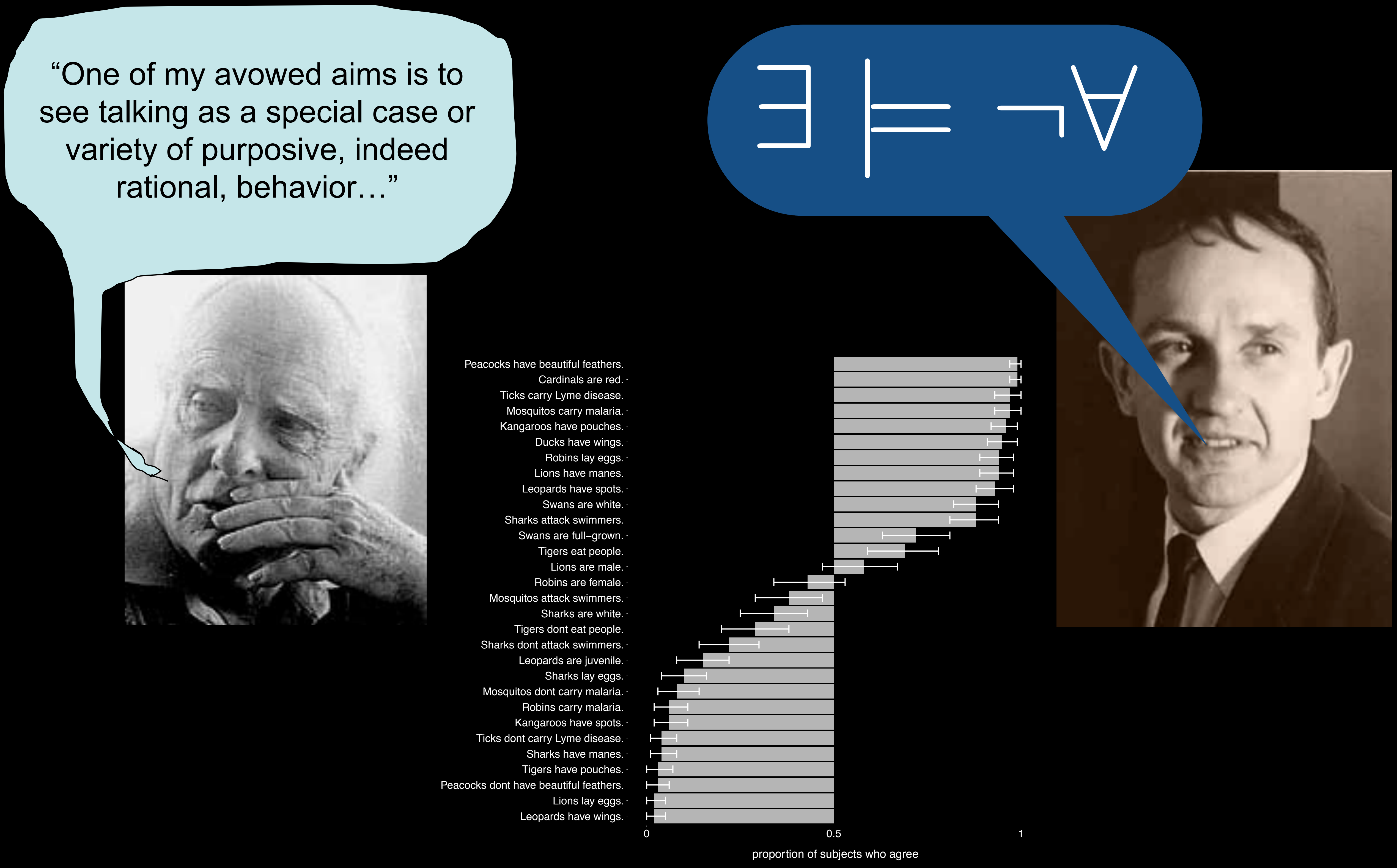
Computation & Cognition Lab

ngoodman@stanford.edu

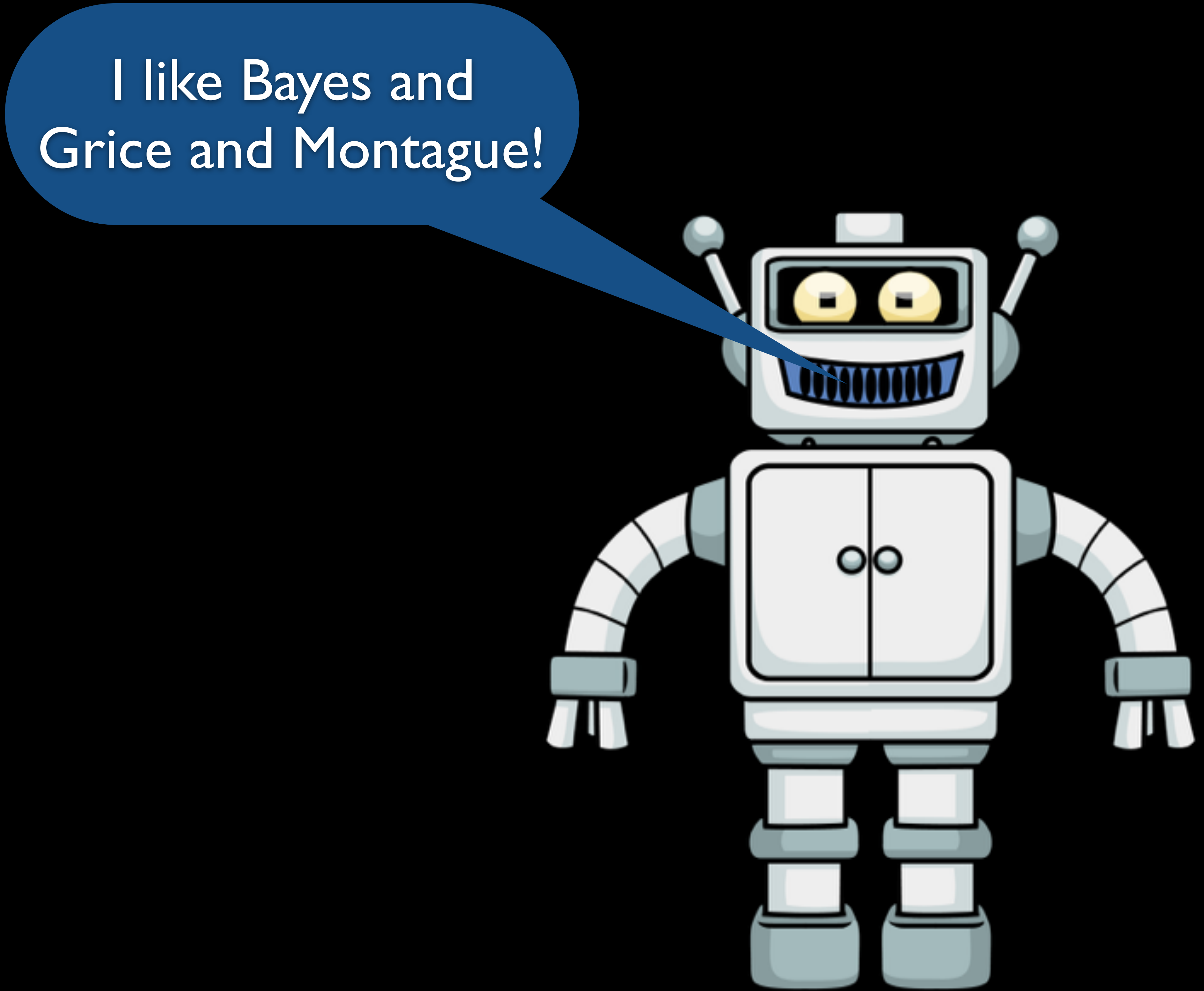
Some thoughts on and examples of How Language Works

Noah D. Goodman
Stanford University

KogWis 2018, Darmstadt
September 05, 2018



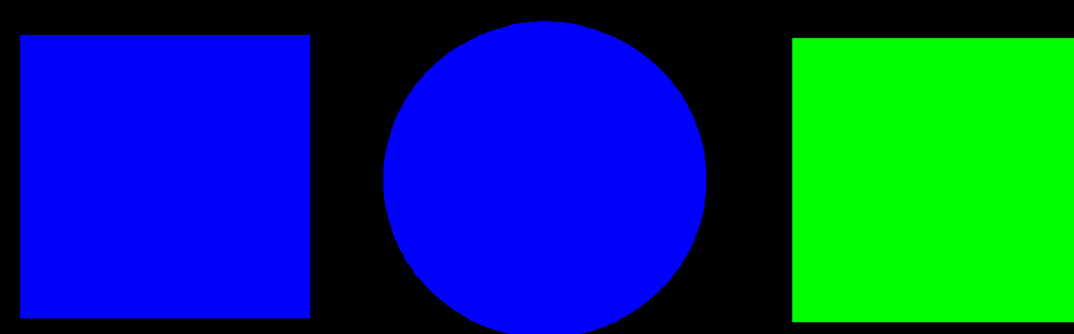
How can we make quantitative formal models of human language use?



How can we make robots talk better?

Reference games

Speaker: Imagine you are talking to someone and want to refer to the middle object. Would you say “blue” or “circle”?



Listener: Someone uses the word “blue” to refer to one of these objects. Which object are they talking about?

Frank and Goodman (2012)

Social reasoning

utterance

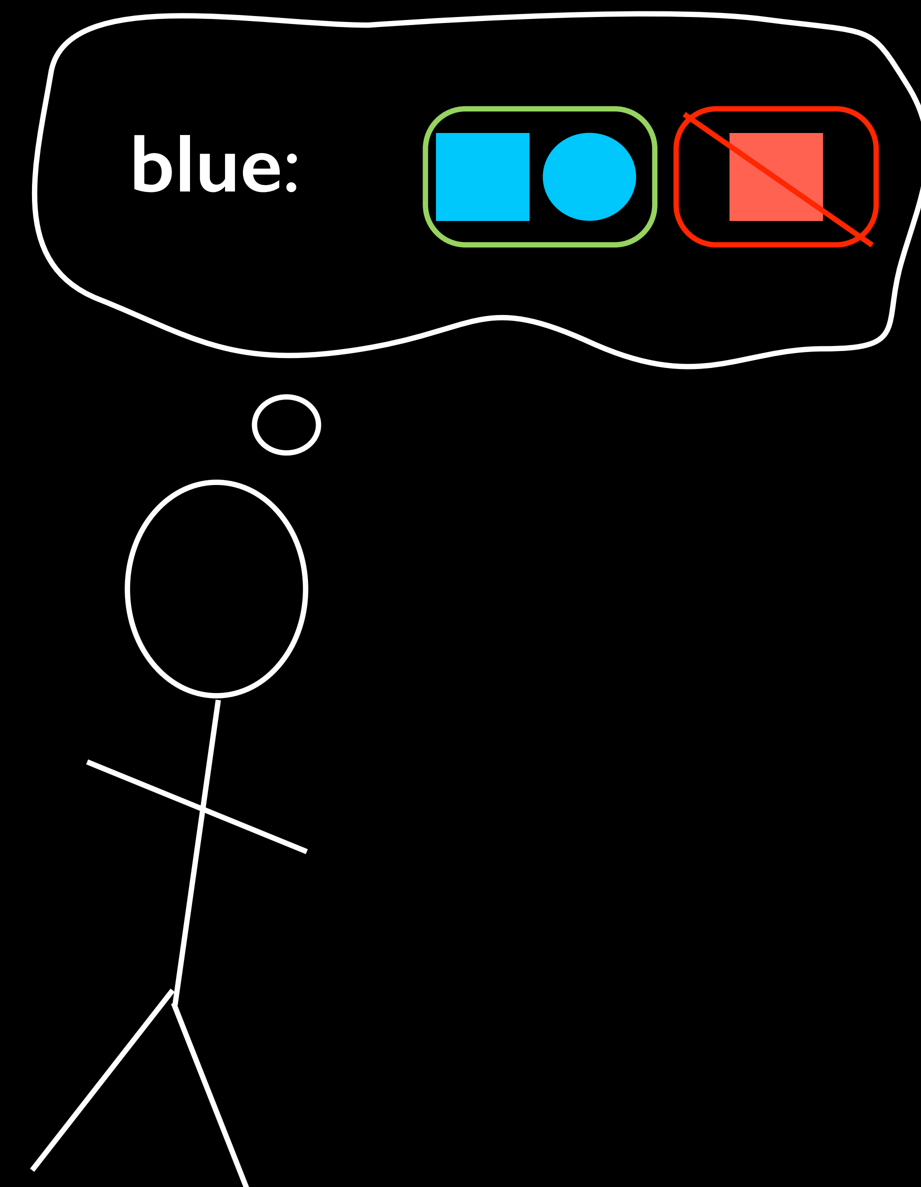
objects in context

$$L_0(o|u, C) \propto P_C(o)\mathcal{L}(o, u)$$

target object

meaning:
is *u* true of *o*?

$$\mathcal{L}(o, u) \in \{0, 1\}$$

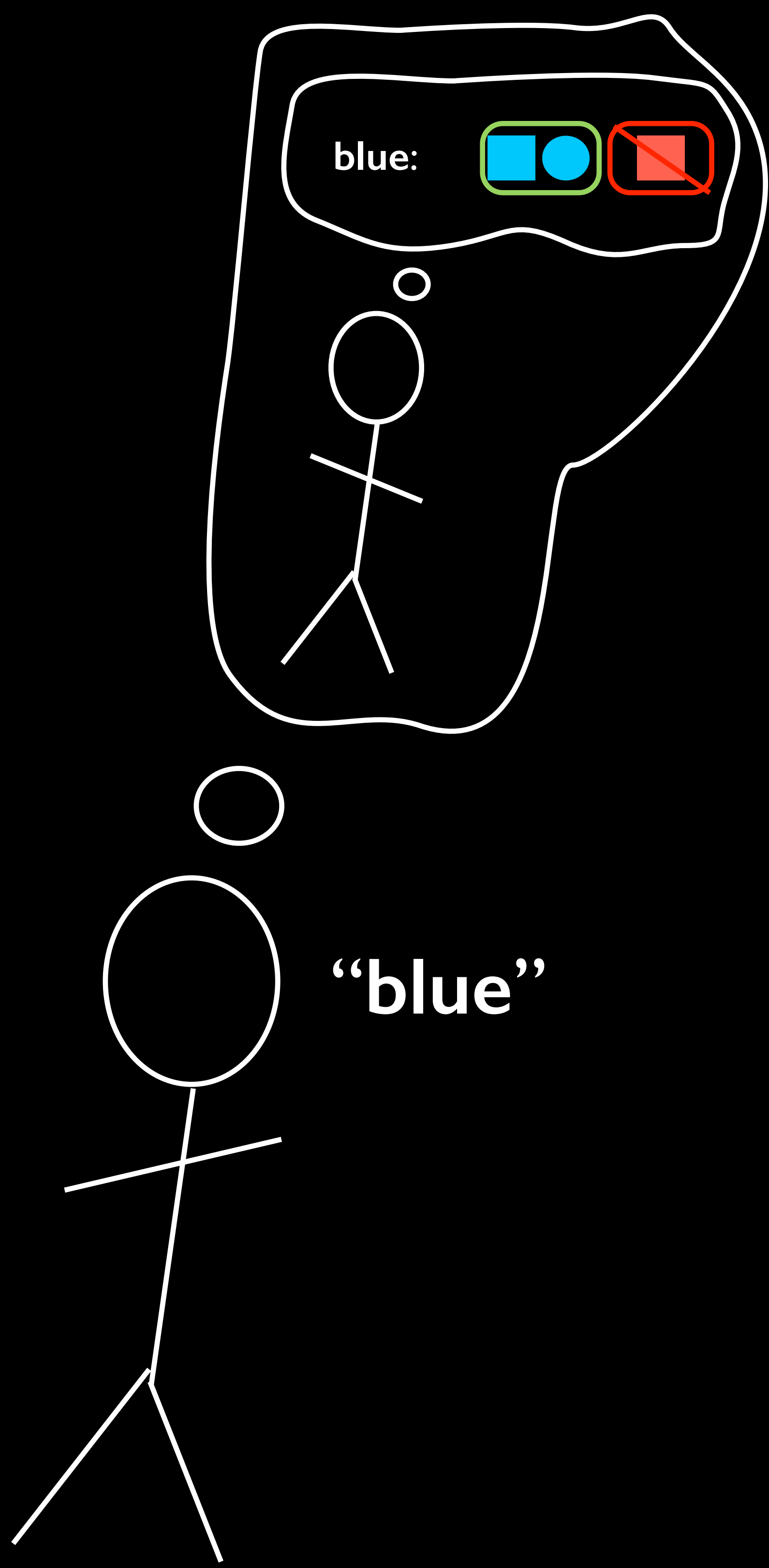


Social reasoning

$$L_0(o|u,C) \propto P_C(o)\mathcal{L}(o,u)$$

$$S_1(u|o,C) \propto e^{\alpha U(u,o,C)}$$

$$U(u,o,C) = -\ln L_0(o|u,C) - \mathrm{cost}(u)$$



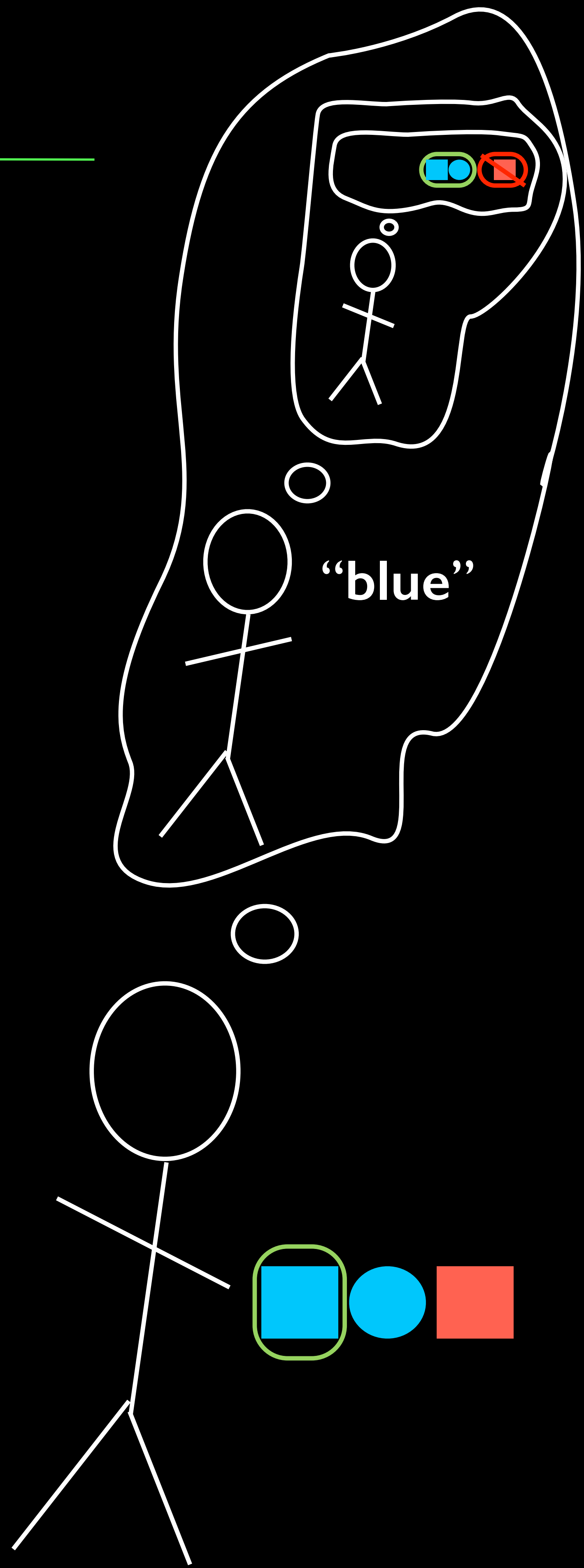
Social reasoning

$$L_0(o|u,C) \propto P_C(o)\mathcal{L}(o,u)$$

$$S_1(u|o,C) \propto e^{\alpha U(u,o,C)}$$

$$U(u,o,C) = -\ln L_0(o|u,C) - \mathrm{cost}(u)$$

$$L_1(o|u) \propto P_C(o)S_1(u|o)$$



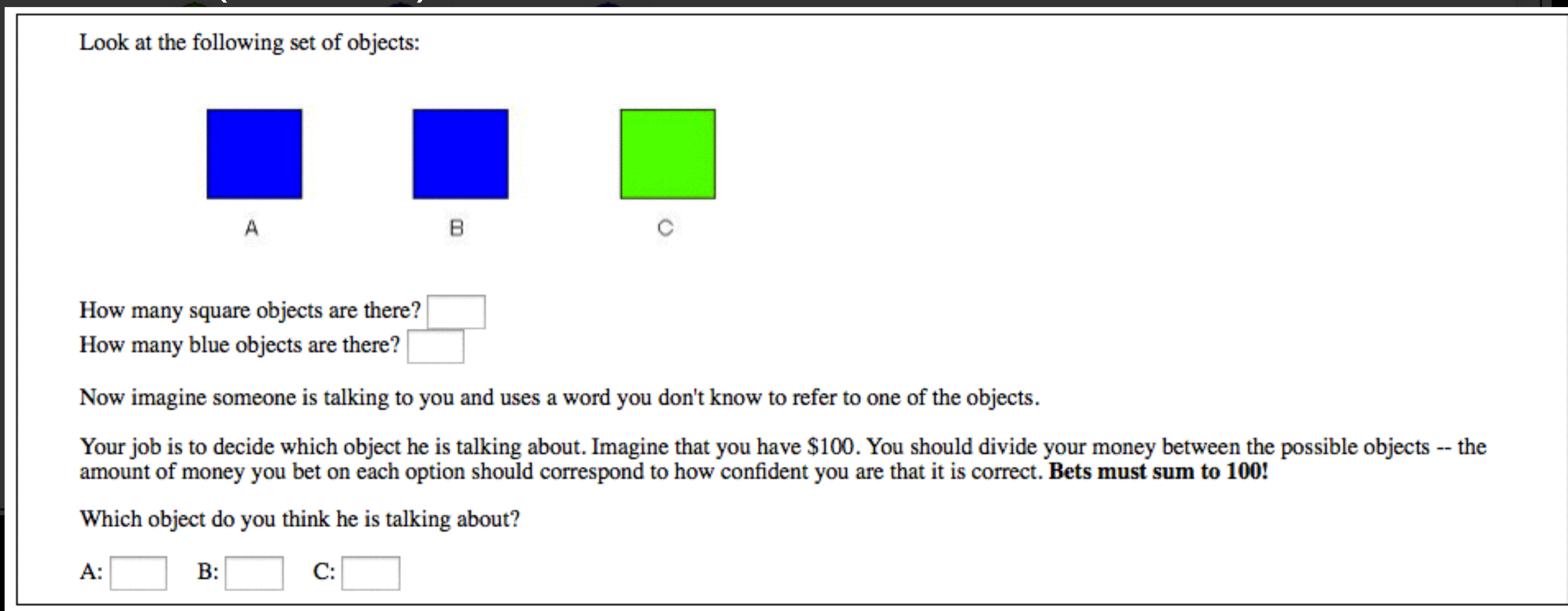
All models implemented as probabilistic programs in WebPPL.

