interaction with the "nothing follows" response (see Figure 4).

Quantitatively, the model provides a better fit to the data (Figure 5(c)) $r = 0.94$.

5.1.2 Generalized quantifier experiments revisited

I also examine the fit to the 2 Chater and Oaksford (1999) experiments using the generalized quantifiers *most* and *few*. The pragmatics model exhibits an increased fit to both datasets, $r_{CO-exp1} = 0.83$ and $r_{CO-exp2} = 0.67$. As well, the model qualitatively exhibits more gradedness in responses (see Figure 8, more variability in the X-axis variable). It's remarkable than such a good fit can be achieved by a very simple and conservative model of the meanings of *most* and *few*. The fit to CO - Experiment 2 is likely worse than the fit to CO - Experiment 1 because of the higher rate of "nothing follows" responses in Exp 2. I will re-examine this in future work.

5.1.3 A different pragmatic inference

In addition to determining where the inference "bottoms out", the `depth` parameter isolates the part of the inference that is thought to be determined recursively[5]. In the model above, note how in the conditioning

[5]For this exercise, it's particularly important to keep track of (parentheses).

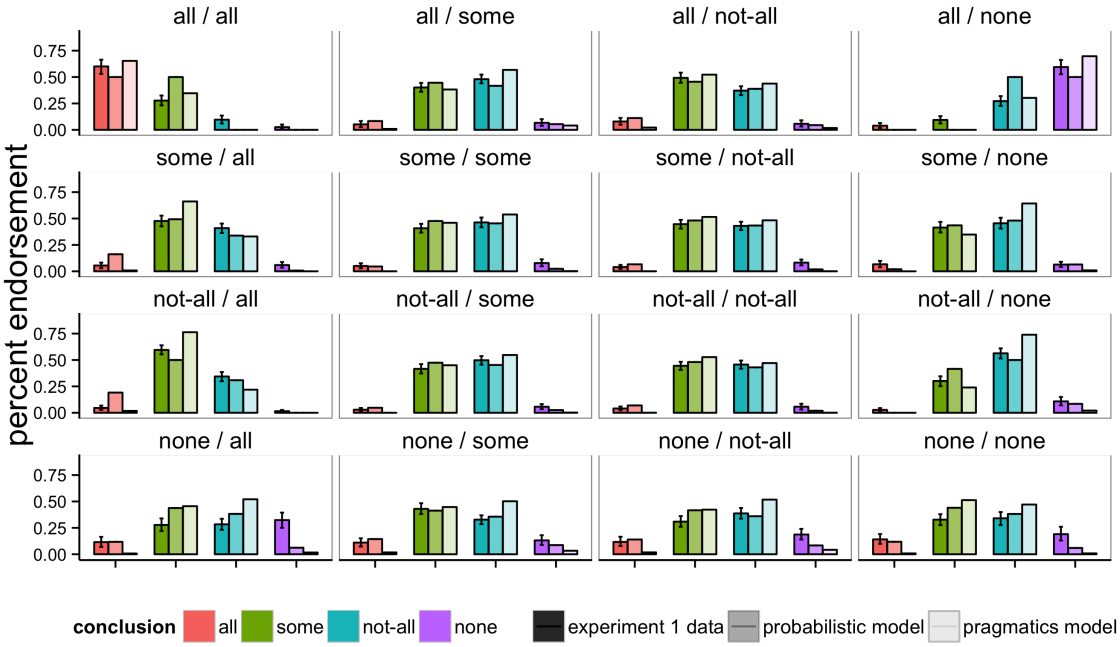


Figure 7: Experiment 1 – data, probabilistic, and pragmatics model predictions. The pragmatics model has a preference among valid conclusions (row 1, columns 1 & 4). The pragmatics model shows preferences in some invalid syllogisms (e.g. row 3, columns 1 & 4). Syllogistic premises are of the form: A–B / B–C. Conclusions are: A–C.

statement, the `conclusion` is set outside the scope of the recursion. This is saying that the premises are interpreted as coming from a pragmatic agent, but the reasoner's conclusion should be produced without appeal to recursive reasoning. We could instead define the conditioning statement like this:

```
(and ((first premises) A B)
     ((second premises) B C)
     (if (= depth 0)
          (conclusion A C)
          (equal? premises (experimenter conclusion (- depth 1)))))))
```

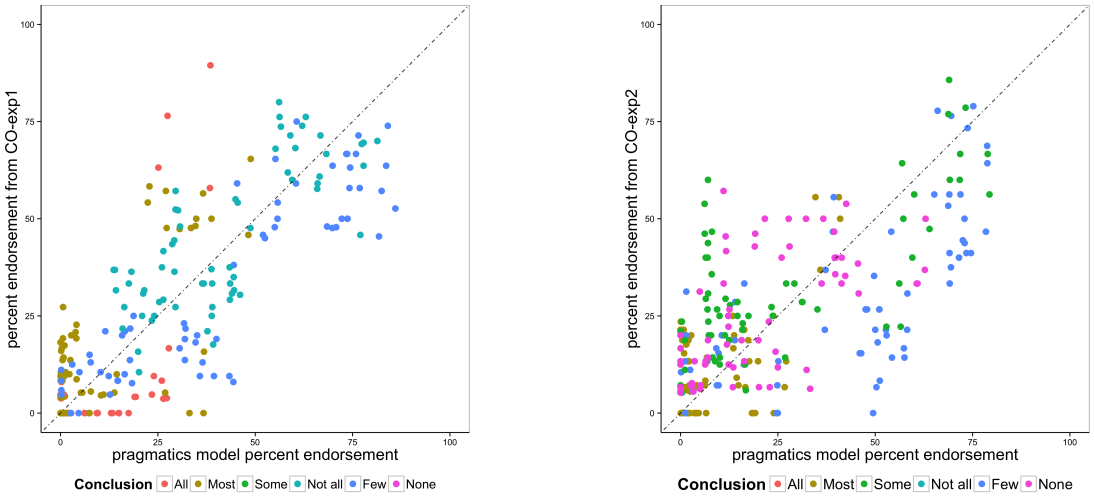
The interpretation of this model would be slightly differently. This would say the premises are interpreted literally, but the conclusion is produced pragmatically[6]. How can we understand the pragmatic production of a conclusion? In RSA, the speaker is said to convey utterances uniquely-illustrative of the world-state she has observed. In this model, the reasoner would be acting as a speaker, conveying conclusions illustrative of the premises she has received. The experimenter, then, would be thought of as a listener, trying to reconstruct the world (or, in this case: the premises) from the utterance (or, conclusion) he has received. In the reasoner's mind, the role of the experimenter would no longer the careful crafter of questions but the grader or evaluator of the reasoner's answers. The reasoner asks herself, "is this answer a good answer, given that I have these alternatives?"

I will not present a full evaluation of this alternative pragmatics model in this article.

5.1.4 Relationship to *Rational Speech-act* theory

In RSA, the speaker's access was a current state of the world. By being informative with respect to a world-state, the speaker is able to communicate enriched meanings (e.g. scalar implicature – that "some" may

[6]This distinction between production and interpretation comes from the different query (or, return) expressions for each function. The `reasoner`, e.g., is said to *interpret* premises and *produce* conclusions, because `premises` is the function's argument and `conclusion` is what the function returns.



(a) CO – Experiment 1 $r = 0.83$

(b) CO – Experiment 2 $r = 0.67$

Figure 8: Pragmatics model fit to experiments using generalized quantifiers. Experiment CO-1 used the quantifiers {*all*, *most*, *few*, *not-all*}. Experiment CO-2 used the quantifiers {*most*, *few*, *some*, *none*}.

also imply “not all"). It is known, however, that “local" scalar implicatures do a poor job qualitatively of accounting for reasoning with syllogisms (M. J. Roberts et al., 2001). Indeed, a preliminary analysis of a standard “local" pragmatic-listener model in this framework was consistent with this account.

However, a listener (our reasoner) may consider the premises in a wider, conversational setting: she may ask herself why the experimenter chose to give these particular premises, as opposed to alternative arguments. This requires a closer look at what the reasoner believes to be at stake in this “conversation”—the Question Under Discussion, or QUD (C. Roberts, 2004). In a syllogistic context, we take the QUD to be “what is the relationship between A & C (the end terms)?", very often the actual context in which the experiment is presented.

In this setup, pragmatic inferences will differ from the standard local implicatures found in RSA; for instance, “Some A are B” may not lead to a “Not all A are B” implicature if “All A are B” provides no additional information about the A–C relationship. The enriched meanings then come from the following counter-factual consideration: “why did this experimenter present me with this argument and not any other argument?" The pragmatic reasoner draws conclusions that are more uniquely determined by the particular argument the experimenter provides.

The A-C QUD is naturally captured by a `reasoner` who considers an `experimenter` who considers the conclusions the `reasoner` would draw about A & C (not the reasoner’s inferences about the whole world-state, which would include information—superfluously—about B).

5.2 The *mu* utterance

We are now ready to consider the problem of “nothing follows" responses (also referred to as: No Valid Conclusion). The quantifiers used in classical syllogisms have important relations to each other. One very important property of the set of the quantifiers in classical syllogisms is that the set is composed of two pairs of contradictions (see Figure 9).

A pair of contradictions is a pair of relations such that the relations cannot both be true but one must be true. Thus, every situation that the probabilistic model generates will have exactly two conclusions that will be true (one from each set of contradictions). This poses a problem for expressing the fifth conclusion option: “none of these conclusions follows", since each generated situation has some statement that is true of it (actually 2 statements).

Thus far, we have presented syllogisms with respect to the probabilistic support they provide for various

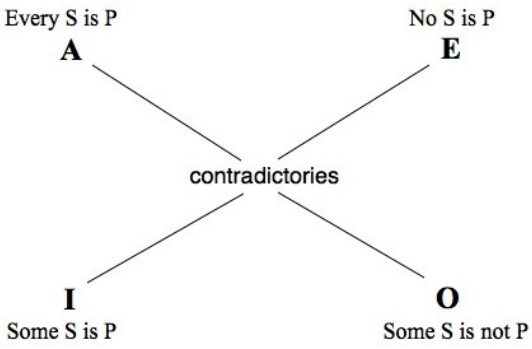


Figure 9: Square of opposition

conclusions. What would it mean, then, to conclude that “nothing follows"? One possibility is that the argument (or in Bayesian framework: the evidence) provides little support for any conclusion beyond that which is provided by the conclusion alone (the prior). The extreme case of this corresponds intuitively to the notion of a “vacuous argument".

We can attempt to capture patterns of responses in support of a “nothing follows" conclusion by introducing a fifth conclusion into the set of possible conclusions. I call this fifth conclusion “mu", borrowed from the Chen and Zen traditions meaning roughly *without*; or *not have*, but has also been taken to mean *un-ask the question* (Pirsig, 1974). Support for the *mu* conclusion should be highest for the syllogisms that convey the least information, and visa versa. The mechanism for determining the informational content of syllogisms with respect to conclusions is already available in the recursive model described in Listing 2. Thus, we can simply append the list of possible conclusions with `mu`, a conclusion which is always true.

```
(define mu (lambda (A B) true))
```

In the `(= depth 0)` model, `mu` will always receive $\frac{1}{3}$ endorsement[7]. In the `(= depth 1)` models (either the interpretation or the production model), this value will be higher for syllogisms will lower informational content (i.e. more vacuous arguments) and be lower for syllogisms with higher informational content (e.g. valid syllogisms).

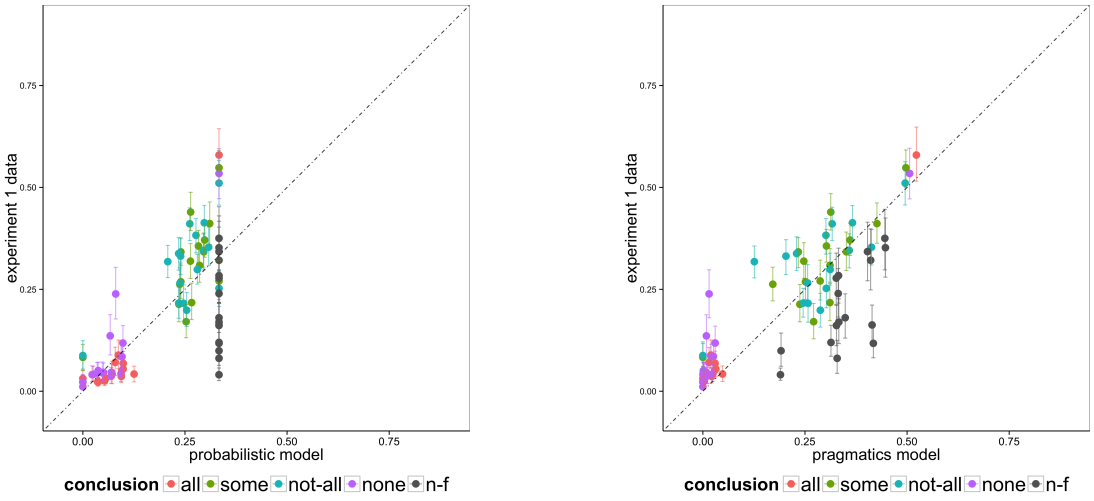
I leave for future work the precise mathematical formulation of this idea but I believe something like the Kullback-Leibler divergence between the conclusion-only model (the prior) and the probabilistic model (the posterior) could serve as an alternative definition of “informational content/gain" with which to compare “mu" endorsement.