Experiment

Speaker (N=206)

Listener (N=263)

Prior (N=276)



Look at the following set of objects:

A B C

How many square objects are there?

How many blue objects are there?

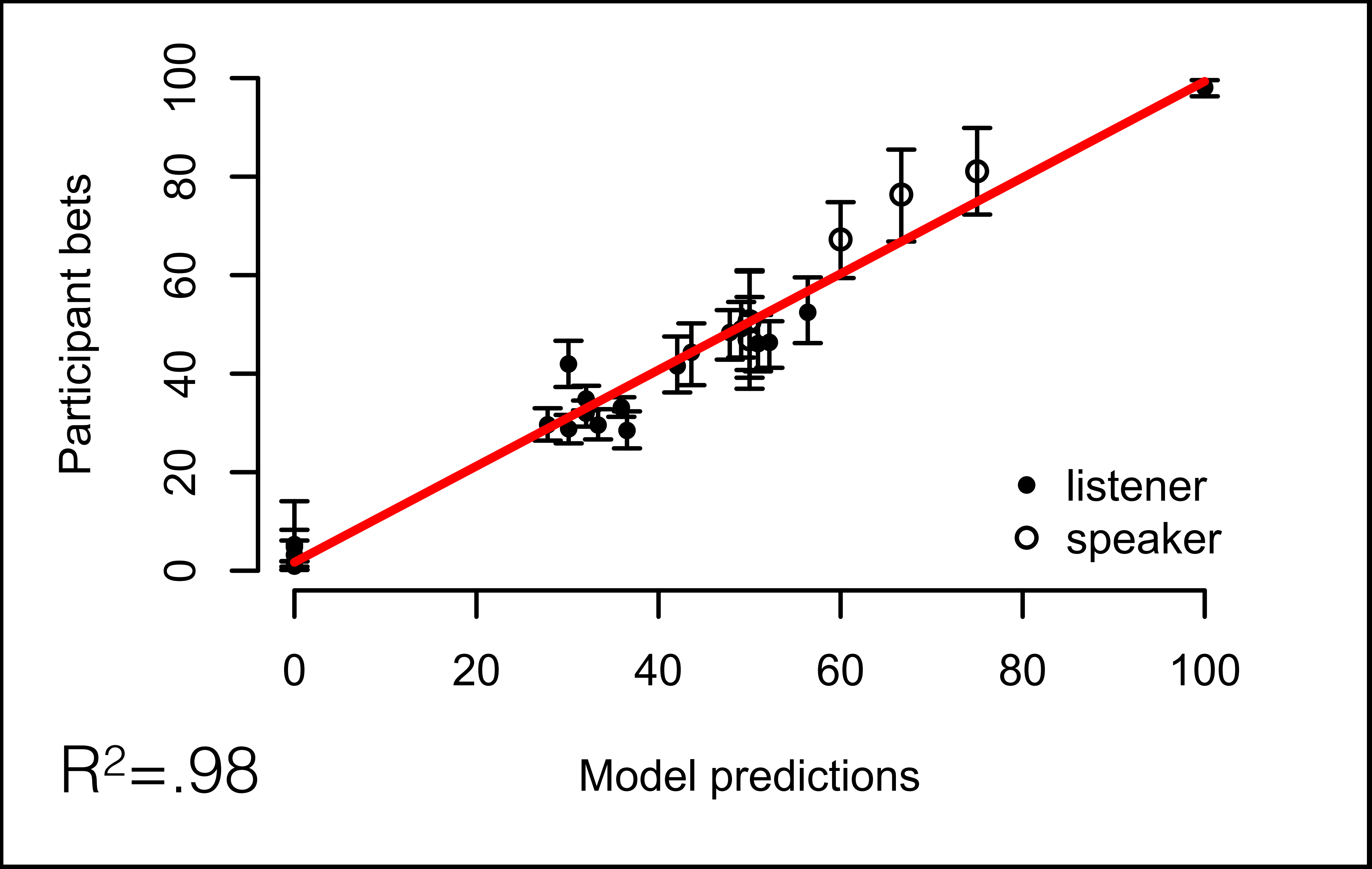
Now imagine someone is talking to you and uses a word you don't know to refer to one of the objects.

Your job is to decide which object he is talking about. Imagine that you have $100. You should divide your money between the possible objects -- the amount of money you bet on each option should correspond to how confident you are that it is correct. **Bets must sum to 100!**

Which object do you think he is talking about?

A: B: C:

Results



Participant bets
0
20
40
60
80
100
0
20
40
60
80
100
Model predictions
listener
speaker
R^2=.98

Rational speech acts

- Rational Speech Act models
  - Understanding is a Bayesian inference using context and model of speaker.
  - Production is a rational(ish) decision with goals to be informative, efficient, etc.
  - This extends to many language understanding phenomena, Goodman & Frank 2016 for a review.

Nonliteral language

But soft! What light through yonder window breaks? It is the east, and **Juliet is the sun.**

www.yelp.com › Restaurants › American (New) ▾ Yelp, Inc. ▾
★★★☆☆ Rating: 3 - 76 reviews - Price range: $$
Oh and need I mention that **it took a million years to** get the waiter at our table to order.

My phone is a hundred years old.

I told you a thousand times already.

A latte at that hipster place costs ten dollars.

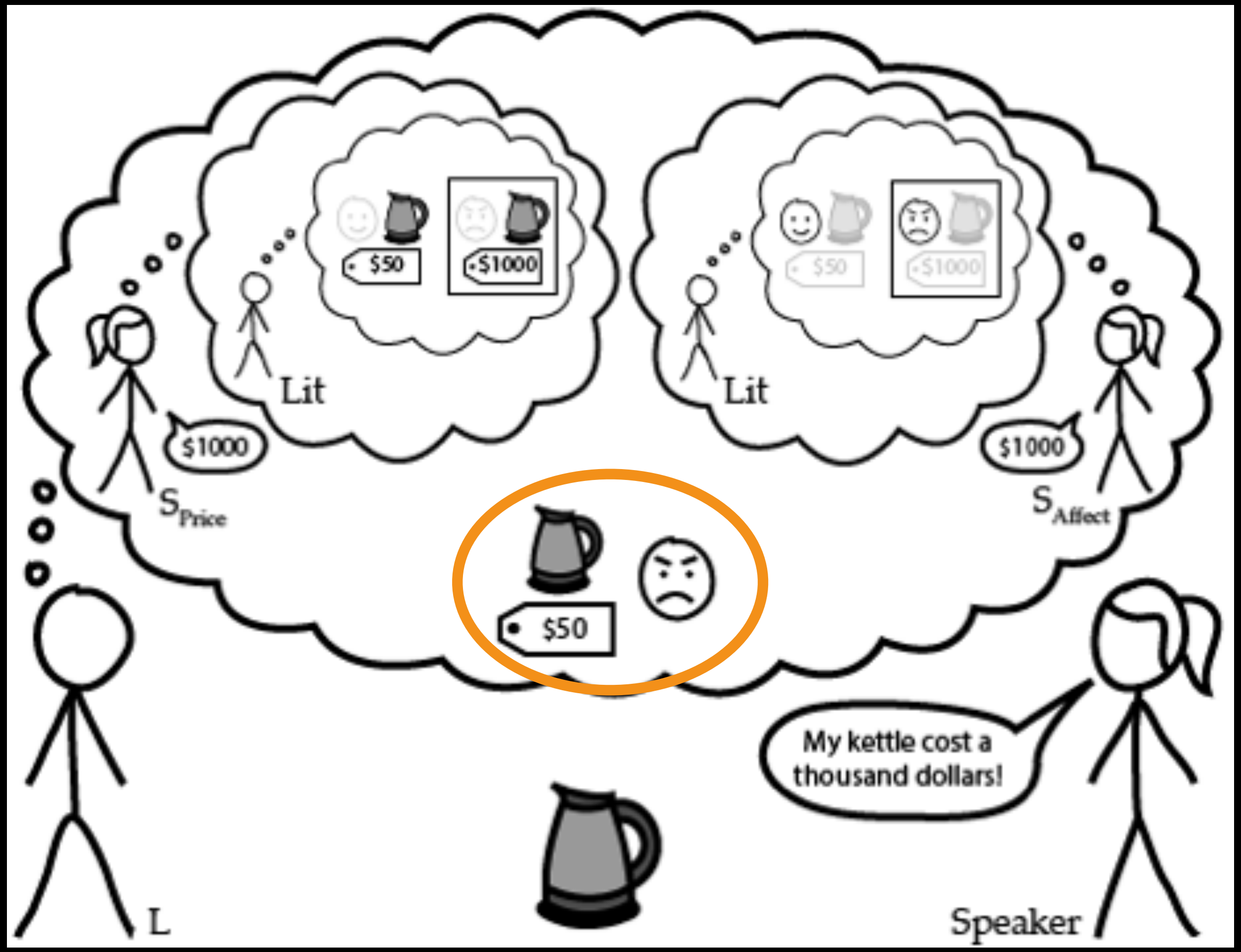
Nonliteral language

- Puzzle: understanding must start in conventional, literal meaning; but language is often used in ways that are literally* false.
  - Why is exaggeration not lying?
- The simplest RSA model can't predict non-literal usage.

*I mean "literally" literally, not non-literally as "figuratively".

Model

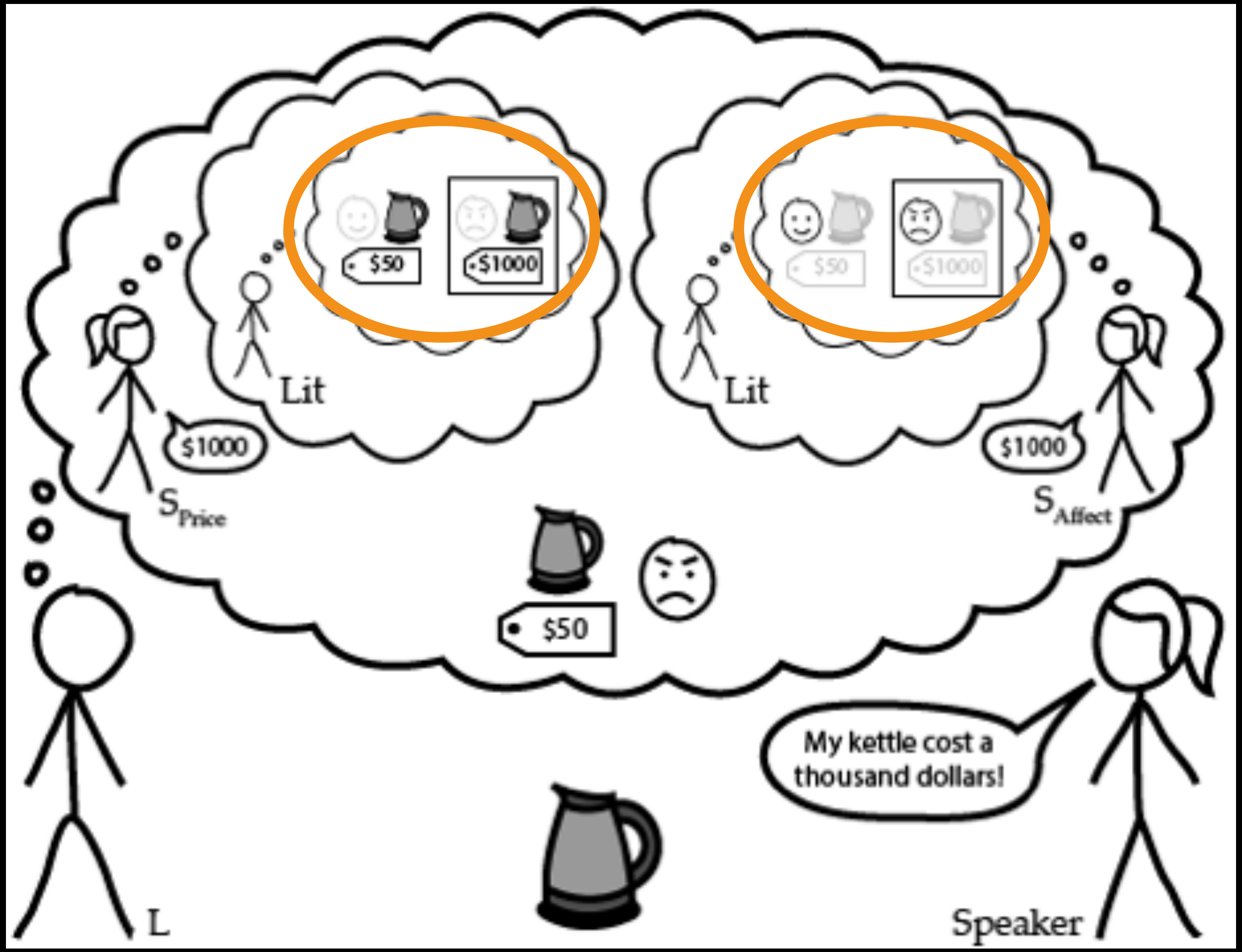
- Non-literal interpretation often conveys information about opinion, beyond the objective world state.
  - Extend the "world" of interpretation to have these opinion dimensions. w=(state, affect)



Kao, Wu, Bergen, Goodman (2014)

Model

- The speaker may have a goal of conveying only opinion, and not care about the actual state.
  - Allow listener to reason about the topic of conversation (QUD).



Kao, Wu, Bergen, Goodman (2014)

Model

- A QUD is a function mapping worlds to *relevant* information.

$$Q_{\text{affect}}((s,a)) = a$$
$$Q_{\text{state}}((s,a)) = s$$
$$Q_{\text{both}}((s,a)) = (s,a)$$

Kao, Wu, Bergen, Goodman (2014)

Model

- A QUD is a function mapping worlds to *relevant* information.
- A speaker aims to be informative about the projected world.

$$S_1(u|w,Q) \propto e^{U(u,w,Q)}$$

$$U(u,w,Q) = -\ln \boxed{\sum_{w'} \delta_{Q(w)=Q(w')} L_0(w'|u)} - \mathrm{cost}(u)$$

Kao, Wu, Bergen, Goodman (2014)

Model

- The pragmatic listener isn't sure which speaker she is hearing from.
- She jointly infers the interpretation and the QUD.

$$S_1(u|w,Q) \propto e^{U(u,w,Q)}$$

$$U(u,w,Q) = -\ln \sum_{w'} \delta_{Q(w)=Q(w')} L_0(w'|u) - \text{cost}(u)$$

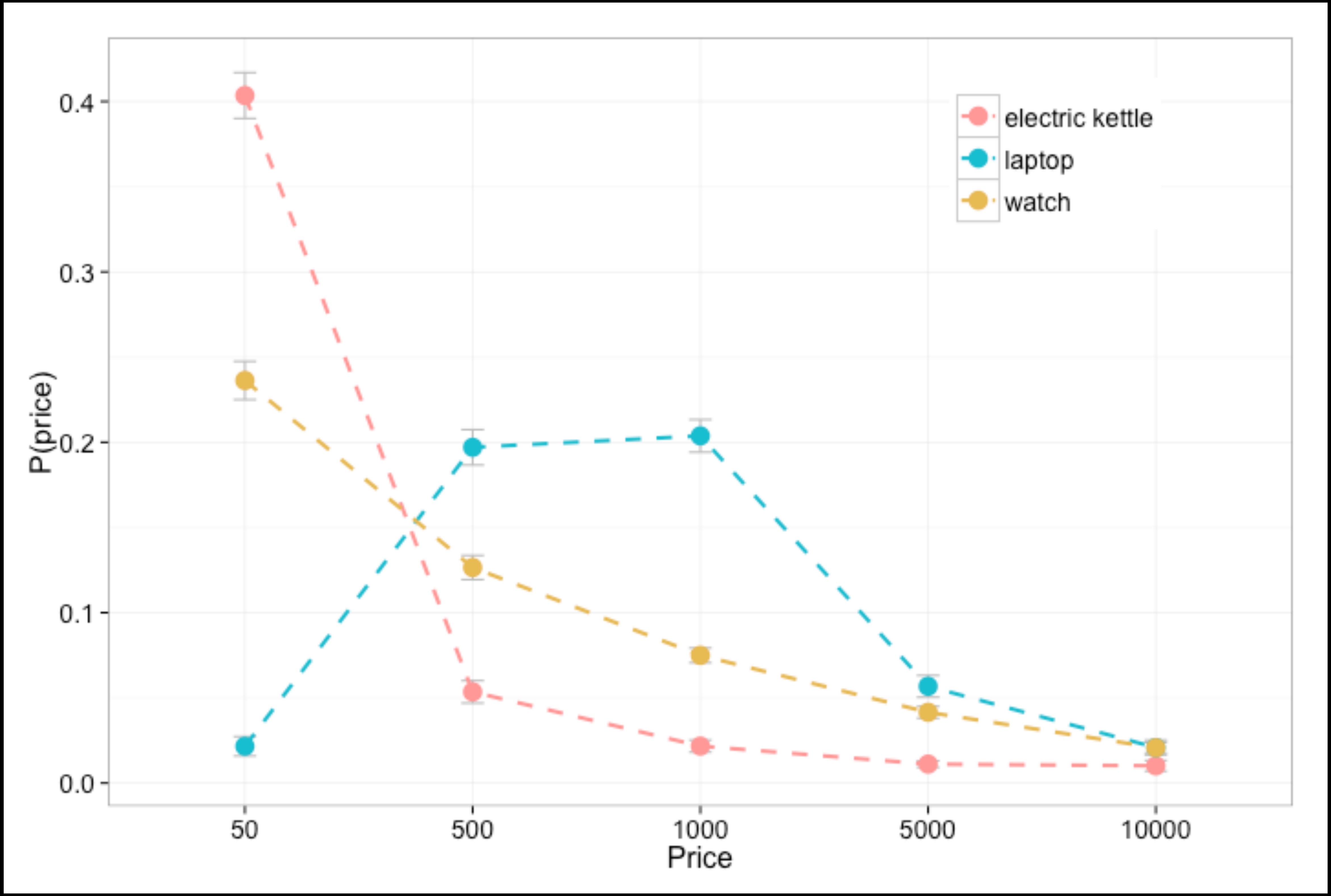
$$L_1(\boxed{w,Q}|u) \propto P(w)P(Q)S_1(u|w,Q)$$

Kao, Wu, Bergen, Goodman (2014)

Price priors

30 participants

Price priors



Affect priors

Eric bought a new ***electric kettle***.

It cost 10,000 dollars.

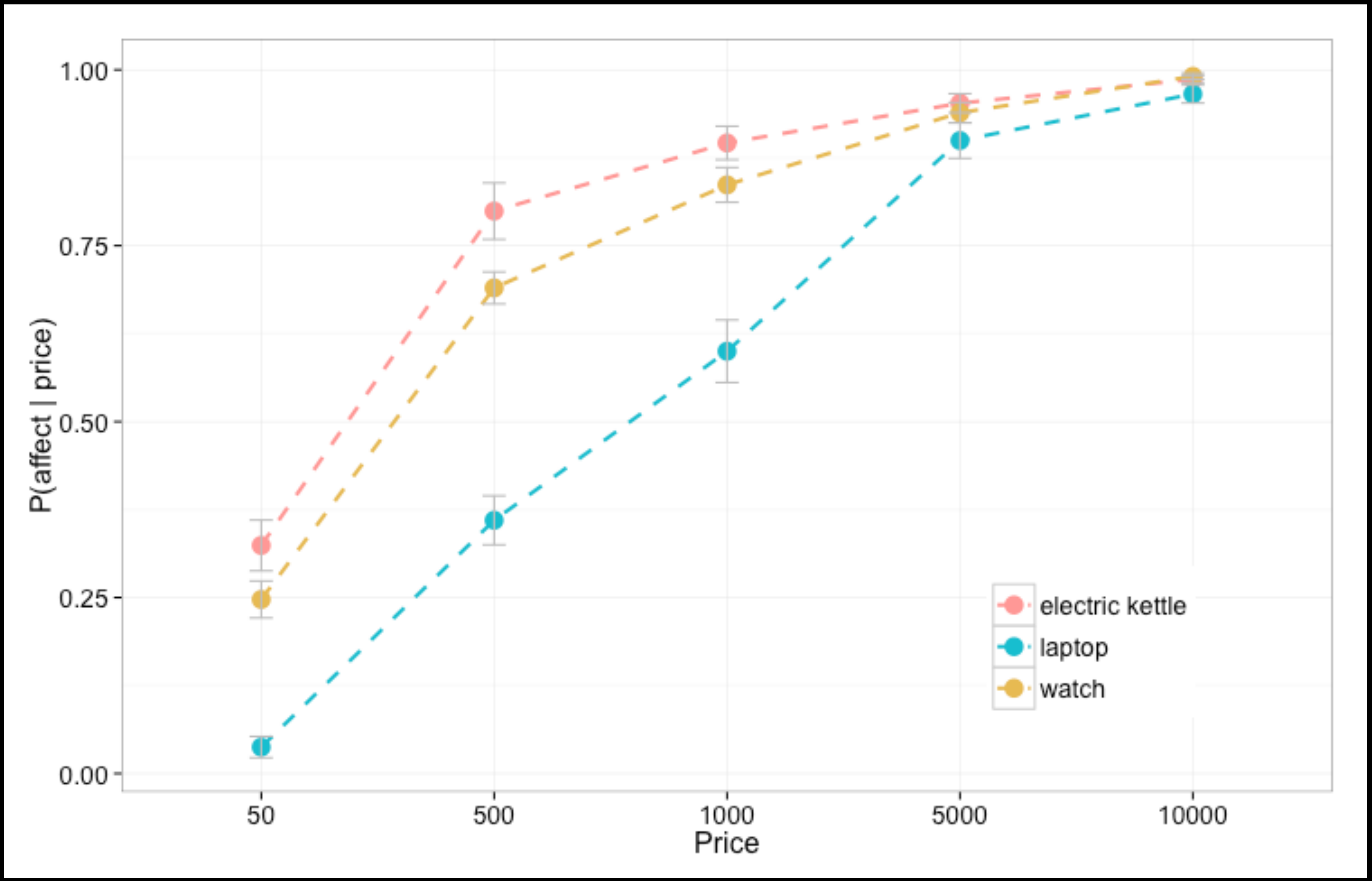
How likely is it that Eric thinks the electric kettle was expensive?