I’ll briefly examine Experiment 1 data with the “nothing follows" responses using the “mu" utterance. With the “nothing follows" responses in the data set, the number of data points increases from 64 to 80. We would expect the probabilistic model fit to be worse, as “mu" can only receive gradedness via pragmatics (per my discussion above). The overall fit for the probabilistic model is: $r_{probabilistic} = 0.73$ (Figure 10(a)).

The recursive reasoning (pragmatics) model modulates the endorsement probability by the degree to which the syllogism uniquely (i.e. compared to other syllogisms) supports the conclusion. The “mu" conclusion is something which is always true, and hence should be modulated by the degree to which the distribution of the other 4 conclusions is uniquely determined by the syllogism (the evidence). Another way of putting it is the higher the informational content of the syllogism, the lower endorsement of the “mu" response. I have not provided a formal proof here or even simulation evidence. For now, I will only compare it to the “nothing follows" responses from the experimental data.

The overall correlation between the pragmatics model including a “mu" conclusion and the data from Experiment 1 is $r_{pragmatics} = 0.87$ (Figure 10(b)). Looking at only the sixteen data points that correspond to the “nothing follows" response, the “mu" conclusion in the pragmatics model is closely associated with it $r_{mu} = 0.65$ while the probabilistic model has no association (since “mu" receives constant endorsement).

[7]Endorsement is always normalized. With the addition of `mu`, each situation will have 3 conclusions that are true: 1 from {*all*, *not-all*}, 1 from {*some*, *none*}, and *mu*. Hence, *mu* will receive $\frac{1}{3}$ of the probability mass.



(a) probabilistic model with *mu* $r = 0.73$

(b) pragmatics model with *mu* $r = 0.87$

Figure 10: Model fits with “mu" utterance – experiment 1

This suggests that participants’ responses that “nothing follows" is negatively associated with the extent to which the syllogism supports conclusions beyond that which the prior suggests.

We can take a closer look at the data by plotting all 16 syllogisms as we did before. We can see that the modulations of “n-f" due to pragmatics are subtle (Figure 11, medium grey vs. light grey bars). The largest modulations occur for the valid syllogisms with two valid responses (row 1, columns 1 & 4). As well, there is an increase in endorsement for the entire bottom row, which are relatively uninformative syllogisms. Many patterns for the “nothing follows" response are not captured (e.g. row 3, col 1; row 1, col 2 & 3). Yet the correlation between “mu" and “nothing follows" ($r_{mu} = 0.65$) suggests at least some interesting patterns of variance are being captured with this formalism.

6 Model extension 3: Background knowledge

Thus far we have assumed the properties or terms related in the syllogisms are independent and identically distributed: `A`, `B`, and `C` were all defined by `(flip br)`. This assumption may have minor effects given the diversity of the stimuli used in Experiment 1. However, it is known that the content of syllogisms affects reasoning in profound ways (Wilkins, 1928; Evans, Handley, & Pollard, 1983; Oakhill, Johnson-Laird, & Garnham, 1989; Oakhill & Garnham, 1993; Newstead, Pollard, Evans, & Allen, 1992; Newstead & Evans, 1993; Cherubini, Garnham, Oakhill, & Morley, 1998) — an effect referred to as “belief bias". There has been much discussion as to possible sources of belief bias in syllogistic reasoning. Many of these qualitative theories rely on categorical distinctions between *a priori* believable and unbelievable statements. The probabilistic model introduced in this paper has a natural way of incorporating background knowledge, without having to make categorical distinctions.

Often, the contents of the syllogism are devised so as to minimize semantic relations between the terms. In reality, of course, this is impossible. Consider for a moment the oft used example: artists, bakers, and chemists. Though at first glance, it may seem that these are unrelated, there are actually strong semantic relations between the terms. Bakers and chemists both use recipes and ingredients in their work, and some bakers would even consider themselves artists (e.g. The ACME Bread Company in San Francisco).

Incorporating background knowledge into the probabilistic model requires treating the properties as non-independent.

```
(define ABC (mem (lambda (x)
                   (multinomial (list 'ABC 'AB_ 'A_C '_BC 'A__ '_B_ '__C '___) background-prior))))
```

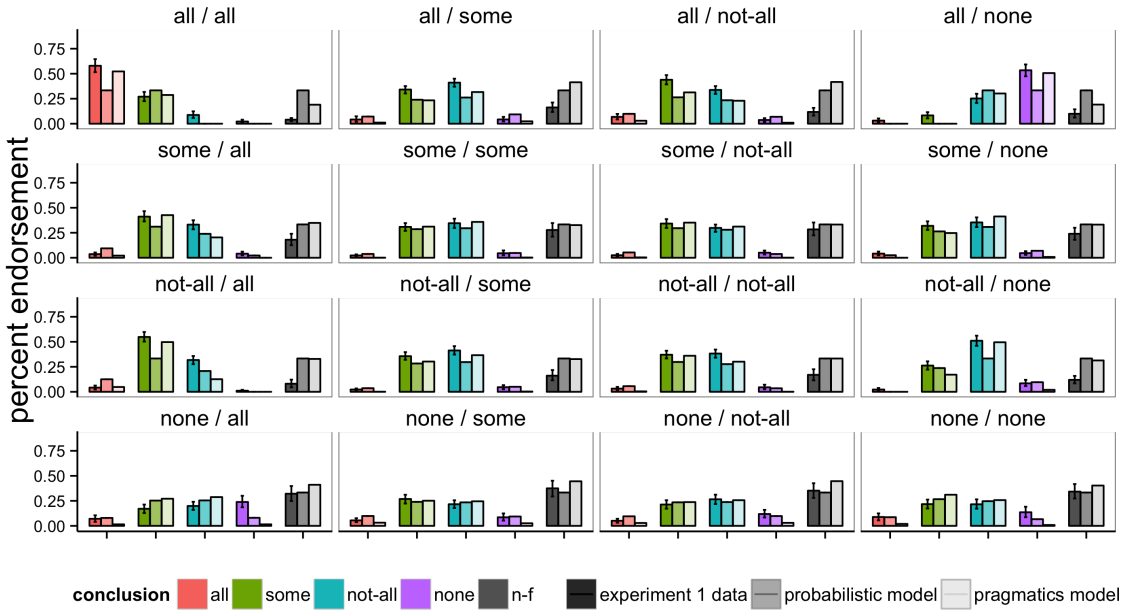


Figure 11: Experiment 1 – data and model predictions including “nothing follows" responses. Syllogistic premises are of the form: A–B / B–C. Conclusions are: A–C.

Here, an object’s properties are represented jointly: each letter denotes the presence of the particular property while the _ represents the absence of a property. `background-prior` is the joint probability distribution over properties `A`, `B`, `C`, represented as a list per Church convention; it implicitly represents the correlations among the properties under consideration.

6.1 Experiment 2

Experiment 2 sought out to see if structured background knowledge played an important role in syllogistic reasoning.

6.1.1 Methods

50 participants located in the United States were recruited using Amazon’s Mechanical Turk (MTurk) platform. All participants had over a 95% approval rate for MTurk submissions. Participants were compensated for their participation.

6.1.2 Materials

Participants were assigned to a “content condition" (List A or List B) and shown 6 syllogisms with content from the assigned list. The lists were loosely designed to be either “meaningful" or “arbitrary" content, but in practice this distinction was not relevant; I consider the lists as representing 12 different domains, and each syllogism has 2 possible domains. The syllogisms and content used in this experiment are shown in Table 2.

The syllogistic forms were selected to match those used by Oakhill et al. (1989). These were chosen by the original authors to elicit a range of *belief biases* in participants responses. That is, some syllogisms were expected to be more susceptible to “bias" than others.

In the second half of the experiment, participants were shown combinations of the properties from their list and asked to rate the likelihood of people existing with those properties. All possible combinations of the three terms were presented with a slider bar ranging from “Very unlikely" to “Very likely". Participants were required to touch all the slider bars before proceeding.

Experiment 2 Materials					
Syllogism	List A	List B	Syllogism	List A	List B
	Golfers	Actresses		Lazy students	Scientists
All A are B	Retired	Hikers	Some A are B	Dean's list students	Bowlers
No B are C	Full-time employees	Jugglers	All B are C	Good students	Hikers
	Married tenants	Artists		Wine drinkers	Firemen
No B are A	Bachelors	Bakers	No B are A	Italians	Italians
All B are C	Renters	Canoeists	Some B are C	Saudis	Saudis
	Trainers	Movers		Church-goers	Nurses
All B are A	Can bench 3x body weight	Bakers	All B are A	Religious	Attendants
Some B are C	Weight lifting competitors	Quakers	All C are B	Priests	Chemists

Table 2: Content domains used in Experiment 2 syllogisms.

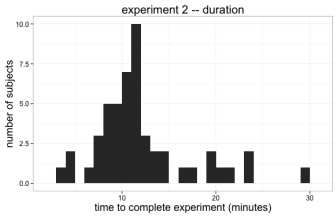
6.1.3 Procedure

The procedure was the same as in Experiment 1. There were 4 conclusion sliders, amounting to the 4 quantifiers in the C–A direction. In addition, a fifth slider bar was used for "None of these conclusions follows" as in Experiment 1.

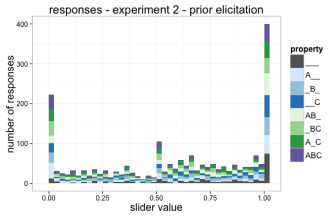
The prior elicitation section began right after the syllogistic reasoning section[8].

6.1.4 Results

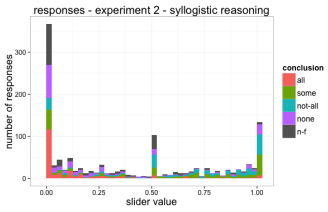
The raw data of total experiment duration and binned slider values for both parts of the experiment are shown in Figure 12. Generally, participants reported that sets of properties either were or were not plausible configurations. Figure 12(c) replicates the overall pattern of responses in Experiment 1 (Figure 2(b)).



(a) Time to complete study.



(b) Prior elicitation slider values.



(c) Syllogistic reasoning slider values.

Figure 12: Experiment 2 raw data

6.1.4.1 Prior elicitation

The elicited prior distribution over properties are shown in Figure 13. One domain is absent due to an error in the experiment implementation. It's clear that a wide range of prior distributions over properties was elicited. List A appears to have more variance than List B. This is somewhat expected from the supposed "meaningful" / "arbitrary" distinction, though all that matters for my purposes is that a range of priors was elicited.

6.1.4.2 Syllogistic reasoning

Patterns of responses for all six syllogisms are shown in Figure 14. Different syllogisms elicit different patterns of responses, showing that participants were sensitive to the logic of the arguments. Different domains, however, modulated responses in a number of syllogisms (e.g. bottom row of Figure 14). Interestingly enough, the bottom row of the table of responses corresponds exactly to the syllogisms that Oakhill et al.