Impossible | Not very likely | Neutral | Very Likely | Absolutely certain

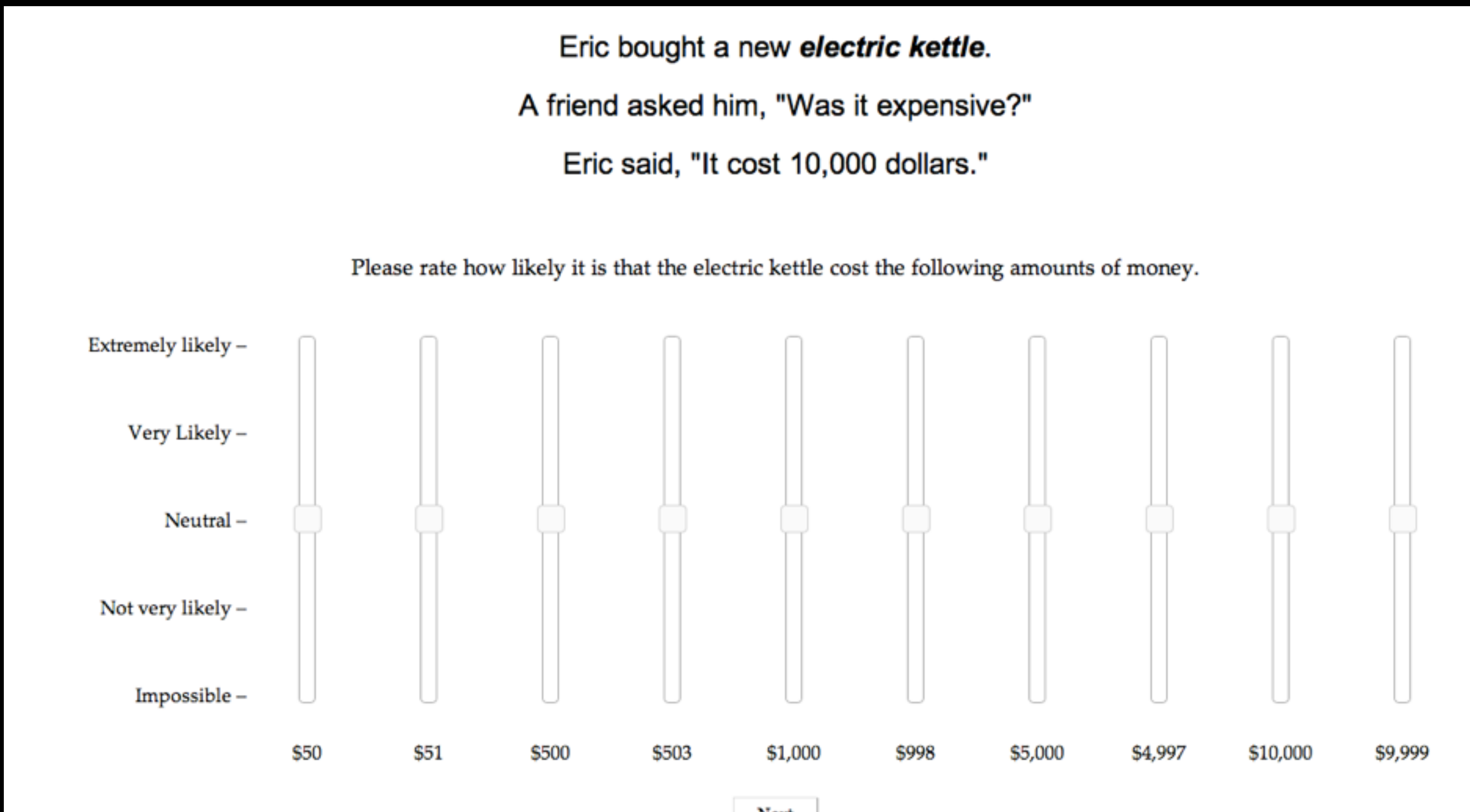
Next

30 participants

Affect priors

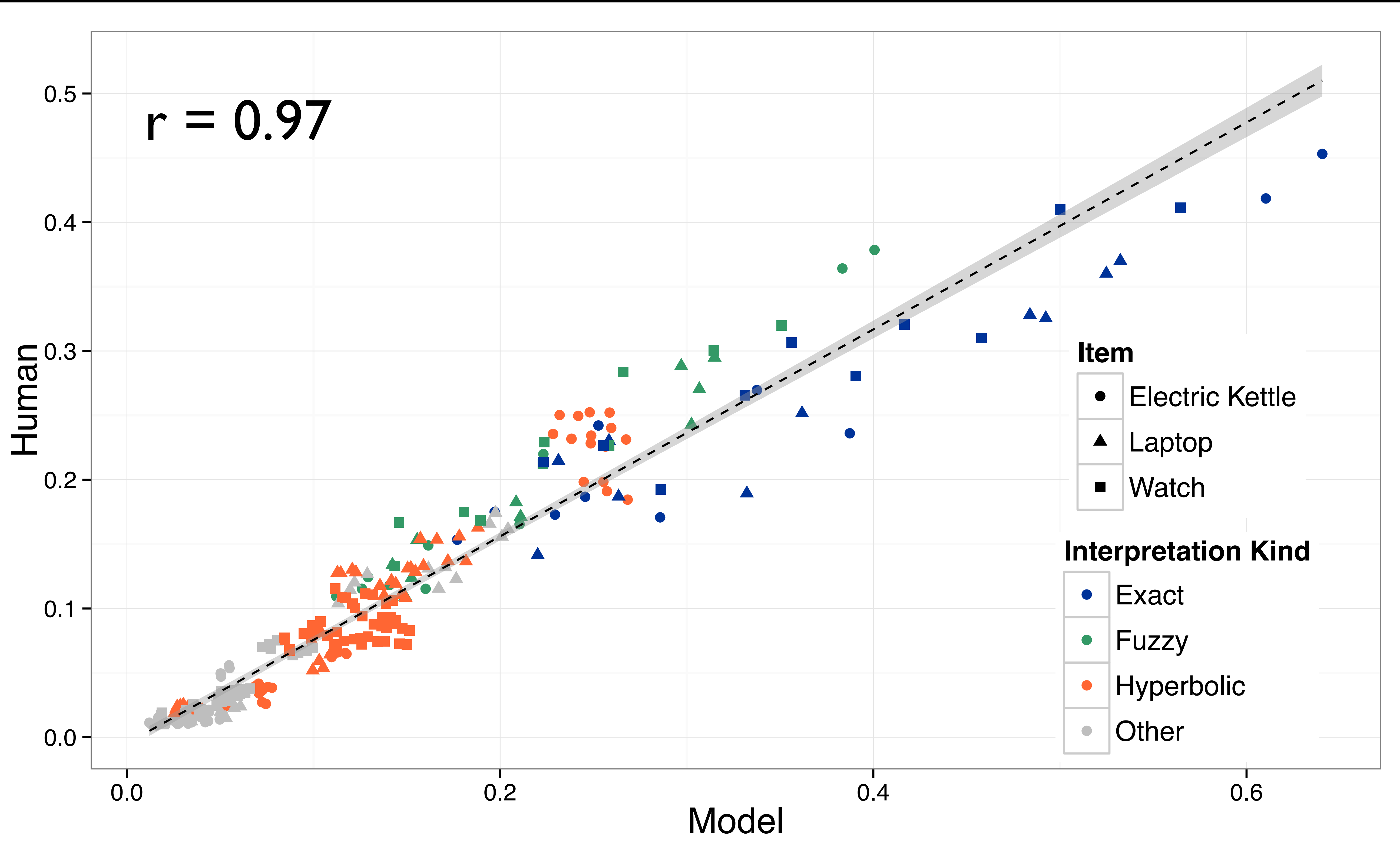


Hyperbole experiment



120 participants

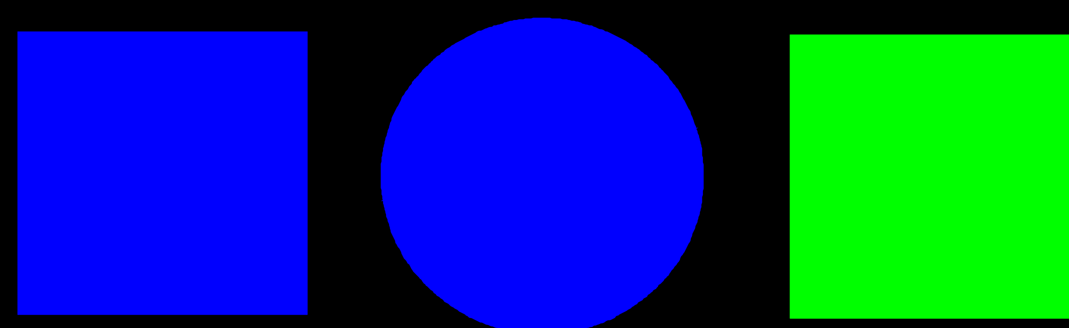
Results



r = 0.97
Human
Model
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
Item
Electric Kettle
Laptop
Watch
Interpretation Kind
Exact
Fuzzy
Hyperbolic
Other

Redundant reference

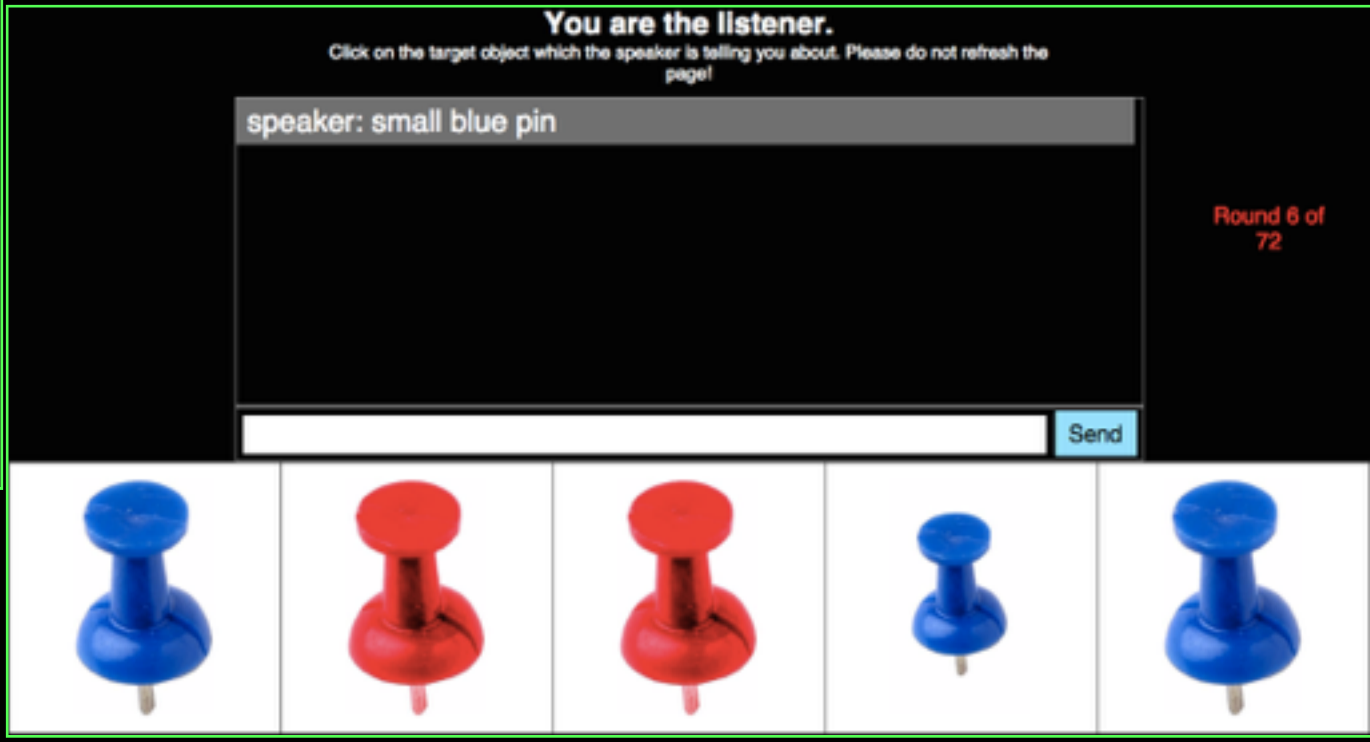
Speaker: Imagine you are talking to someone and want to refer to the right object. What would you say?



“green square”

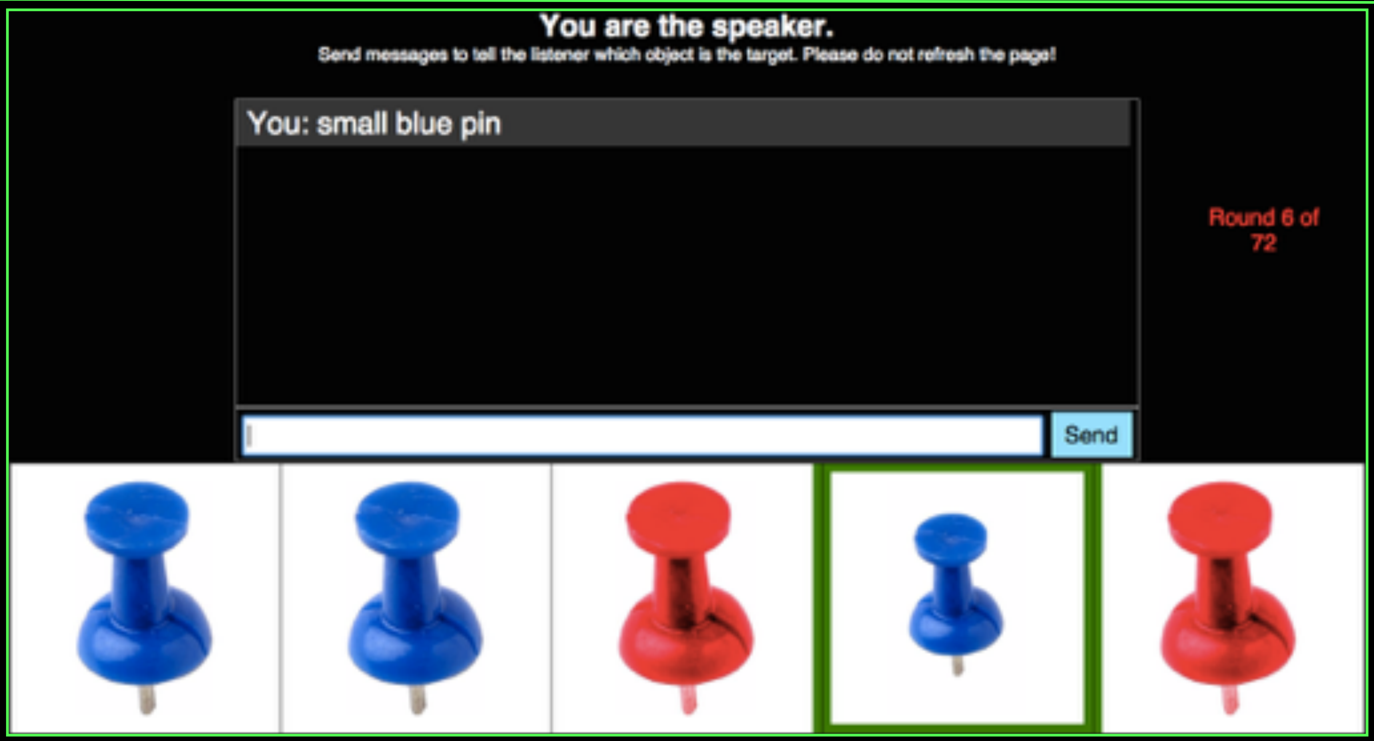
- When given the chance, people often produce redundant information.

Free-production games



- Real-time two player language games with free use of chat window.
- Lots of situated language use data!

RSA with soft semantics



- Basic RSA doesn't explain redundant reference (Cf. Gatt, et al).
- If the modifiers were noisy, redundancy would be a safe bet….

RSA with soft semantics

$$L_0(o|u, C) \propto P_C(o)\mathcal{L}(o, u)$$

meaning: real valued, instead of Boolean $\mathcal{L}(o, u) \in [0, 1]$

- Assume compound modifiers compose by multiplication: $\mathcal{L}(o, u_1 u_2) = \mathcal{L}(o, u_1)\mathcal{L}(o, u_2)$
- Assume some "fidelity" for color and for size, that moderates the truth values:

`meaning("small blue",`  `)`
`=fid(size)*fid(color)`

`meaning("small red",`  `)`
`=fid(size)*(1-fid(color))`

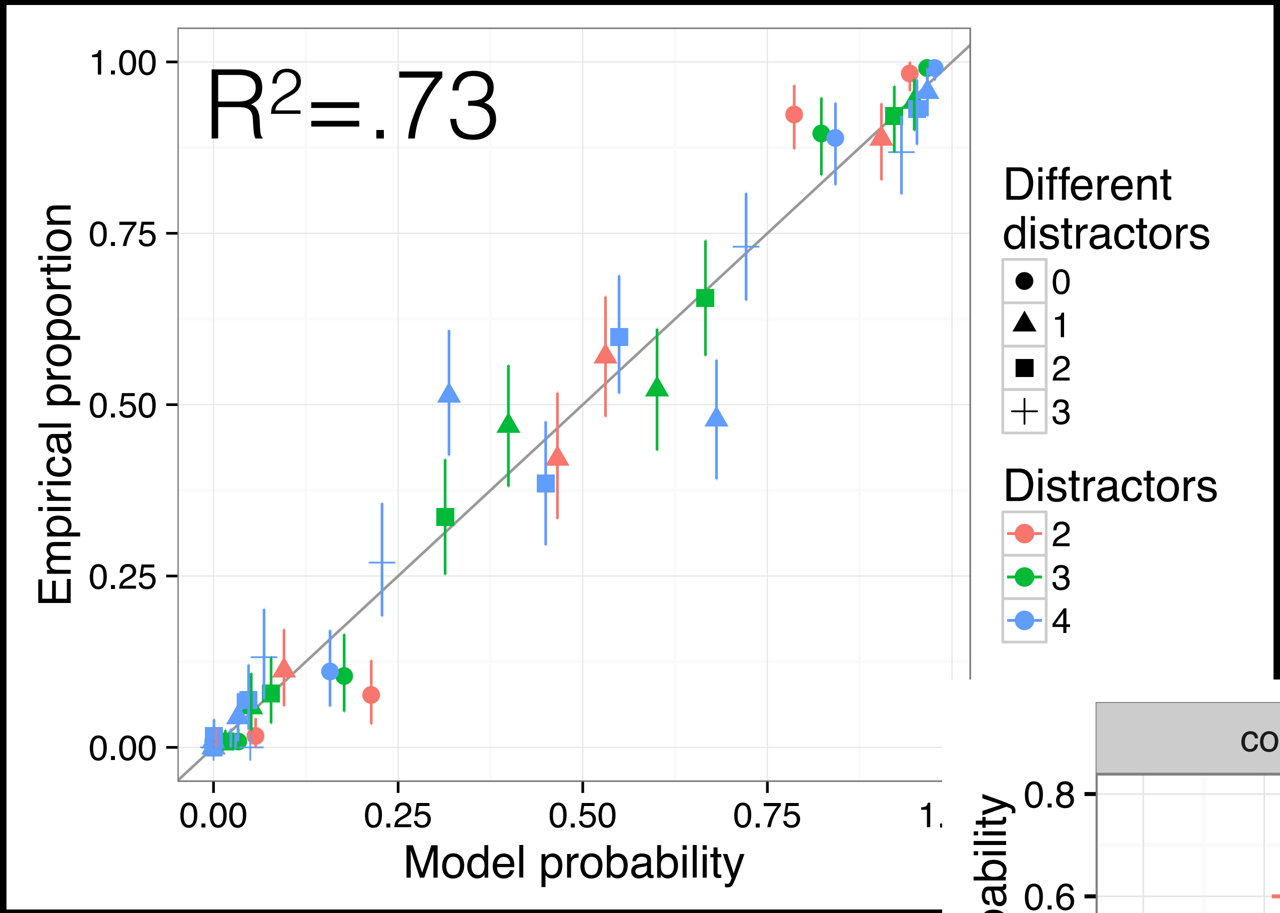
Experiment

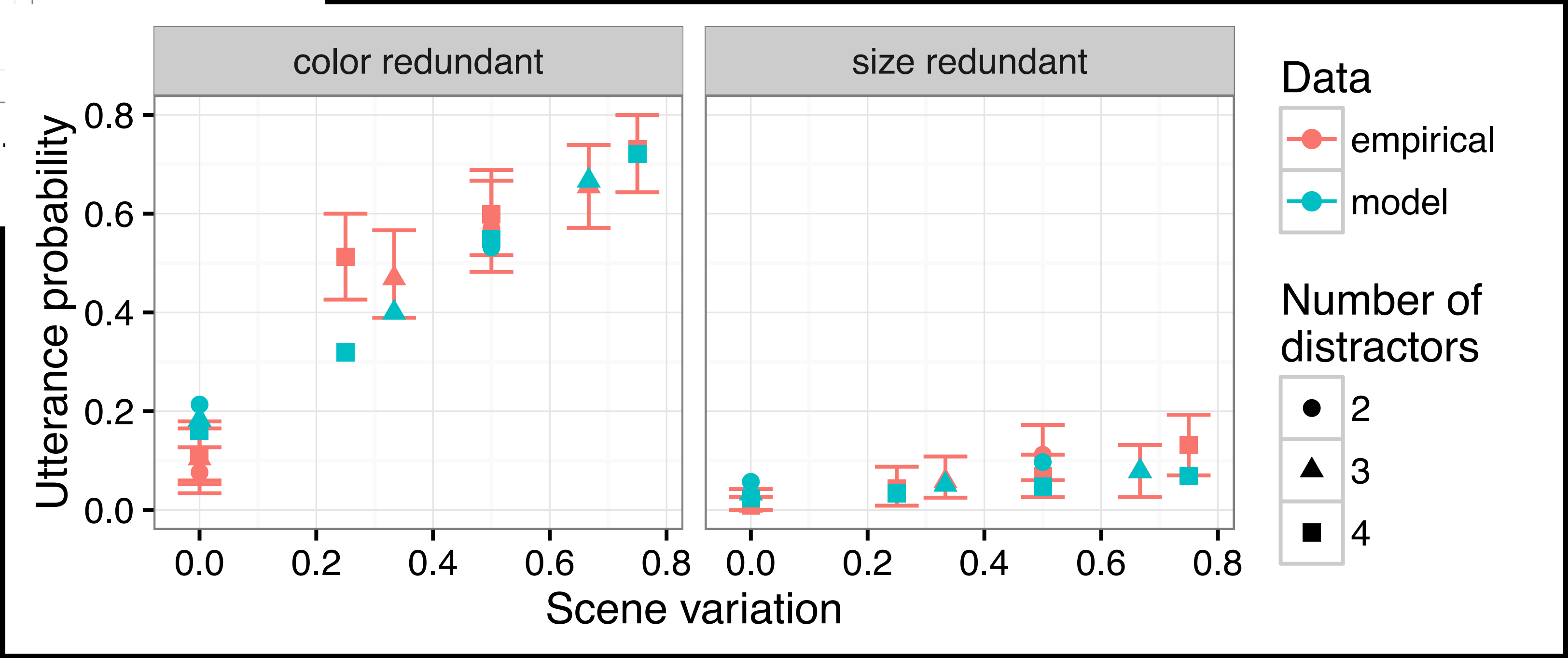
- 58 pairs of participants on Mechanical Turk
- random assignment to speaker/listener role
- 72 trials (half targets, half fillers)
- 36 object types
- on all target trials, one of size or color was sufficient
- **scene variation manipulation:**
  - total number of distractors (2, 3, 4)
  - number of distractors that shared the insufficient feature value with target

Degen, Hawkins, Graf, Goodman (in prep)

Posterior predictive

Posterior predictive from Bayesian data analysis:

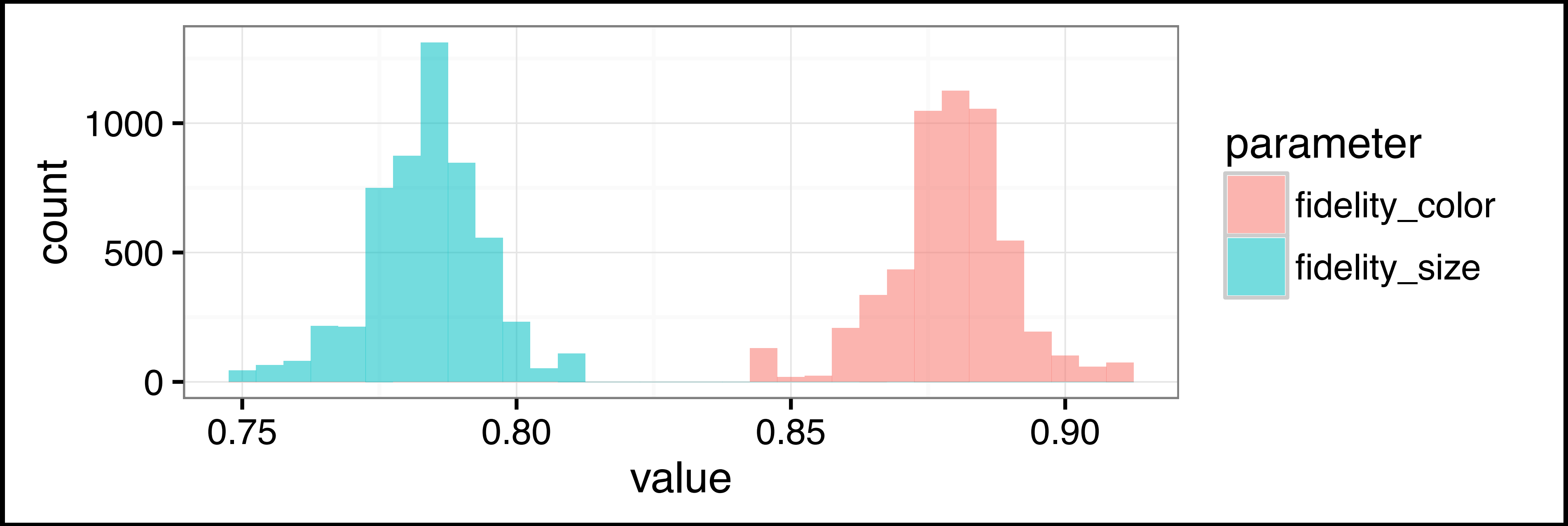




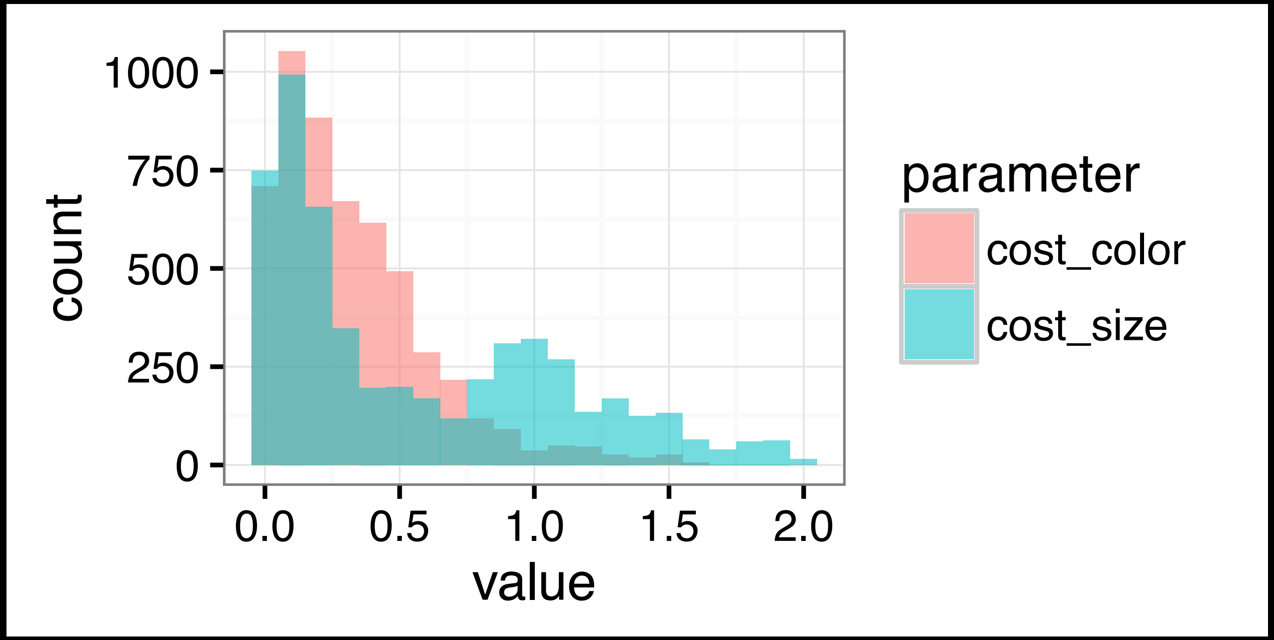
Degen, Hawkins, Graf, Goodman (in prep)

Inferred parameters

Meanings:
inferred size fidelity lower than inferred color fidelity

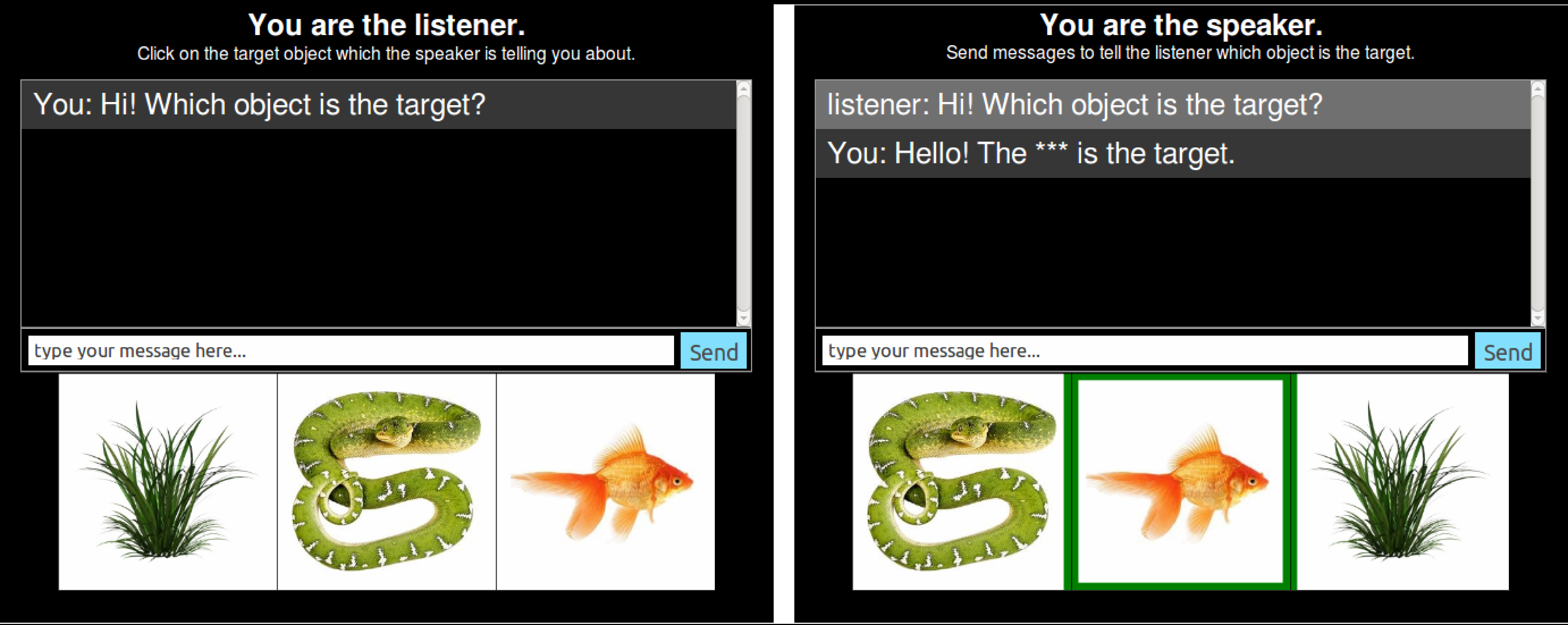


Cost:
inferred size and color costs similar



Degen, Hawkins, Graf, Goodman (in prep)

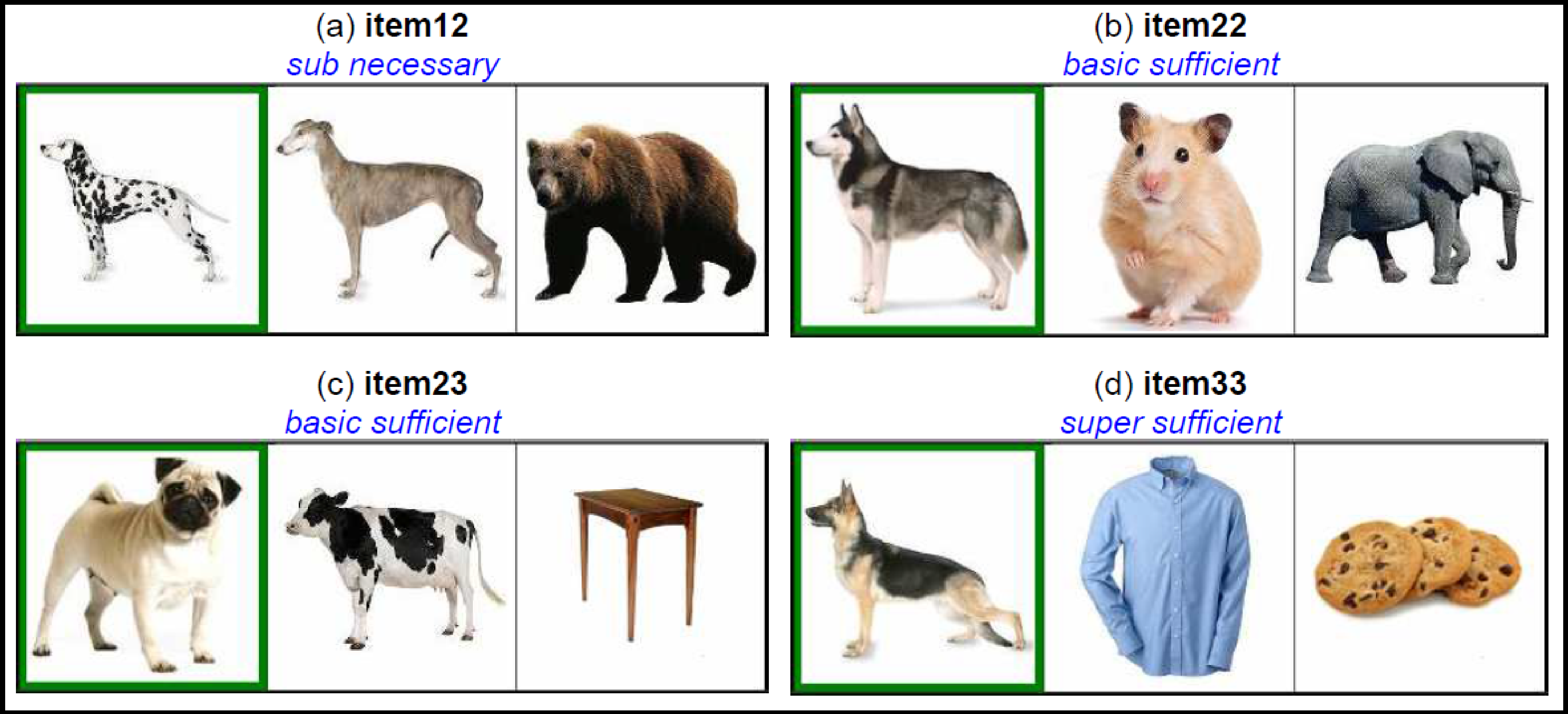
Noun-level choice



- At what level of reference will people refer to an object?
- "goldfish" vs "fish" vs "animal"

Graf, Degen, Hawkins, Goodman (2016)

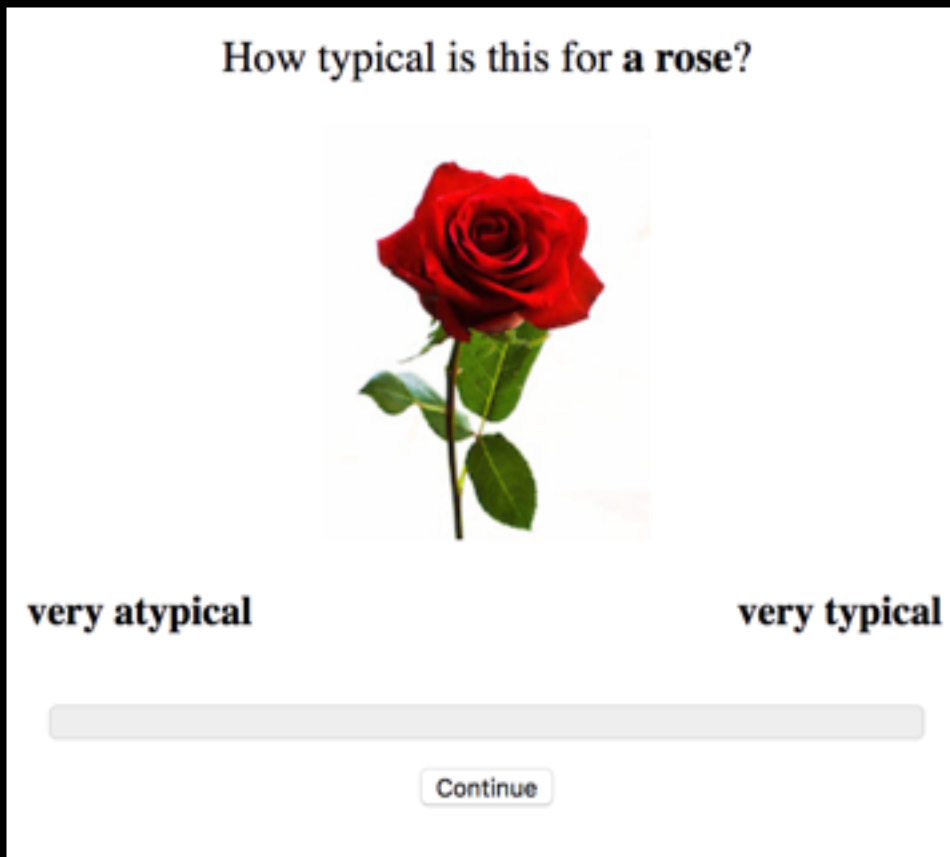
Noun-level choice

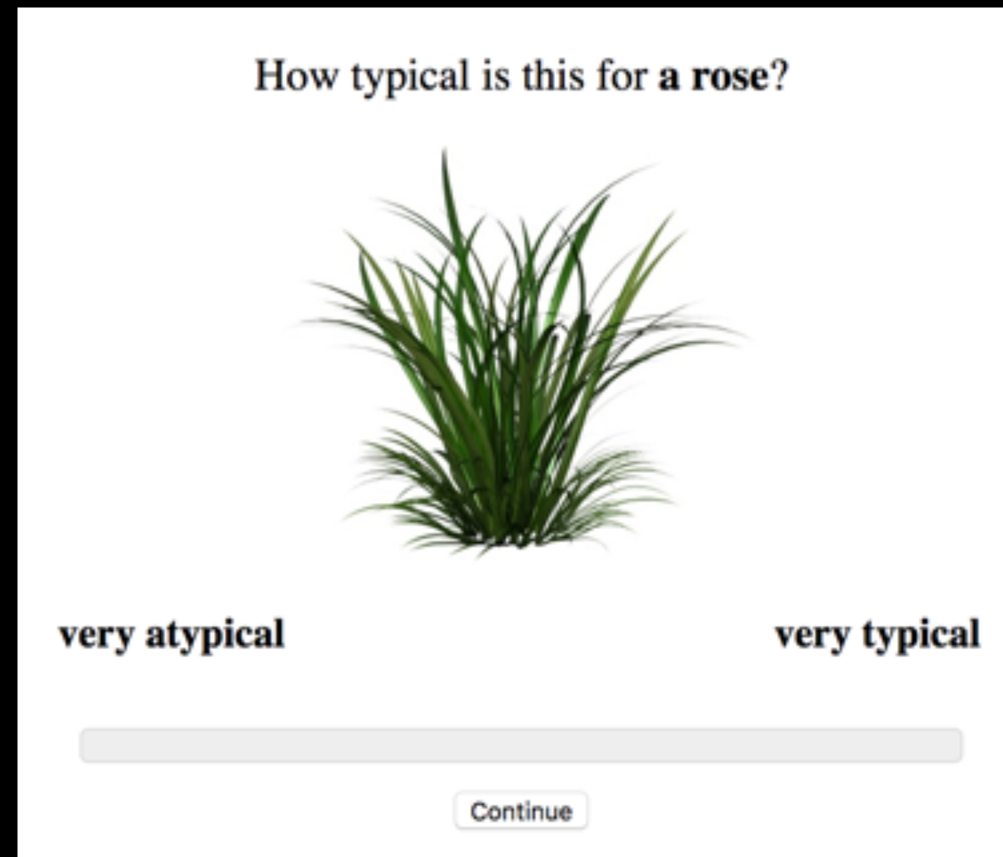


- 56 Ps (28 pairs), 36 trials each.
- Speaker utterances annotated as sub / basic / super / other.