[8]This experiment can be viewed at: `http://stanford.edu/~mtessler/syllogism00/syllogism0.html`

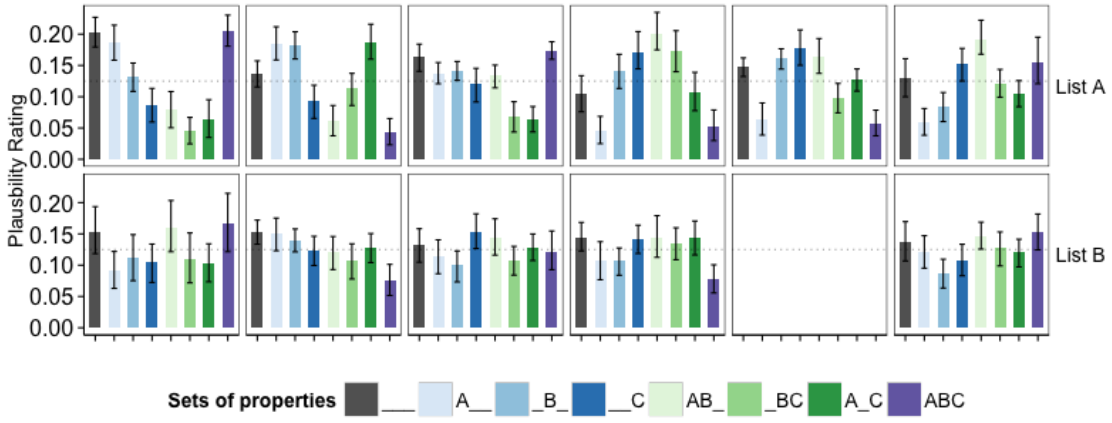


Figure 13: Experiment 2 priors. The horizontal lines indicate a uniform distribution. In the legend, letters denote the presence of a property and blanks denote the absence of a property. Each column corresponds to a different pairing of domains. Domains were paired with respect to a syllogistic form, such that each syllogism could be instantiated by one of two domains in a pair. The rows correspond to the two lists of domains, which correspond loosely to "meaningful" vs. "arbitrary" domain types. It's particularly clear in List A that a range of structured background knowledge was elicited. One domain is absent due to an error in the experiment implementation.

(1989) predicted would have the strongest effects of belief bias. Their argument derives from a feature of the mental models framework which is also present in our probabilistic model framework (see Section **??**).

Background knowledge naturally interfaces with the probabilistic model that reasons over concrete situations populated by objects with properties. In Figure 15, two probabilistic models are compared to the human data. The first model uses a "naive" binomial prior, identical (modulo parameter settings) to that used in Figure 4. This model has two parameter, `n_objects` and `br`, which were fit by maximum likelihood to 6 and 0.30, respectively. The second model uses an empirical prior measured in the prior elicitation section of this experiment. I use the full joint probability distribution specified by participants' responses to questions of plausibility of the various properties co-occurring, described above. This model, then, has only one parameter: `n_objects`, which was fit by maximum likelihood to 6.

The naive probabilistic model fits the data well: $r_{naive} = 0.88$. The probabilistic model that incorporates background knowledge fits even better: $r_{empirical} = 0.91$. The empirical prior model is able to capture many of the qualitative belief effects (e.g. Figure 15, columns 3-6). It should be noted that the prior alone is not doing the work: a model of the conclusion-only using the empirical prior has a mediocre fit: $r_{prior} = 0.59$. Background knowledge interacts with logic to produce the patterns of responses observed.

6.2 Experiment 3

Experiment 3 sought out to see if more subtle background knowledge played an important role in syllogistic reasoning. The modeling work is still ongoing and I will only introduce the results of the experiment here.

6.2.1 Methods

410 participants located in the United States were recruited using Amazon's Mechanical Turk (MTurk) platform. All participants had over a 95% approval rate for MTurk submissions. Participants were compensated for their participation.

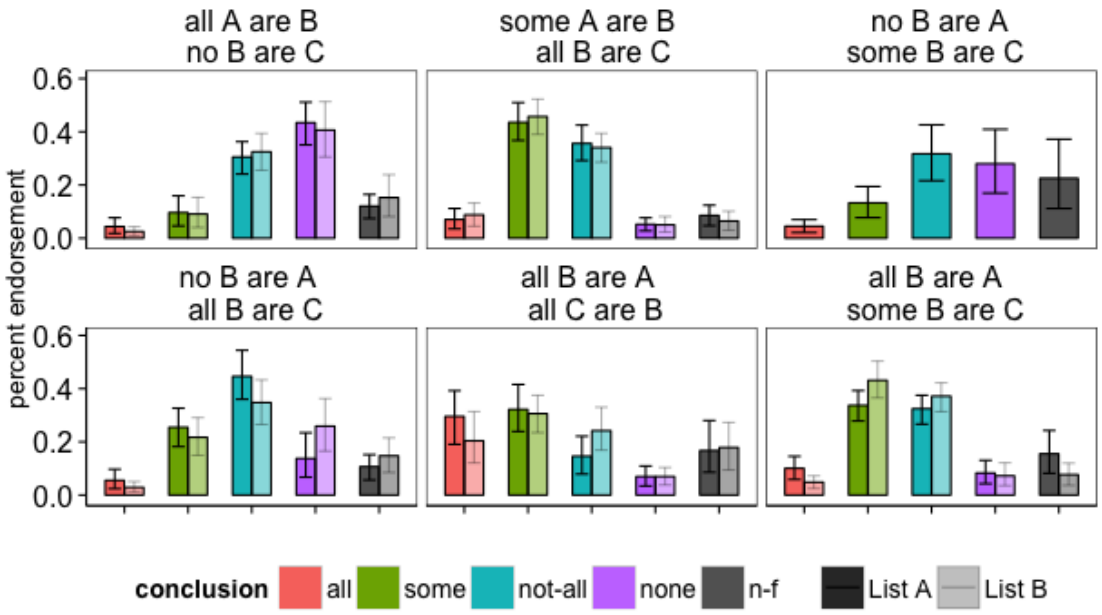


Figure 14: Experiment 2 data. Different opacities correspond to different lists of domains (see text). Reasoning over different domains elicits different reasoning patterns (e.g. bottom row). Error bars correspond to bootstrapped 95% confidence intervals.

Experiment 3 Domains			
Object	A	B	C
Bottle	Green	Empty	Large
Basket	Purple	Soft	Old
Cushion	Colorful	Soft	Small
Microwave	White	Large	Powerful

Table 3: Content domains used in Experiment 3 syllogisms.

6.2.2 Materials

60 participants performed the prior elicitation task only. The remaining 350 participants performed the syllogistic reasoning task.

4 domains were used. The 4 domains were selected from a large set of 16 used in pilot work. The domains were selected to elicit the largest variability in prior elicitations.