Graf, Degen, Hawkins, Goodman (2016)

Typicality semantics





N=240 Ps, 35 ratings each

`meaning(category,object)=`
mean of empirically measured typicality ratings, in [0..1]

`meaning(“rose”,`  `)=0.99`

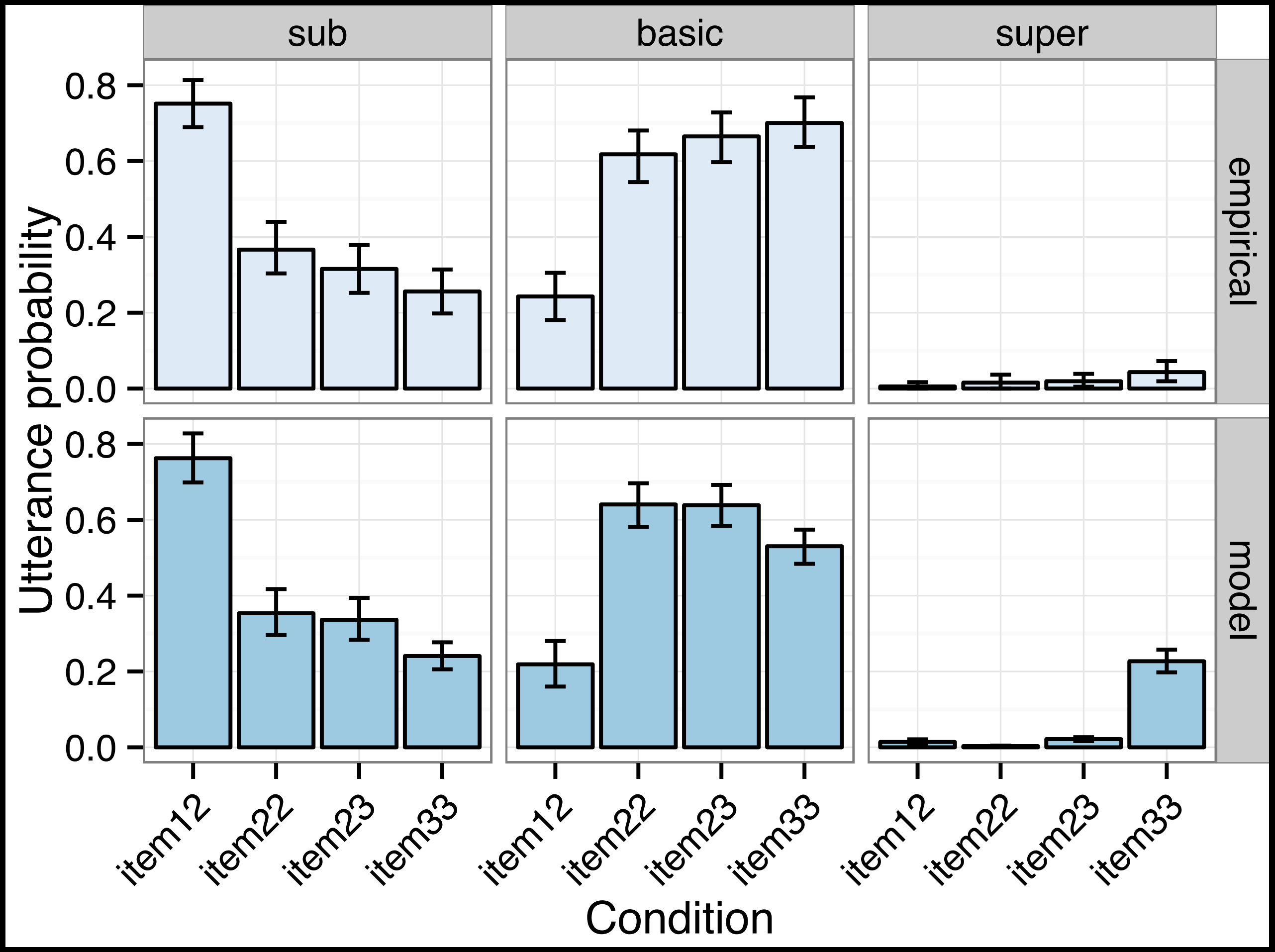
`meaning(“rose”,`  `)=0.01`

`meaning(“rose”,`  `)=0.15`

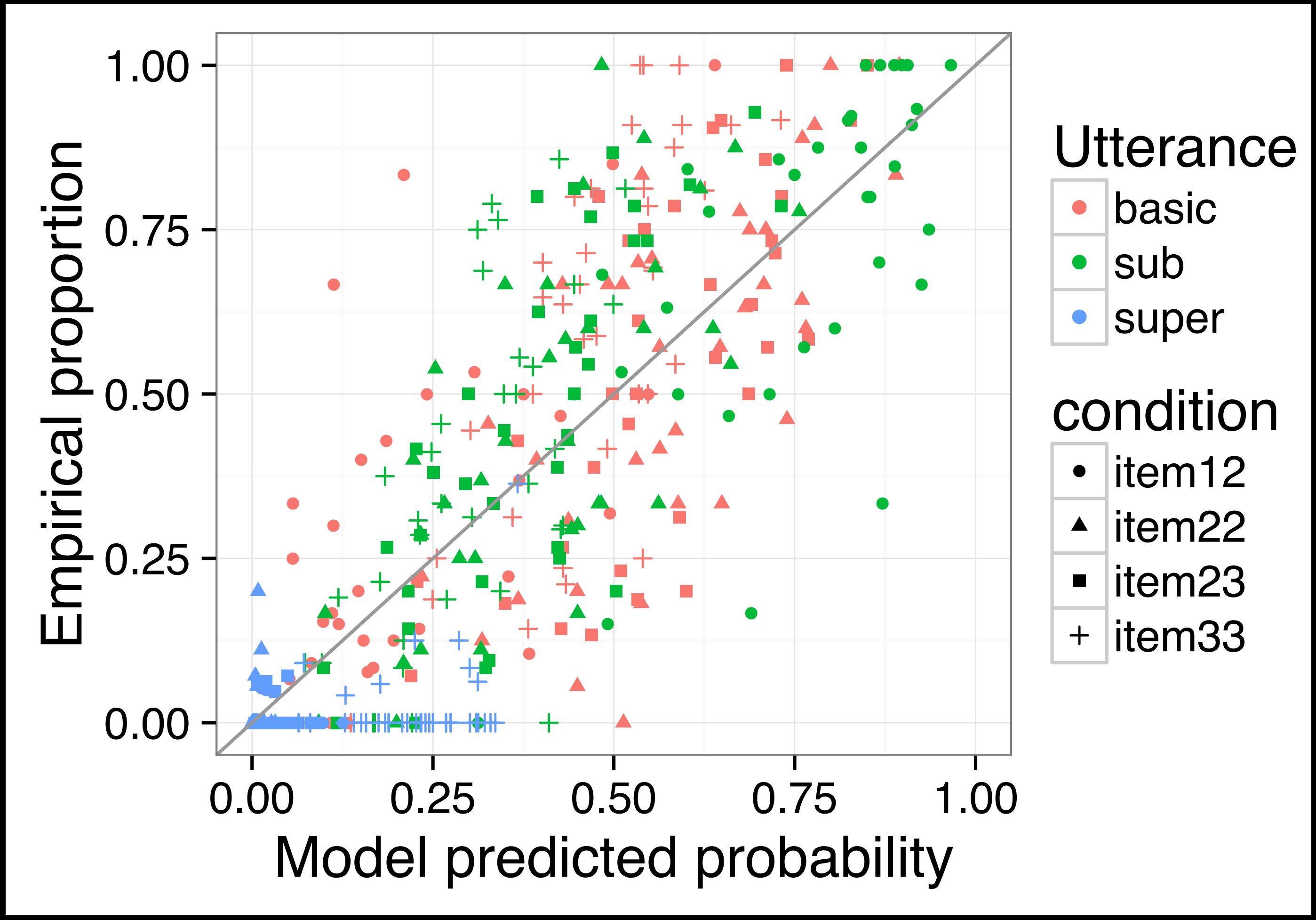
Empirical cost

- We allow the cost of each utterance to depend on number of syllables and empirical frequency (in Google1T).
  - Captures "basic-level" preference: "dog" is much more frequent and shorter than "dalmatian".
  - Inferred weights on these cost terms.

Results



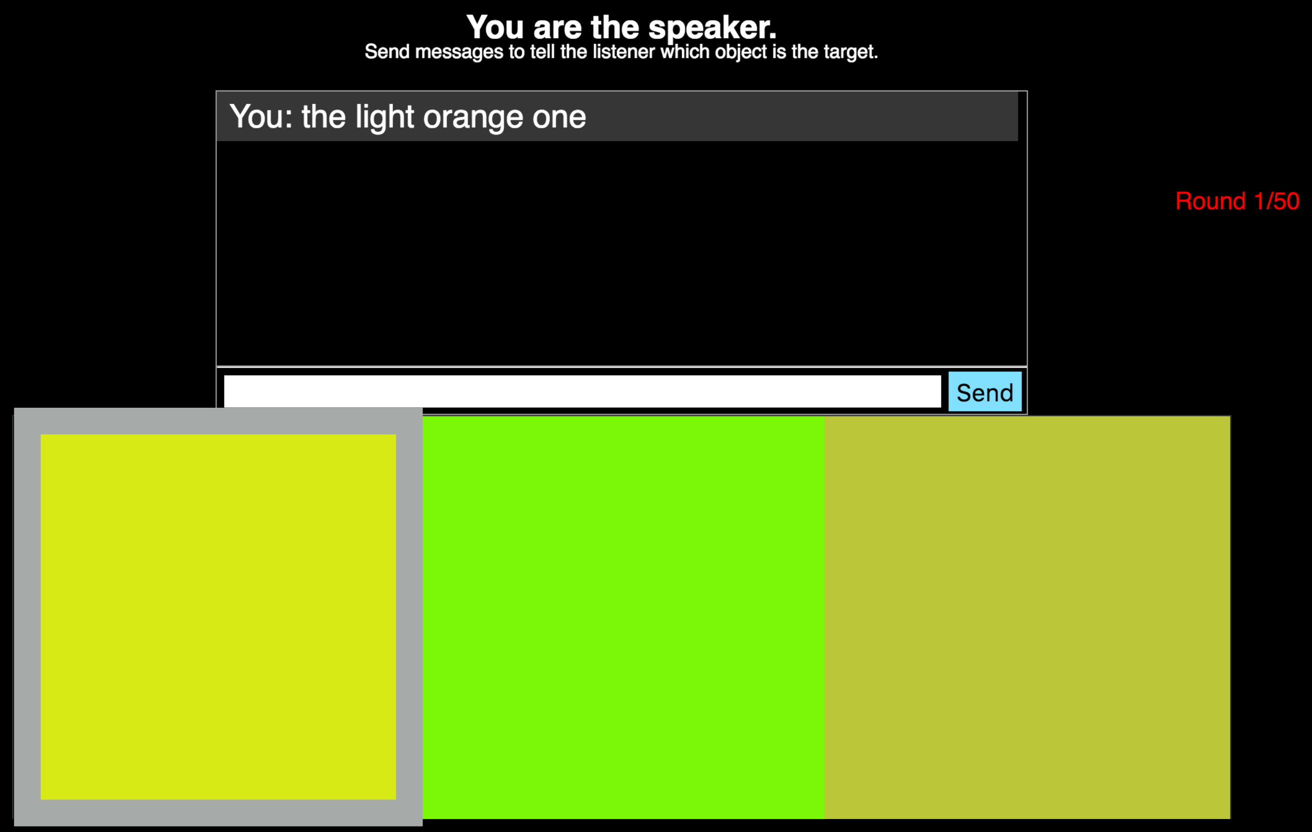
Item-level comparison:

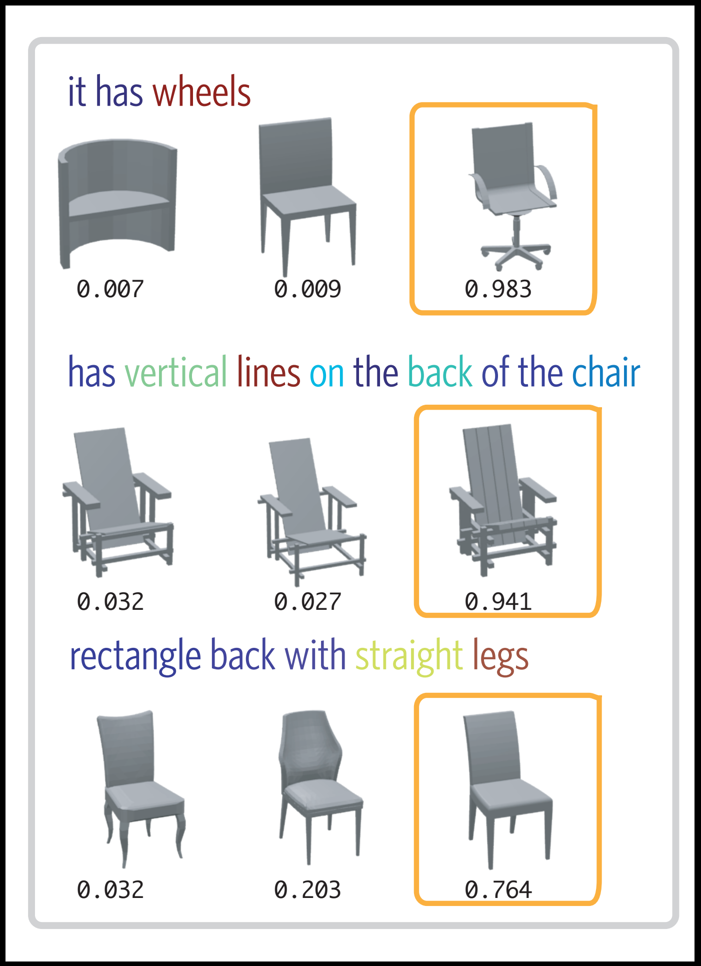


Graf, Degen, Hawkins, Goodman (2016)

Learning language

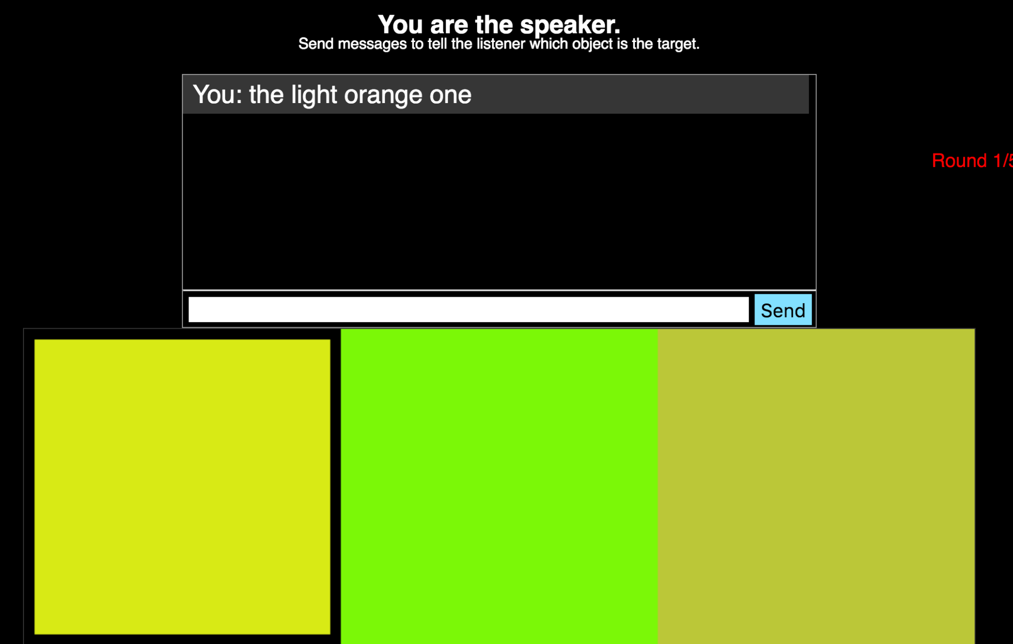
- RSA works very well for explaining language use if we have (or directly measure) the literal semantics.
- Can we learn semantics from a language game corpus?





The colors corpus

- Approx. **50k** trials.
- Three conditions:
  2 far distractors,
  2 close distractors,
  split far/close distractors.



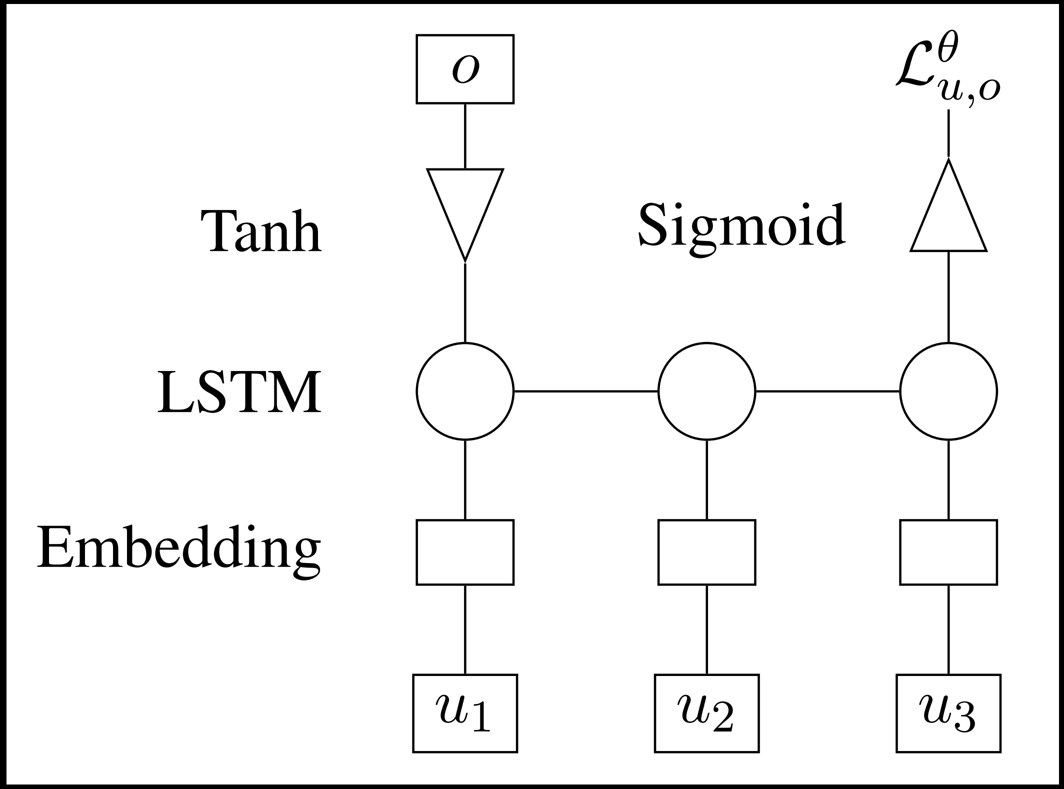
	Context			Utterance
1.				darker blue
2.				Purple
3.				blue
4.				blue

	far	split	close
# Chars	7.8	12.3	14.9
# Words	1.7	2.7	3.3
% Comparatives	1.7	14.2	12.8
% High Specificity	7.0	7.6	7.4
% Negatives	2.8	10.0	12.9
% Superlatives	2.2	6.1	16.7

Monroe, Hawkins, Goodman, Potts (2017)

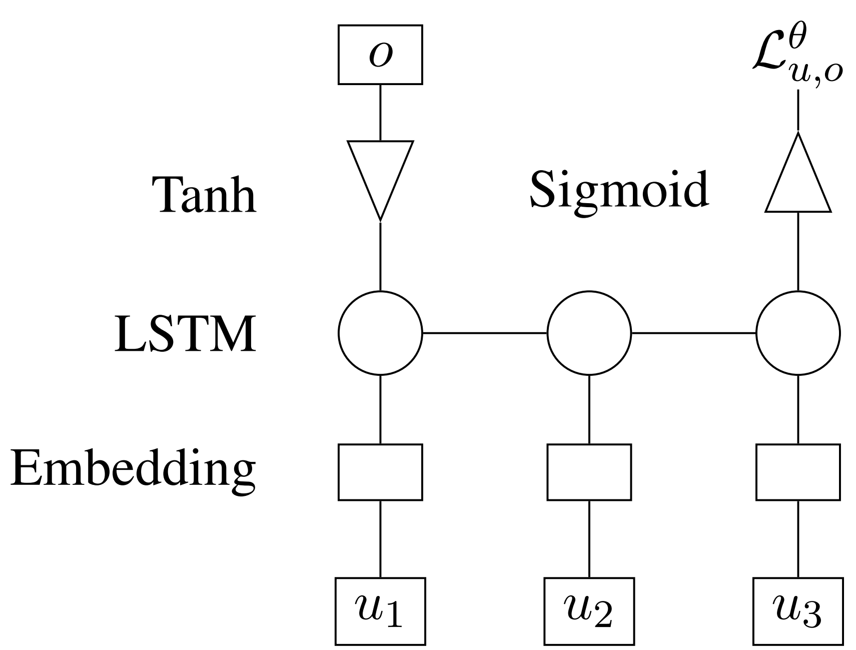
Neural semantics

- We want a flexible, and learnable space of meaning functions.
  - Use recursive neural net (LSTM) from NLP.
  - Colors represented in 3-dim CIELAB color space.
  - Text tokenized and embedded w/ GloVe.



Pragmatic training

- RSA is differentiable wrt parameters of the meaning function… Can learn by gradient descent.
  - Approximate the set of utterances via sequential Monte Carlo with a pre-trained language model prior.