In the prior elicitation task, participants were asked to imagine they had come across one of the objects listed in Table 3. They were then asked to rate the likelihood of that object having the properties listed. All possible combinations of the the presence and absence of the 3 properties were listed alongside a slider bar ranging from “Very unlikey" to “Very likely".

Participants performing the syllogistic reasoning task were shown a random subset of 4 syllogisms from the total of 16 syllogisms of a particular term ordering. The term ordering used in this experiment was the same as in Experiment 1:

A – B
B – C
—
A – C

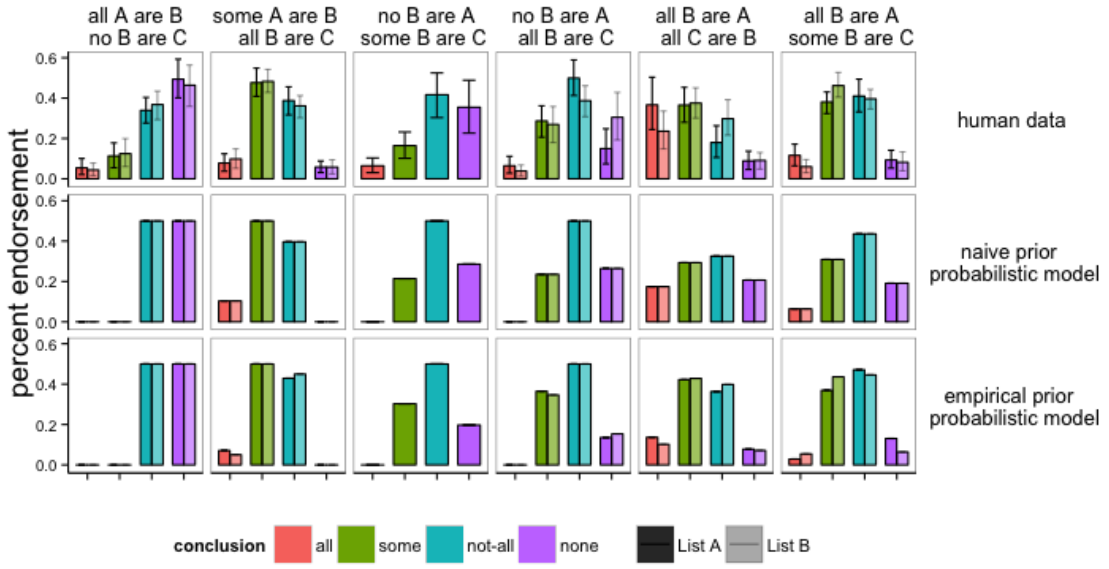


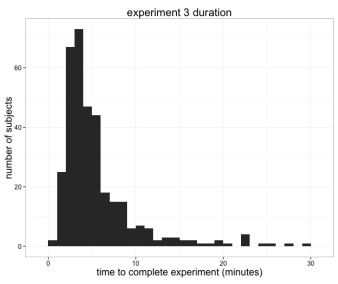
Figure 15: Experiment 2 data and predictions from two models. Different opacities correspond to different lists of domains (see text). Reasoning over different domains elicits slightly different reasoning patterns. The model that uses a naive prior (middle row) is unable to capture this subtlety, while the model that incorporates structured background knowledge (bottom row) is able to capture many of the subtle effects. Error bars correspond to bootstrapped 95% confidence intervals.

6.2.3 Procedure

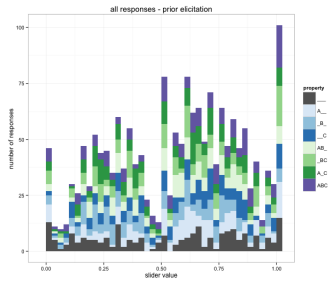
The procedure was same as for the previous experiments[9].

6.2.4 Results

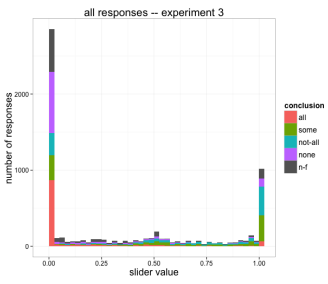
The distribution of time taken on the experiment is shown in Figure 16(a)) with mean 6.2 (median 4.8) and standard deviation of 5.0 minutes. 14 subjects were excluded for taking over 16.2 minutes (mean + two standard deviations) to complete the task, leaving a total of 336 subjects.



(a) Time to complete syllogistic reasoning experiment.



(b) Prior elicitation slider values



(c) Syllogistic reasoning slider values.

Figure 16: Experiment 3 raw data

[9]The experiment can be viewed at: `http://stanford.edu/~mtessler/syllogism02/syllogism02-4f_4.html`

6.2.4.1 Prior elicitation

The distribution of raw slider values for the prior elicitation experiment is shown in Figure 16(b). Compare this with the distribution in Experiment 2 (Figure 12(b)). Experiment 2 used content with obvious correlational structure (e.g. religious people and priests, Figure 13.). Experiment 3 used more banal domains (e.g. baskets: purple, soft, and old) yet still elicited a range of prior distributions (Figure 17).

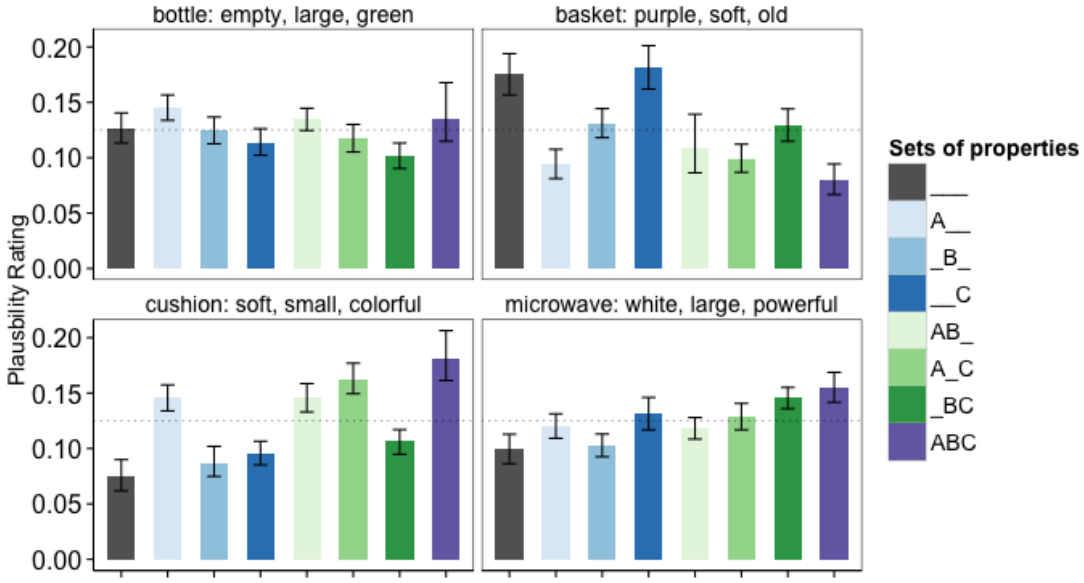


Figure 17: Experiment 3 priors. The horizontal lines indicate a uniform distribution. In the legend, letters denote the presence of a property and blanks denote the absence of a property. Each plot corresponds to a different domain of common objects with common properties.

6.2.4.2 Syllogistic reasoning

The distribution of raw slider values is shown in Figures 16(c). This is similar to the distribution of raw slider values observed for the other two syllogistic reasoning experiments.

The correlation between the maximum slider value and the “None of the above" slider value across all syllogisms and subjects was $r = -0.85$, suggesting that subjects responses were relatively internally consistent.

The human data is shown in Figure 18. Each data point includes responses from at least 15 participants. There is some variability across the domains, and it will be interesting to see if the variability in responses is replicable and corresponds to differences in background knowledge.

7 Relationship to other theories

A recent meta-analysis carved the space of reasoning theories into three partitions: those based on models or diagrammatic reasoning, those based on formal logical rules, and those based on heuristics (Khemlani & Johnson-Laird, 2012). There is another partition we can consider. In one dimension, theories are based on the direct application of derivation rules—be they heuristic or logical—or they are based on the construction of concrete representations or models. In another dimension, theories may take as fundamental: deductive validity or probabilistic support. This theoretical partitioning places the probabilistic reasoning models presented here in a previously unexplored quadrant of the two-dimensional theoretical space described: the model considers probabilistic reasoning over concrete situations.

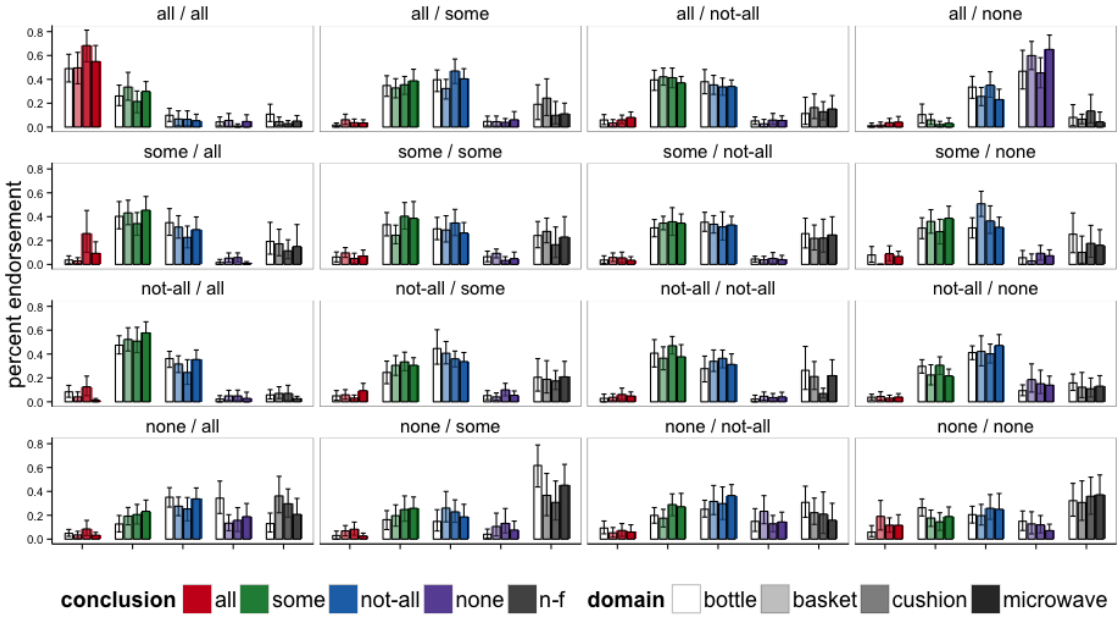


Figure 18: Experiment 3 data broken up by content of each syllogism. Different opacities correspond to different content domains (see text). Error bars correspond to bootstrapped 95% confidence intervals.

7.1 Mental Models

Mental Models Theory (MMT) was offered to capture the intuition that people are able to reason about sets of things explicitly and with respect to context by constructing mental representations of individuals over which to reason. The situations described in the probabilistic models presented here are analogous to mental models. To address the problem of determining which models come into existence, however, MMT relies on a number of complex heuristics. By contrast, we derive a distribution over models (or situations) from natural language semantics and pragmatics, with no further assumptions.

One of the major contributions of the Mental Models Theory was the distinction between “single model" and “multiple model" syllogisms (Johnson-Laird & Bara, 1984). This distinction has to do with the number of distinct situations (models) consistent with the premises. All invalid syllogisms are multiple model syllogisms by definition. A few valid syllogisms, however, are “single model" syllogisms, and thus have only one unique situation consistent with the premises. Participants reason about these syllogisms relatively accurately (Khemlani & Johnson-Laird, 2012).

The probabilistic model described in this article gives a quantitative account of the “single" / “multiple" model distinction. Situations consistent with the premises are derived by a generative process conditioned on the truth of the premises. The “uniqueness" of situations described by MMT is with respect to the possible conclusions[10]. Not only does the probabilistic model make the “single" / “multiple" model distinction, but also makes predictions about the full distribution of conclusions (i.e. which conclusion will be favored first, second, third, fourth). This is something that the basic formulation of MMT does not capture.

The most current account (to my knowledge) of Mental Models Theory of belief bias relies on simultaneous evaluation of logic and believability (Newstead et al., 1992). People are said to construct models consistent with the premises and if the model produces a conclusion consistent with one’s prior beliefs, the search through “model space" ends. If the conclusion is unbelievable (with respect to prior knowledge), people will search for alternative models consistent with the premises.