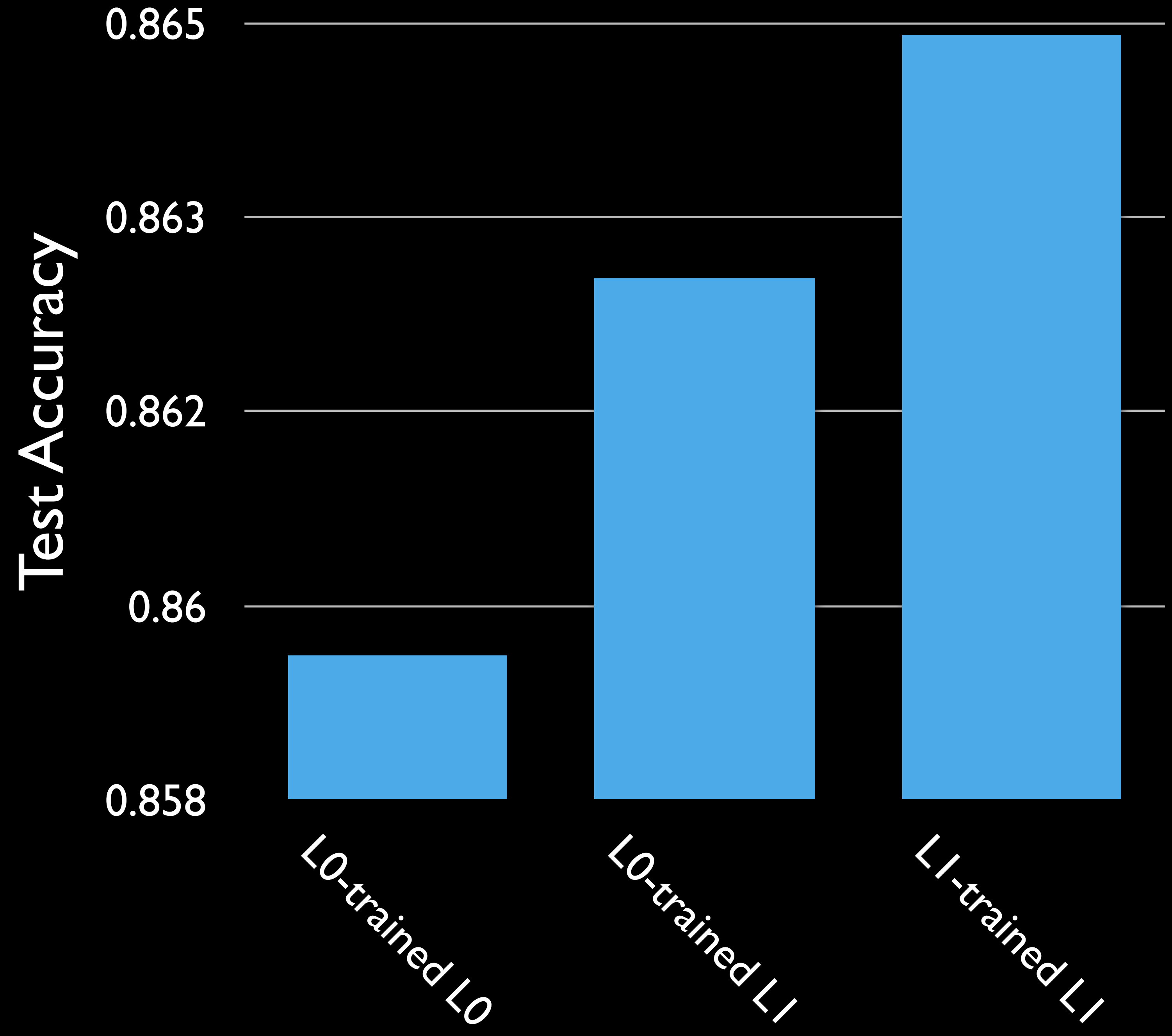


$$
\begin{aligned}
l_1(t \mid u, O; \theta) &\propto s_1(u \mid t, O; \theta) p(t) \\
s_1(u \mid t, O; \theta) &\propto l_0(t \mid u, O; \theta)^{\alpha} p(u) \\
l_0(t \mid u, O; \theta) &\propto \mathcal{L}^{\theta}_{u,O_t} p(t)
\end{aligned}
$$

Pragmatics during training?

- We can use pragmatics during learning and/or when model is used at test.
  - L0-trained L0: no pragmatics.
  - L0-trained L1: pragmatics at test time only.
  - L1-trained L1: pragmatics at training and test.

Results, full data

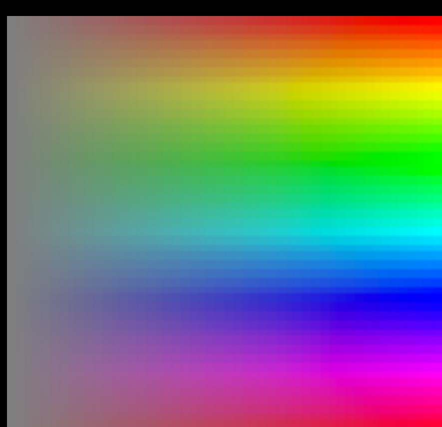


Trained on 15k utterances.

McDowell, Goodman (in prep)

Learned meanings

Color space:



L0-train L1-train

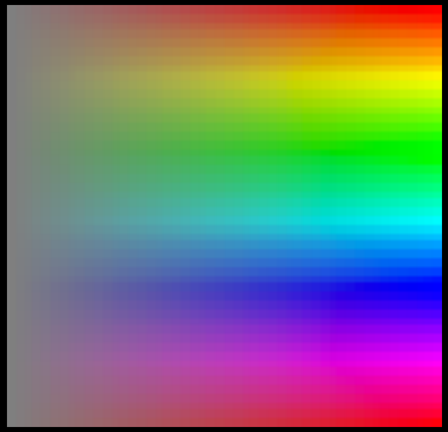
"Blue"

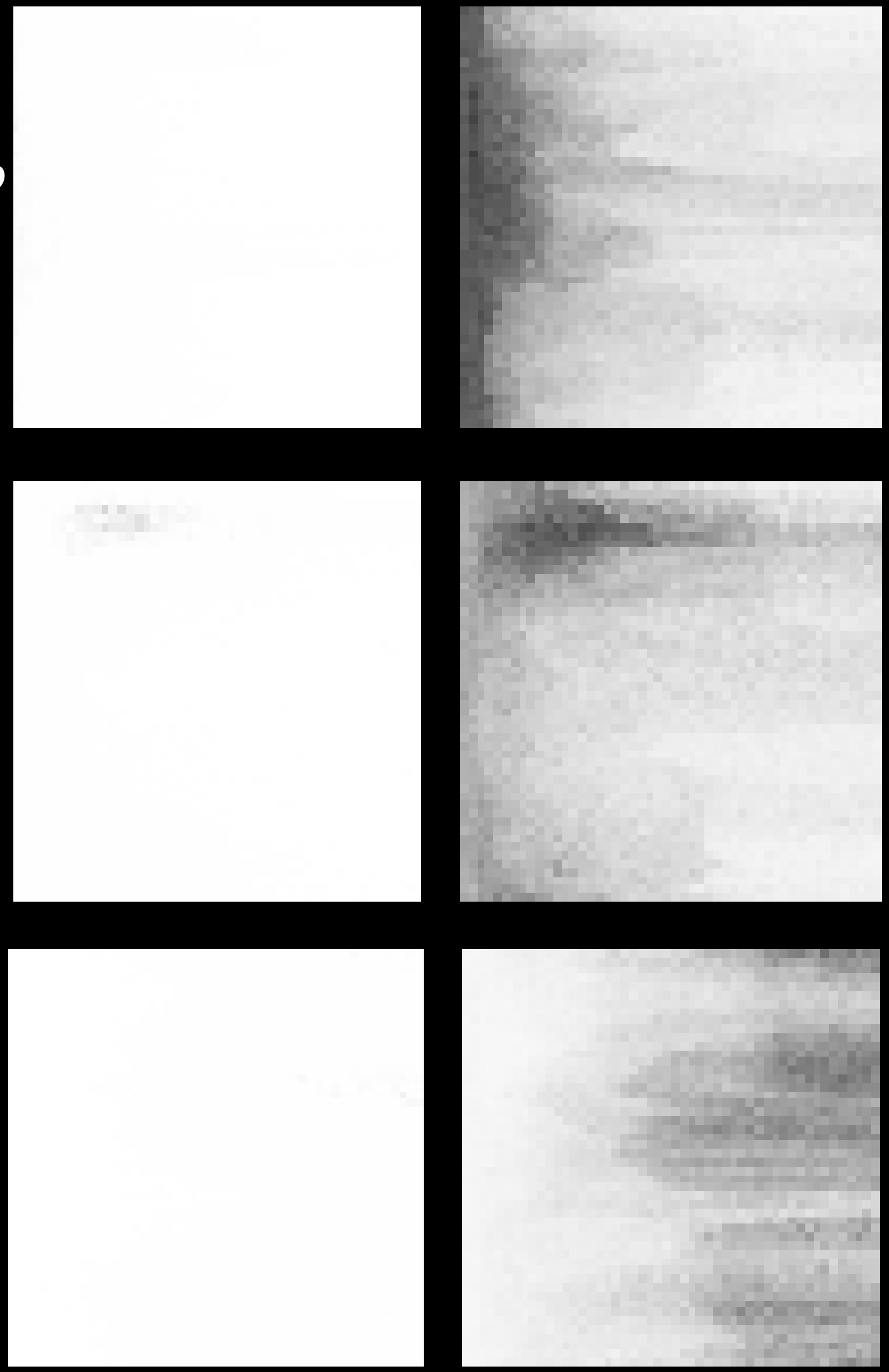
"Purple"

"Red"

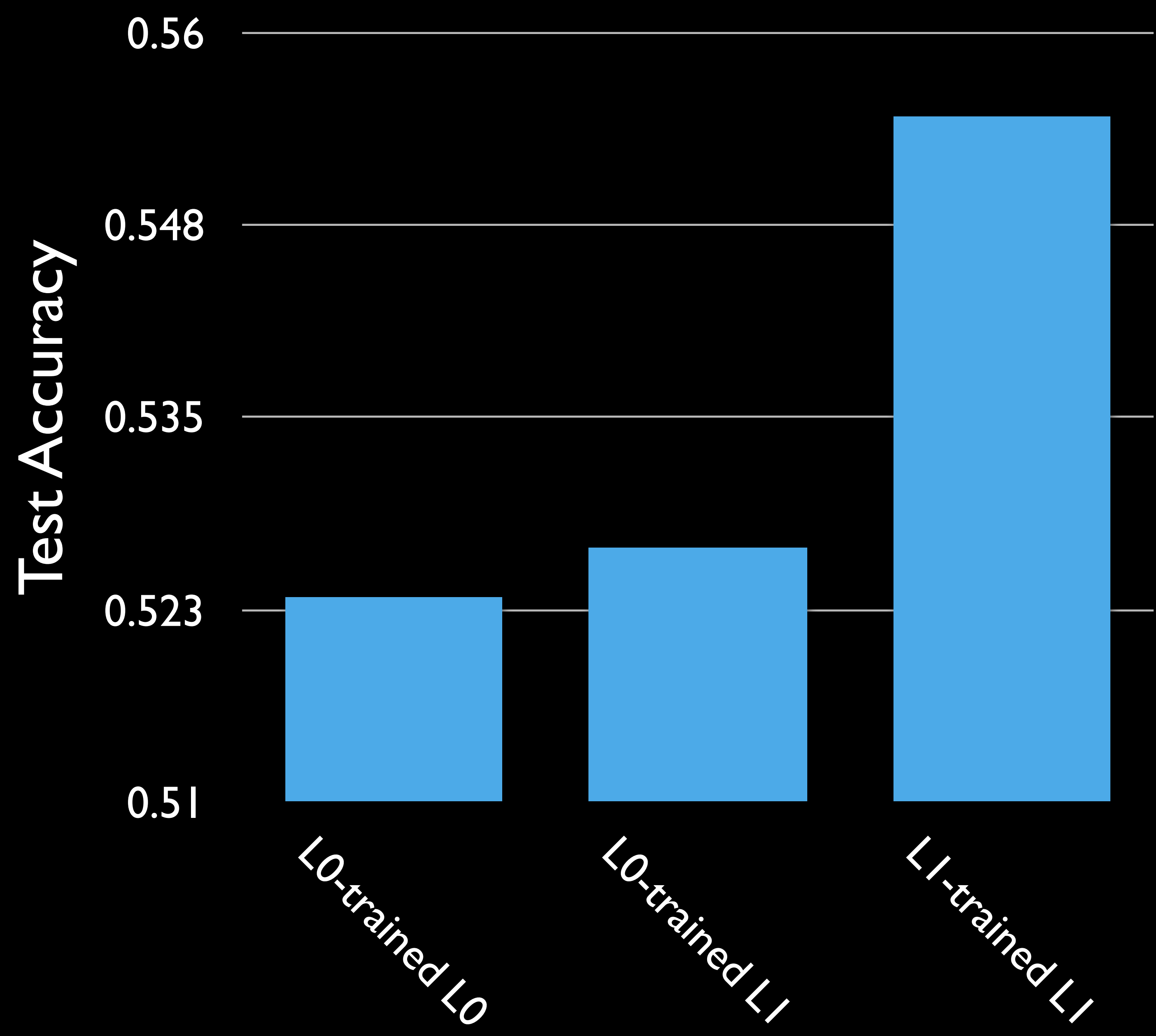
Learned meanings

Color space:





Results, small data

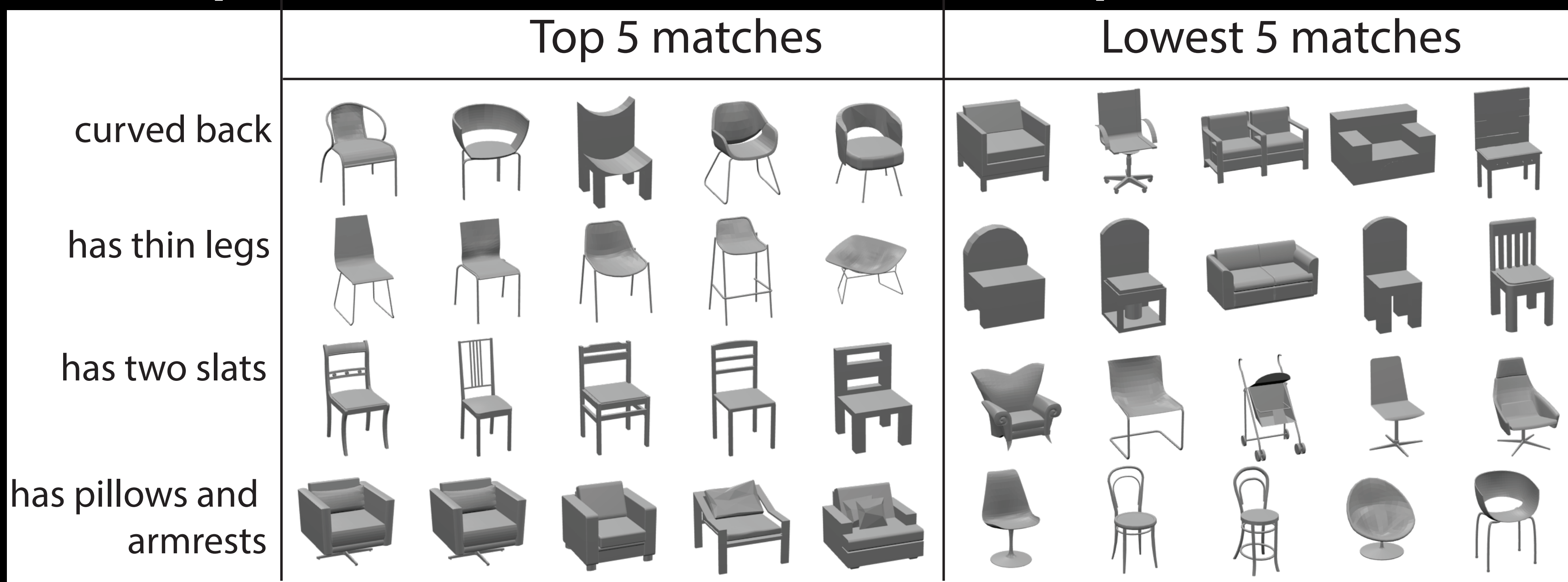


Trained on 128 utterances.

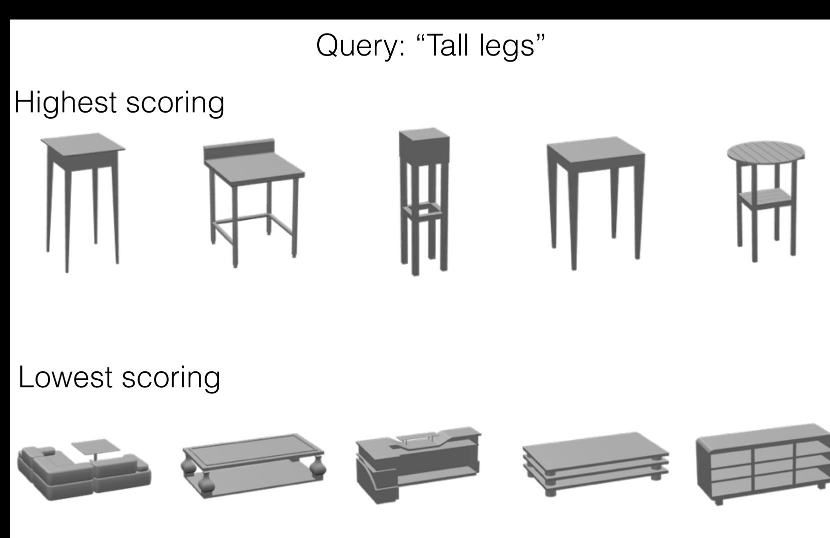
McDowell, Goodman (in prep)

Transfer to new domains

Using trained (L0) listener to search for best exemplar from all chairs in ShapeNet:

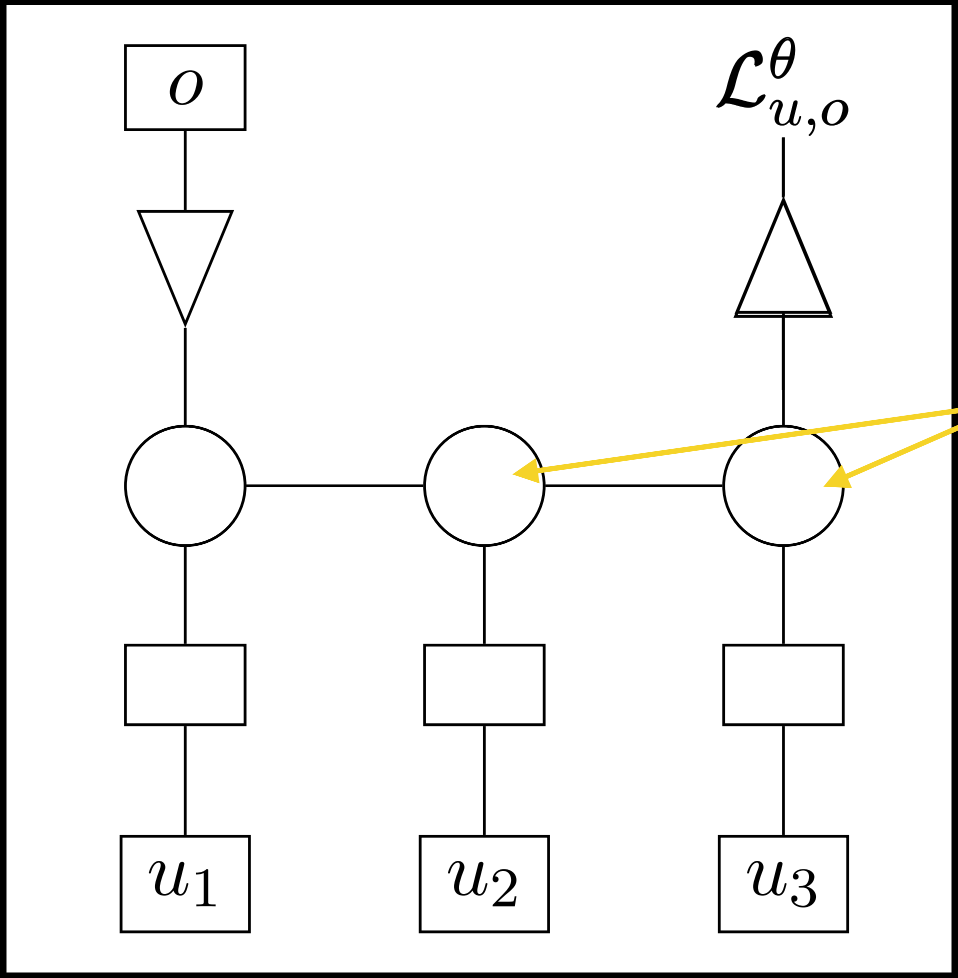


Same listener seems to work for tables:



Incremental pragmatics

- We can read-out early from the neural semantics, applying RSA to get informativity of partial utterances.



These representations have same type, allowing “early” semantic interpretation

Cohn-Gordon, Potts, Goodman (2018)

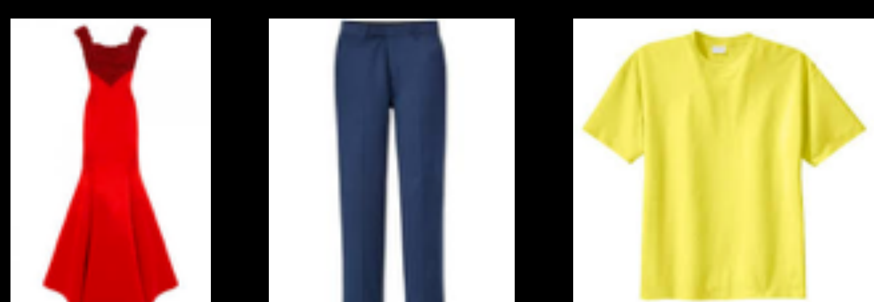
Incremental pragmatics

- This even works character-by-character.

S_0 caption: a double decker bus
S_1 caption: a red double decker bus

Model	Test accuracy
Char S_0	48.9
Char S_1	**68.0**
Word S_0	57.6
Word S_1	60.6

- Perhaps related to human incremental utterance processing?