The probabilistic generative model presented here does not need to posit a particular mechanism of processing. Instead, it says that the proportion of responses for each conclusion will depend on the *a priori*

[10]That is, 2 As are Bs is considered the same as 3 As are Bs, if the same quantifier relations (i.e. some and/or all) in both situations

believability of the conclusion. The models (or situations) are *generated* from prior knowledge. Thus, believable conclusions will be more likely to be produced by virtue of the fact that the corresponding situations are more likely to be generated.

7.2 Probability Heuristics

Chater and Oaksford (1999) introduced the Probability Heuristic Model (PHM) which derives a set of probabilistic rules for syllogistic reasoning. The PHM then augments these probabilistic rules with a complex set of heuristics (for example, informative-conclusion heuristics) to account for what I described above as "pragmatic effects" (i.e. *all* is preferred over *some* when both are valid). The model presented here differs from the PHM in at least two respects. First, the probabilistic "rules" emerge from the semantics of quantifiers by reasoning about situations. Second, inferences are strengthened with respect to informativity by virtue of recursive (or, pragmatic) reasoning. This gives rise to many of the same effects without having to postulate heuristics *de novo*.

Another shortcoming of the PHM is that it engages only with the superficial features of the syllogism. The model bases its predictions only on the quantifiers of the syllogism, and thus treats many sets of logically distinct syllogisms as identical (e.g. see Figure 3: pairs of symmetric, off-diagonal entries). The data presented in Experiments 1 & 3, as well as many previous studies, demonstrate that people engage with the syllogistic arguments in rich ways, driven by both the logic of the quantifiers and the form of the argument.

8 Conclusions

The syllogistic reasoning task involves reading a pair of sentences and producing or evaluating a conclusion. The long literature of syllogistic reasoning provides overwhelming evidence that people do not accord with Aristotelean logic. The gradedness in responses suggests probabilistic support is what underlies reasoning in "formal" reasoning tasks. I have explored a model that explains many different human reasoning patterns in the syllogistic domain by assuming (1) people reason about concrete situations generated probabilistically from prior knowledge and (2) global quantifier interpretation (or, equivalently, full argument interpretation) occurs by way of pragmatic reasoning.

My model is represented in a probabilistic program written in the language of Church. The probabilistic program allows us to model the rich structure of human knowledge. The model considers situations populated by objects with properties. The contents of a particular situation are determined stochastically by sampling. In this way, I derive not only a space of possible situations (as Mental Models Theory has done) but a full distribution.

I have considered the pragmatics of argument interpretation—the problem a reasoner faces when given some sentences. I have incorporated into the model a function called `experimenter`, which represents the participant's theory of the experimenter. This allows the reasoner to arrive at enriched meanings, beyond that which semantics would allow. This has been shown be a fruitful approach in accounting for the reasoning data. In addition to providing enriched meanings to word, the model represents a first step to accounting for experimenter effects in reasoning tasks. By fully specifying the knowledge and the theory of the `reasoner`, the model is able to account for the data well.

It seems that people without formal training in logic treat syllogisms as a sort of puzzle. They incorporate background knowledge into the reasoning process, and consider the arguments as part of a conversation. This resonates with Ong's descriptions of Luria's preliterate people: "To go by your words, they should all be white". Ong (1982) writes:

> "To go by your words" appears to indicate awareness of the formal intellectual structures. A little literacy goes a long way. On the other hand, the chairman's limited literacy leaves him more comfortable in the person-to-person human lifeworld than in a world of pure abstractions: "To go by your words..." It is your responsibility, not mine, if the answer comes out in such a fashion.

References

Barwise, J., & Etchemendy, J. (1994). *Hyperproof*. Stanford, Calif.: CSLI Publications.

Chater, N., & Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. *Cognitive psychology*, *258*, 191–258.

Cherubini, P., Garnham, a., Oakhill, J., & Morley, E. (1998, December). Can any ostrich fly?: some new data on belief bias in syllogistic reasoning. *Cognition*, *69*(2), 179–218. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9894404

Dickstein, L. S. (1978, September). Error processes in syllogistic reasoning. *Memory & cognition*, *6*(5), 537–43. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24203387 doi: 10.3758/BF03198242

Evans, J. S. B. T., Handley, S. J., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

Frank, M. C., & Goodman, N. D. (2012). Quantifying pragmatic inference in language games. *Science*, *336*, 1–9.

Gleick, J. (2011). *The information: A history, a theory, a flood*. Pantheon Books.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church : a language for generative models. *Uncertainty in Artificial Intelligence*.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–84.

GRE Reading Comprehension Practice Test 01 majortests.com. (n.d.). http://www.majortests.com/gre/reading_comprehension_test01. (Accessed: 2014-06-23)

Horn, L. R. (1989). *A natural history of negation*. Chicago: University of Chicago.

Johnson-Laird, P. N., & Bara, B. G. (1984, February). Syllogistic inference. *Cognition*, *16*(1), 1–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6540648

Johnson-laird, P. N., & Steedman, M. (1978). The Psychology of Syllogisms. *Cognitive psychology*, *10*, 64–99.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological bulletin*, *138*(3), 427–57.

Lukasiewiecz, J. (1951). *Aristotle's syllogistic*. London: Oxford University Press.

Newstead, S. E., & Evans, J. S. (1993). Mental models as an explanation of belief bias effects in syllogistic reasoning. *Cognition*, *46*, 93–97.

Newstead, S. E., Pollard, P., Evans, J. S., & Allen, J. L. (1992, December). The source of belief bias effects in syllogistic reasoning. *Cognition*, *45*(3), 257–84. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1490324

Oakhill, J., & Garnham, A. (1993). On theories of belief bias in syllogistic reasoning. *Cognition*, *46*, 87–92.

Oakhill, J., Johnson-Laird, P., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, *31*, 117–140.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word*. London: Methuen and Company Ltd.

Pirsig, R. M. (1974). *Zen and the art of motorcycle maintenance: An inquiry into values*. William Morrow and Company.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Roberts, C. (2004). Information structure in discourse. *Semanatics and Pramatics*(5), 1-69.

Roberts, M. J., Newstead, S. E., & Griggs, R. A. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking & Reasoning*, *7*(2), 173–204.

Wilkins, M. C. (1928). *The Effect of Changed Material on Ability to do Formal Syllogistic Reasoning*.