"the red dress">"el vestido rojo"
Rubio-Fernandez 2016

The → red → dress
The → dress

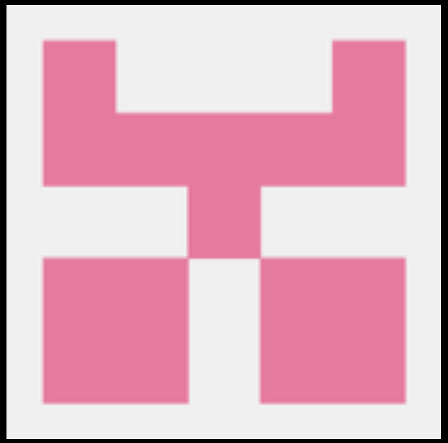
El → vestido

Cohn-Gordon, Potts, Goodman (2018)

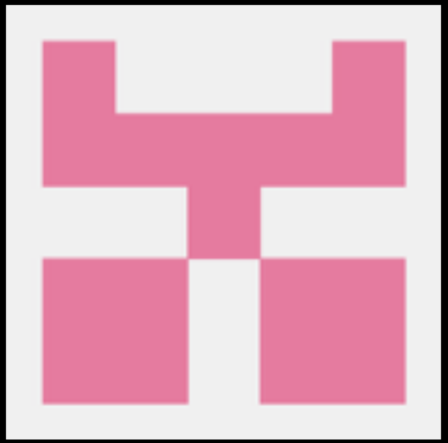
Thanks

Mike Frank

Judith Degen



Bill McDowell



Reuben Cohn-Gordon

Justine Kao

Computation & Cognition Lab

Robert Hawkins
Elisa Kreiss
Caroline Graf
Leon Bergen
Other collaborators…

- Funding from: JSMF, ONR, DARPA.

Learning from language

“Red-spotted mushrooms are poisonous.”

“Lightning causes fires.”

“Feps are gentle.”

- Language is a unique opportunity for learning.
  - The predictive content of a great deal of direct experience can be conveyed in a sentence.
  - This may be central to the “cultural ratchet”, enabling the unique successes of our species.

Generalizations

Robins lay eggs.

Mosquitos carry malaria.

Ravens are black.

- Category generics seem to convey generalizations. What do they actually mean?
- Probability is a universal currency of belief, useful in describing human generalization from examples. (Cf. Shepard, 1987; Tenenbaum & Griffiths, 2001)
- So, maybe generics refer to probability?

Tessler and Goodman (in prep)

Generalizations

$$[\![\text{Some ravens are black}]\!] := \{P(\text{is } \overset{\text{feature}}{\text{black}} \mid \text{is a } \overset{\text{kind}}{\text{raven}}) > 0\}$$

$$[\![\text{Most ravens are black}]\!] := \{P(\text{is black} \mid \text{is a raven}) > 0.5\}$$

$$[\![\text{All ravens are black}]\!] := \{P(\text{is black} \mid \text{is a raven}) = 1\}$$

$$[\![\text{Ravens are black}]\!] := \{P(\text{is black} \mid \text{is a raven}) > \theta\}$$

feature kind

$$[\![generalization]\!] := \{P(f \mid k) > \theta\}$$

50% Robins lay eggs. ✓

50% Robins are female. ✗

> 99% Ravens are black. ✓

< 1% Mosquitos carry malaria. ✓

A vague semantics

The generic utterance is true proportionally to the prevalence probability*:

$$\mathcal{L}(p, \text{“k f”}) = p$$

$$p = P(f \mid k) = \text{prevalence}$$

Standard RSA listener interprets utterance:

$$L_0(p|u) \propto P(p)\mathcal{L}(p, u)$$

Endorsement is modeled as speaker's decision to utter the generic vs staying silent (Cf. Franke, 2014; Degen & Goodman, 2014):

$$S_1(u|p) \propto L_0(p|u)^{\alpha}$$

$$u \in \{generalization, null\}$$

$$\mathcal{L}(p, \text{“k f”}) = p \quad \mathcal{L}(p, \text{“...”}) = 1$$

*This is equivalent to a simple uniformly uncertain threshold semantics.

Prevalence priors

Background knowledge: a prior distribution on prevalence

$$L_0(p|u) \propto P(p)\mathcal{L}(p, u)$$

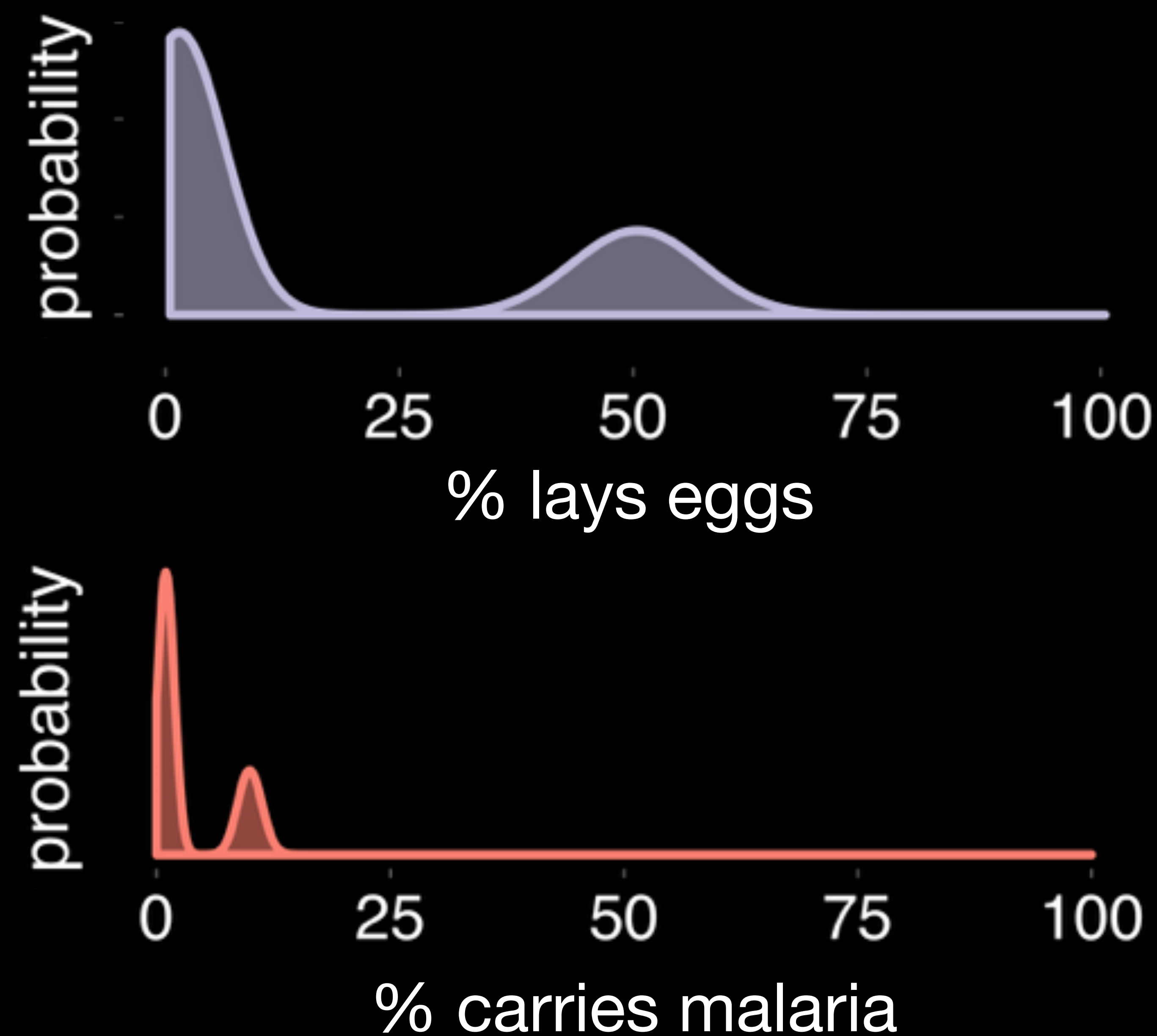
$$p = P(f \mid k)$$

Think of some kind of animal.

What percentage *are female*?

What percentage *lay eggs*?

Hypothetical prevalence priors



Prior elicitation

Category elicitation

supplied to participants

Kangaroos
Robins
Sharks
Mosquitos
Ducks
Ticks

participants generate animal kinds

Continue

Prevalence elicitation

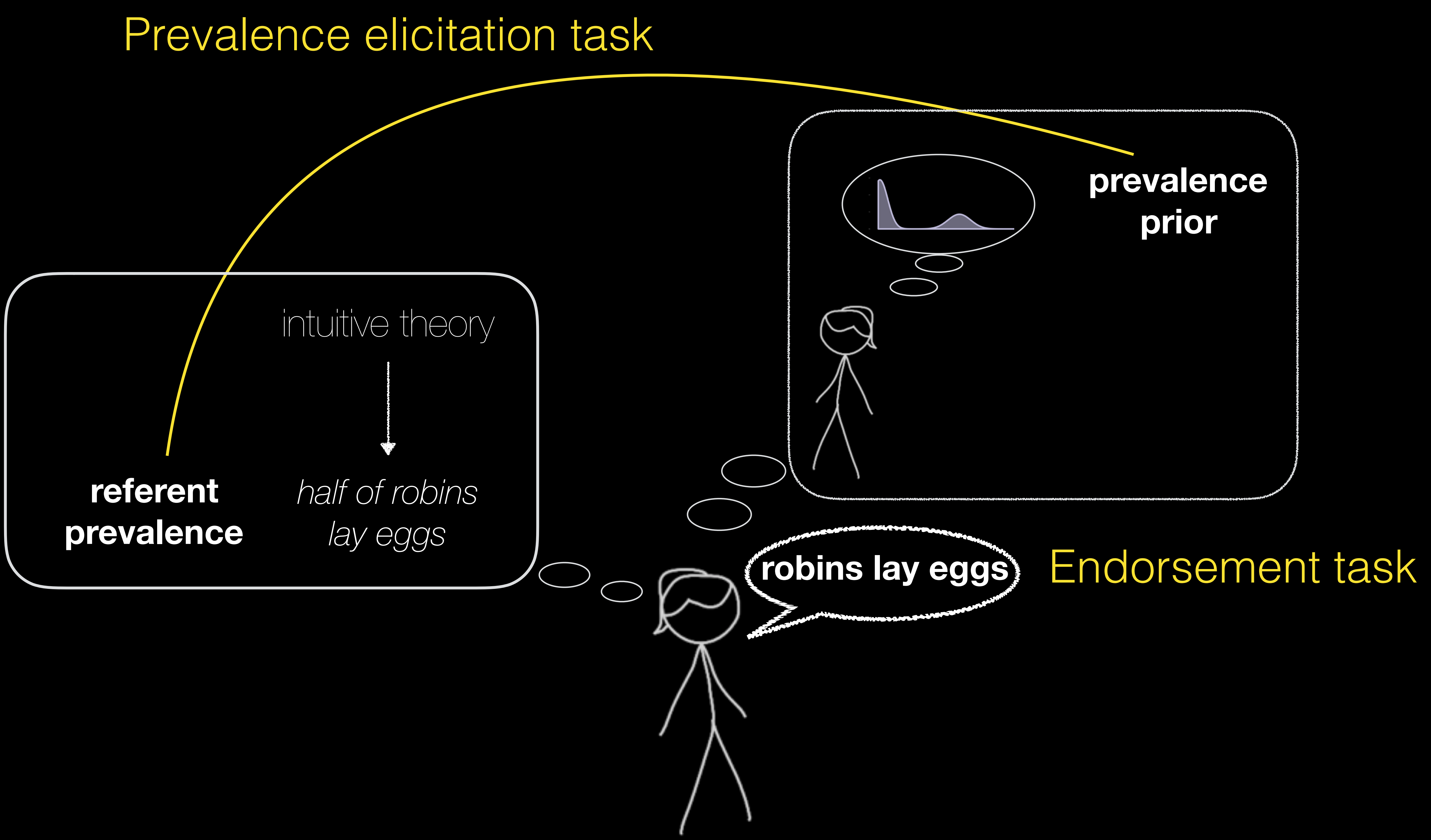
For each kind of animal, what percentage of the species do you think

	carry malaria
Kangaroos	0 %
Robins	0 %
Sharks	0 %
Mosquitos	10 %
Ducks	0 %
Ticks	3 %
dogs	1 %
cats	1 %
geese	0 %
monkeys	1 %
falcons	0 %

n = 60 from Amazon’s Mechanical Turk

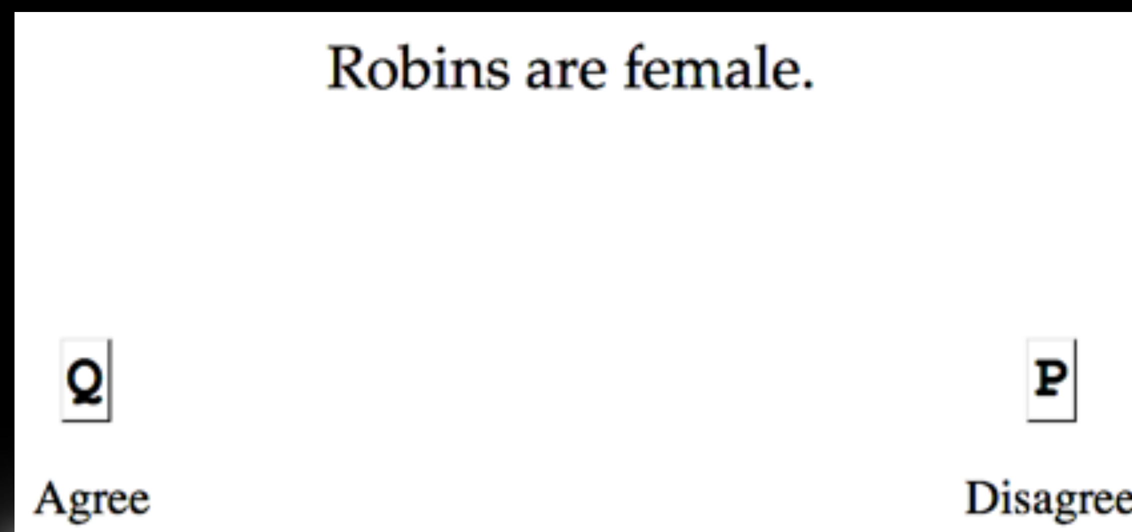
filtering 0% responses

21 properties in total



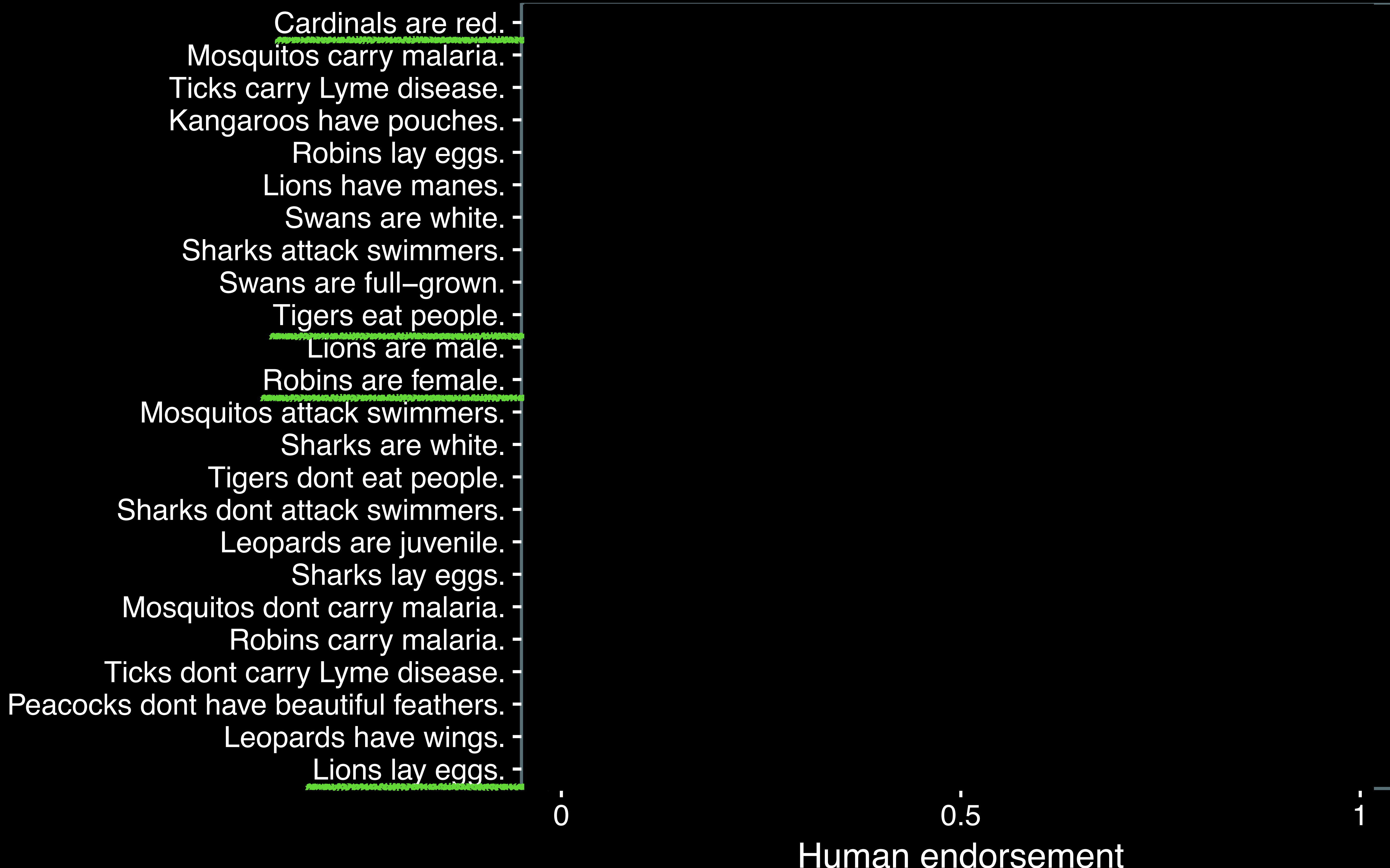
Prevalence elicitation task
prevalence prior
intuitive theory
referent prevalence
half of robins lay eggs
robins lay eggs
Endorsement task

Endorsement task (Generics)



30 items in total

items selected from the linguistic literature +
cover different “conceptual distinctions” (Prasada et al., 2013)



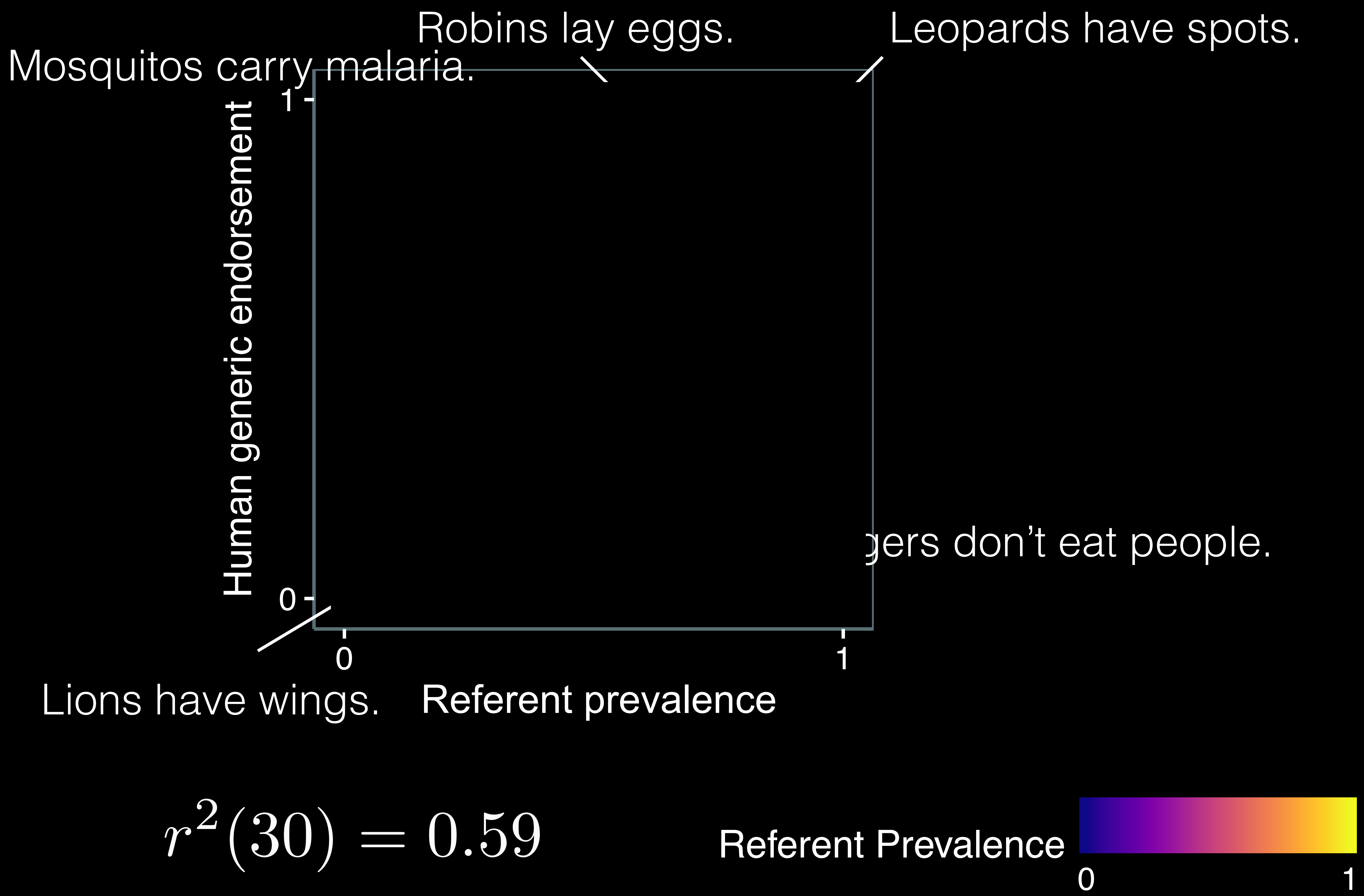
error bars = 95% Bayesian credible intervals

Referent Prevalence

similar in 5-year-olds: Brandone et al. (2012)

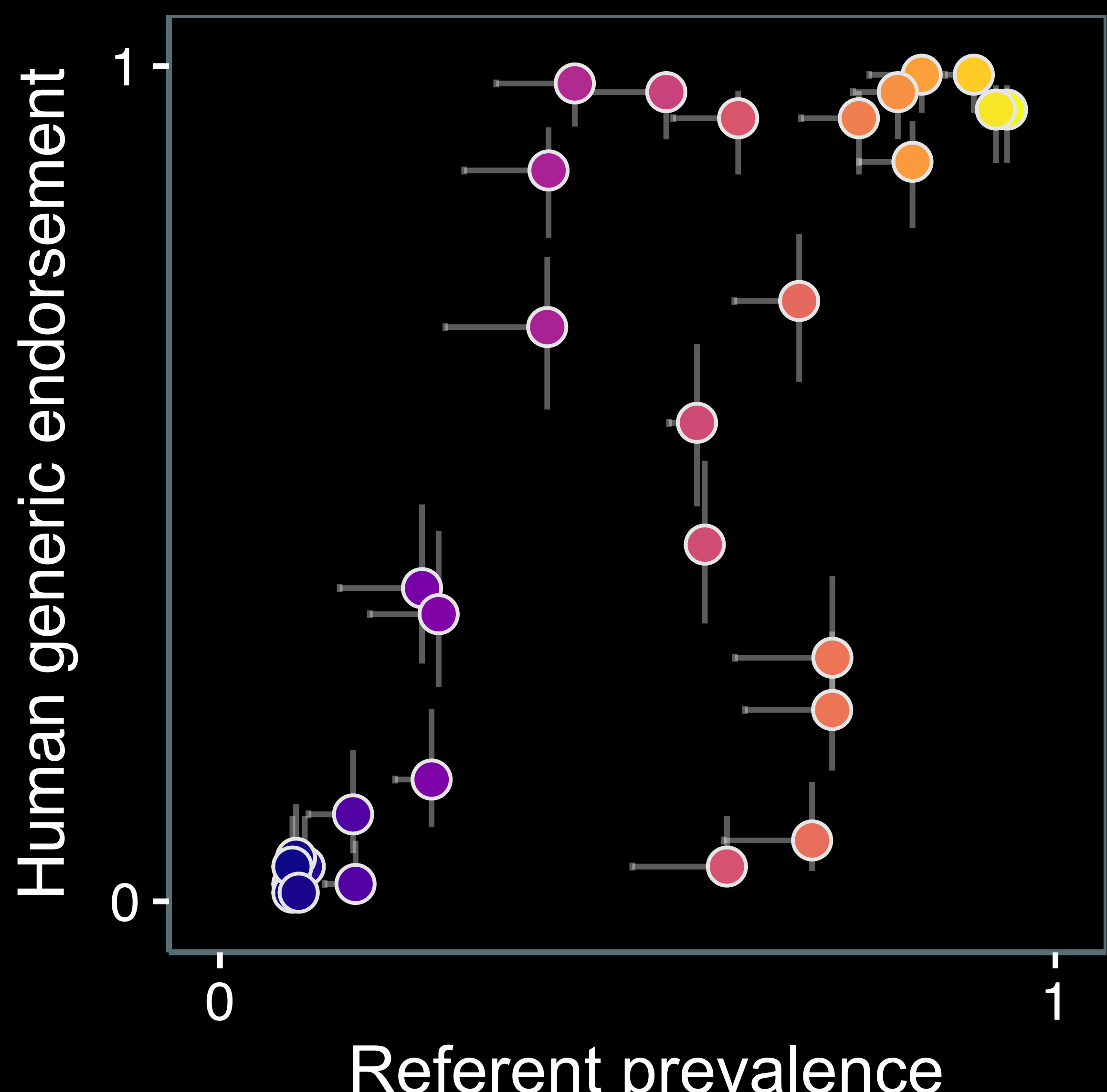
Simple statistical hypothesis

Linear model: Referent prevalence



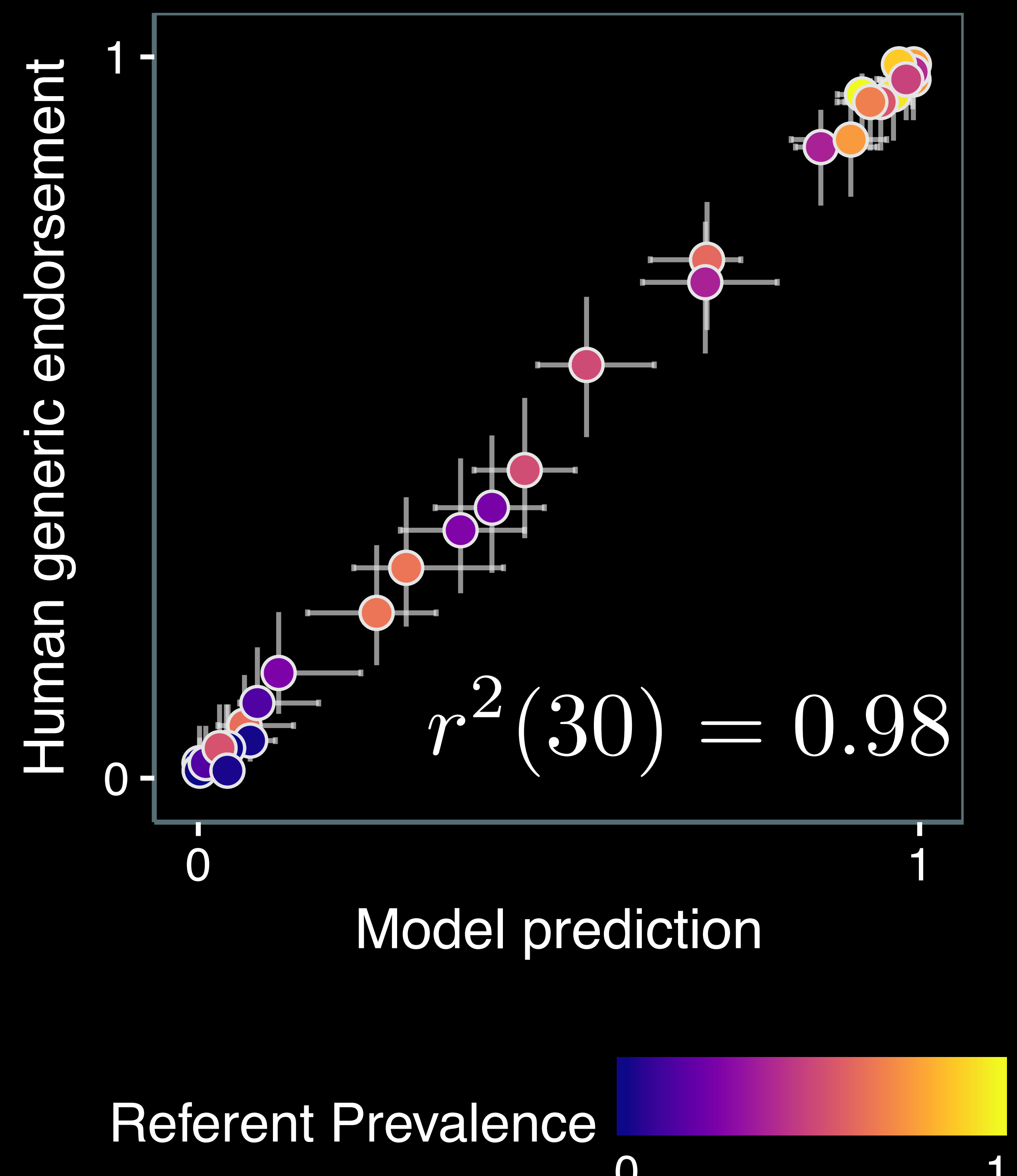
$$r^2(30) = 0.59$$

<u>Simple statistical hypothesis</u>
Linear model: Referent prevalence



$$r^2(30) = 0.59$$

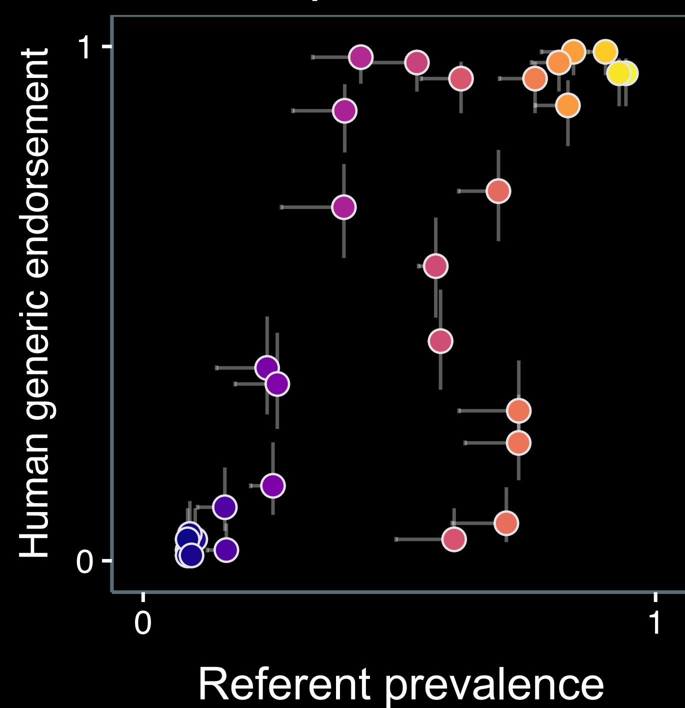
<u>Endorsement model</u>
Probability, vagueness, & context



Simple statistical hypothesis

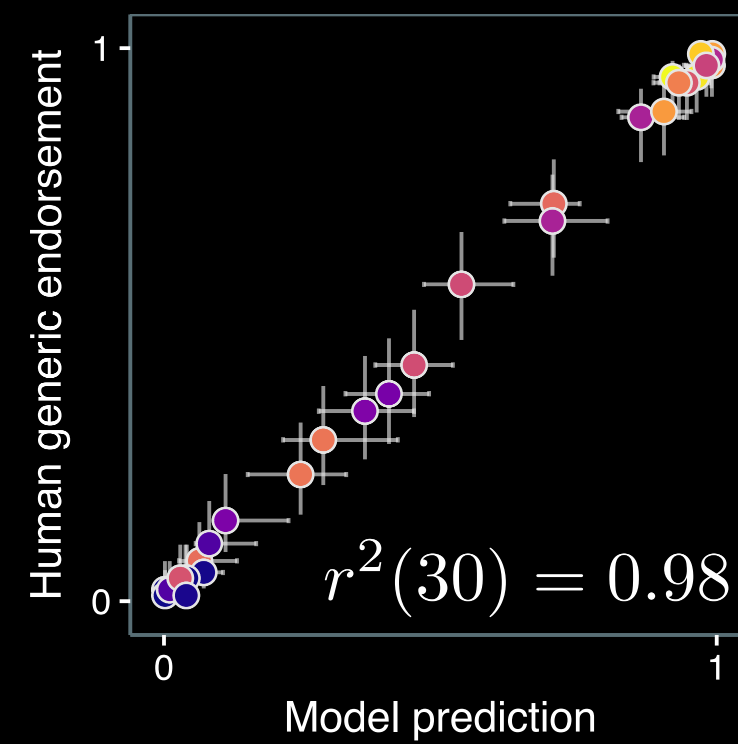
Linear model: Referent prevalence

$r^2(30) = 0.59$



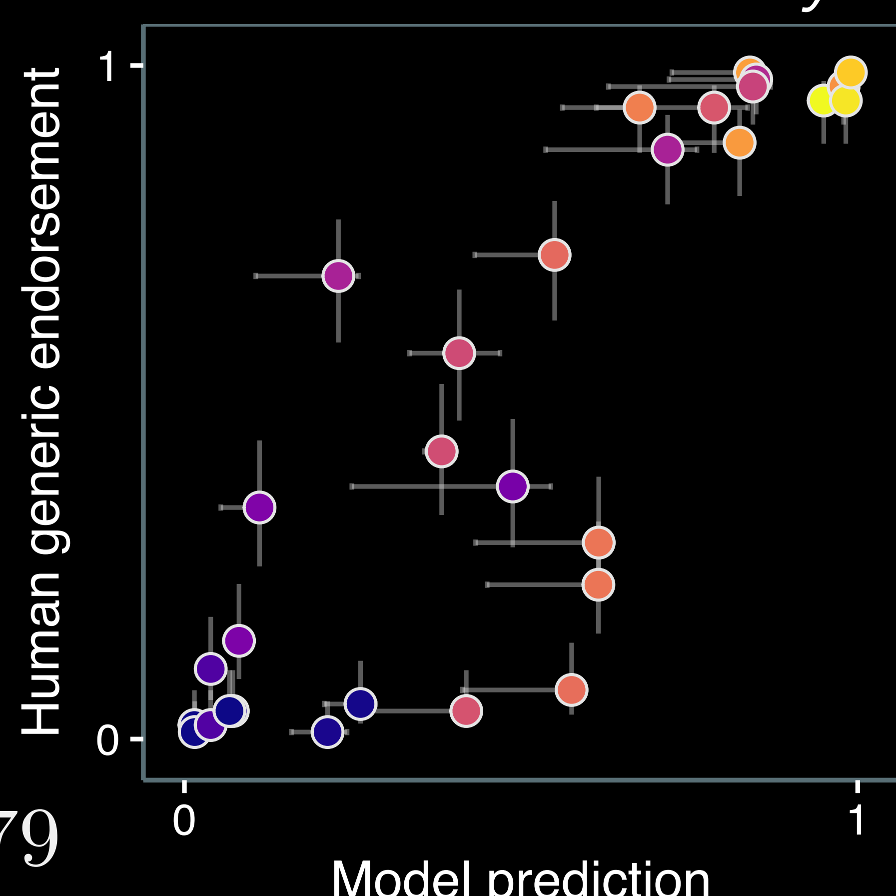
Endorsement model

Probability, vagueness, & context



Alternative statistical hypothesis

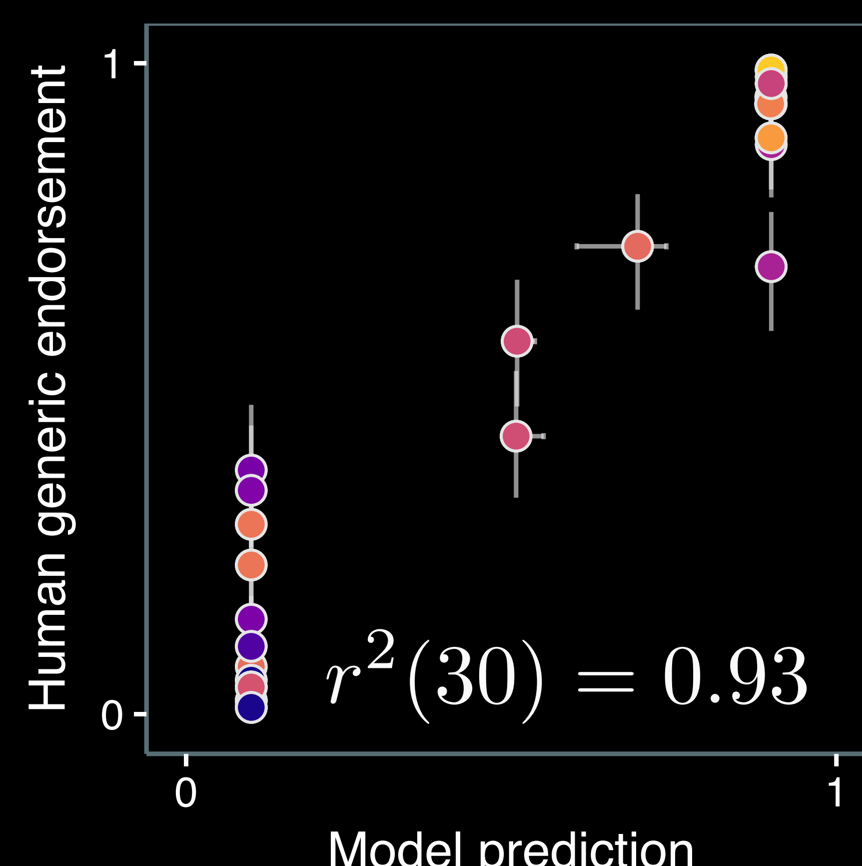
Referent prevalence + *cue validity*



$r^2(30) = 0.79$

Lesioned endorsement model

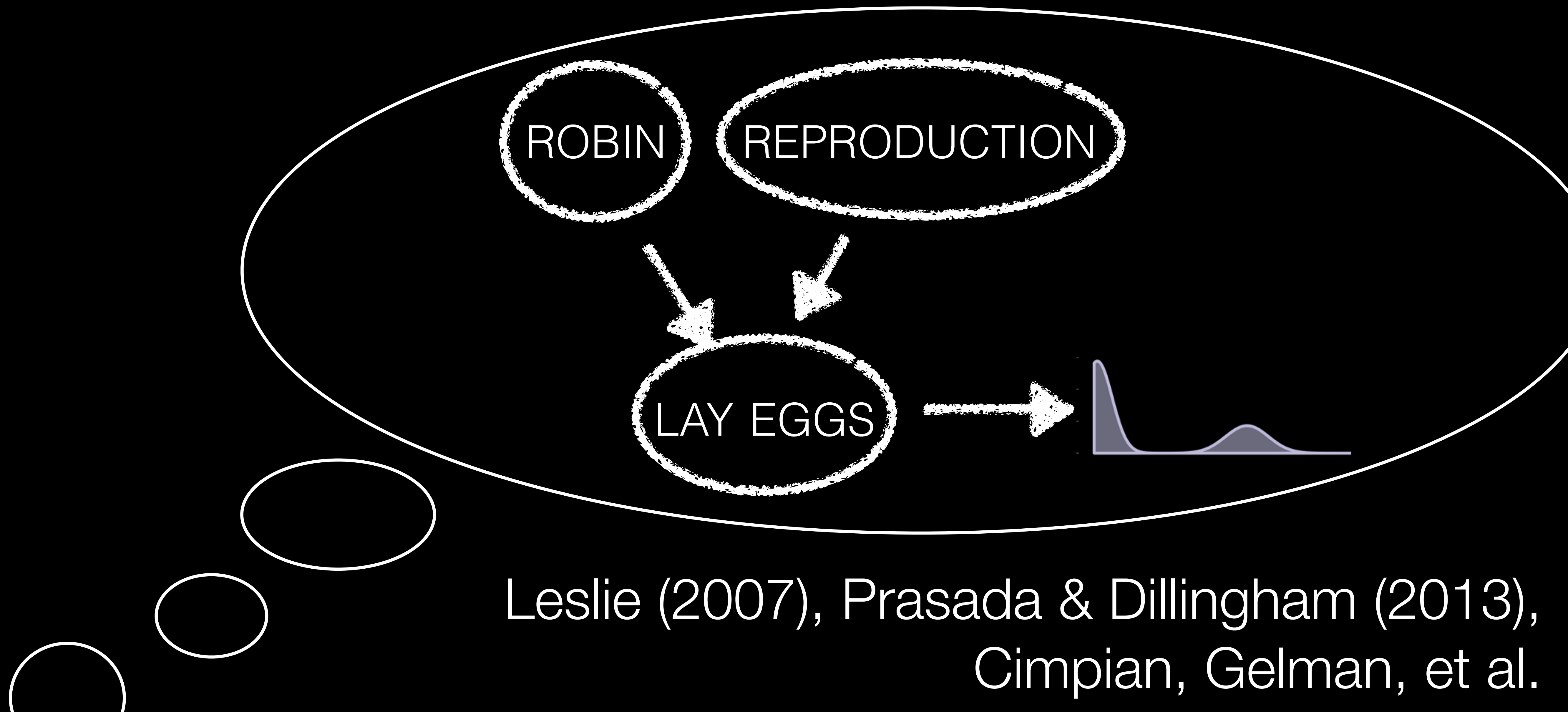
Fixed prevalence threshold



Conceptual structure

- Generic understanding has to do with KIND—PROPERTY relations
- *"Birds lay eggs" conveys the reproductive capacity of birds*

Different KIND—PROPERTY relations give rise to different prevalence priors



Leslie (2007), Prasada & Dillingham (2013),
Cimpian, Gelman, et al.

Language of generalization

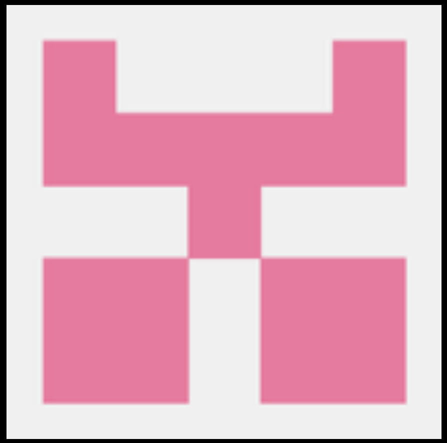
Case studies of generalization	**Categories (generics)**	**Events (habituals)**	**Causes (causals)**
Example	*Robins lay eggs*	*John smokes*	*Drinking moonshine makes you go blind*
Generalizing over	Individual robins	John at particular times	Instance of moonshine causing blindness
Prevalence prior	**Measured**	**Measured**	**Manipulated**
Referent prevalence	**Measured**	**Manipulated**	**Manipulated**

Tessler and Goodman (in prep)

Thanks


Mike Frank


Judith Degen


Bill McDowell


Justine Kao

Computation & Cognition Lab

Reuben Cohn-Gordon
Robert Hawkins
Elisa Kreiss
Caroline Graf
Leon Bergen
Other collaborators…


Michael Tessler

- Funding from: JSMF, ONR, DARPA.