A QUANTITATIVE EXPLORATION OF ADJECTIVE ORDERING PREFERENCES WITH AN INCREMENTAL RATIONAL SPEECH ACT MODEL

Hening Wang & Fabian Schlotterbeck

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

ESSLLI 2023 Pros & Comps Workshop
Ljubljana, Slovenia, 2. Aug. 2023

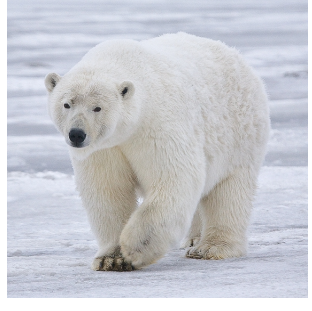
OUTLINE

1. Introduction
    1. Theoretical Background
    2. Motivation
    3. Research Aim
2. Experiment: Preference Ratings
    1. Design
    2. Results
    3. Discussion
3. Computational Model
    1. Rational Speech Act (RSA) framework
    2. Previous Models
    3. A Fully Incremental RSA Model
    4. Qualitative Results
4. Model Implementation
    1. Model Specifications
    2. Inference
    3. Quantitative Results
5. General Discussion

INTRODUCTION

ADJECTIVE ORDERING PREFERENCES

1. a. **big white bear**
   b. white big bear



- Robust across languages (e.g. Sproat, 1991)

EXPLANATIONS FROM DIFFERENT PERSPECTIVES

- Semantic hierarchies (e.g. Dixon, 1982)
- Syntactic mapping (e.g. Cinque, 1993)
- Psycholinguistic explanations: Absoluteness (e.g. Martin, 1969) or Inherentness (e.g. Whorf, 1945)

We focus on **two recent hypotheses** with **experimental support** and common, **rationality-based** theoretical motivation (Scontras, et al. 2017; Fukumura, 2019).

MOTIVATION

SUBJECTIVITY PREDICTS ORDERING

SUBJECTIVITY hypothesis: **Less subjective** adjectives are preferred **closer to the noun** (Scontras et al., 2017).

- **Explanation**: **More efficient** expressions are integrated **earlier in semantic composition** to minimize misidentification of referents.
  (see Scontras et al. 2019, 2020; Simonic, 2018; Franke et al. 2019)
- Subjective adjectives include, e.g., gradable dimension adjectives like *big*.

Subjectivity of an adjective was operationalized by Scontras et al. (2017) as *faultless disagreement*, roughly the degree to which two speakers can disagree about attributing a property to an entity without one of them necessarily being wrong .

DISCRIMINATORY STRENGTH AFFECTS ORDERING

DISCRIMINATORY STRENGTH hypothesis: More discriminatory adjectives are preferred earlier in the linear sequence. (Fukumura, 2019).

- **Explanation**: More efficient expressions are produced earlier in the linear sequence to maximize informativity.
- Adjectives have maximal discriminatory strength in a given context if they single out a referent perfectly.
- It can be defined as $\frac{1}{|[\![w]\!]^c|}$ (cf. Frank & Goodman, 2012) or, more generally, $P(r|w;c)$

COMPARING THE TWO HYPOTHESES

- Both are based on efficient communication.
- Both assume *early* use of informative expressions.
- Both have empirical supports.

HOWEVER...

DIFFERENT PERSPECTIVES TAKEN

Speaker who describes... Listener who identifies...

...an intended referent...

...incrementally... ...sequentially...

...based on...

...linear sequence. ...hierarchical structure.

COMPARING THE TWO HYPOTHESES

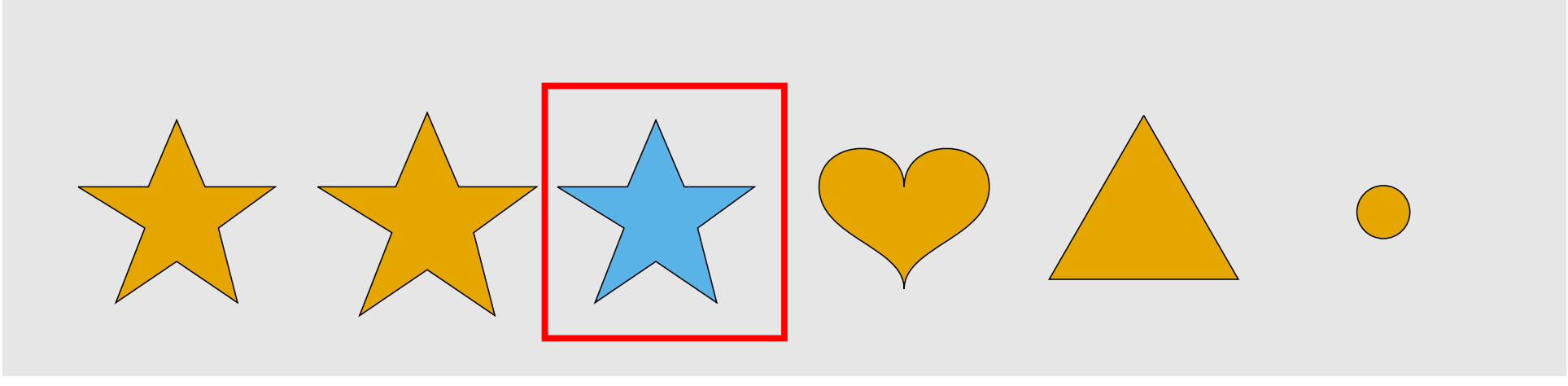
- Both are based on efficient communication.
- Both assume *early* use of informative expressions.
- Both have empirical supports.
- Due to different perspectives (listener vs. speaker) *early* means different things (close to vs. far from noun in prenominal modification).

MAIN EMPIRICAL QUESTION

What happens if the two hypotheses stand in direct **conflict** in referential visual context?

An example would be a context where a less subjective adjective discriminates more strongly between potential referents.

DIRECT CONFLICT BETWEEN HYPOTHESES



2. a. big blue star

b. blue big star

CAN THE TWO PERSPECTIVES BE COMBINED?

- **DISCRIMINATORY STRENGTH** was not tested in the presence of subjectivity gradients.
- **SUBJECTIVITY** was not tested in a referential task.
- Explanations for **SUBJECTIVITY** regarding efficiency are simulations based on referential communication.
- Unclear if **SUBJECTIVITY** would be challenged by 'opposite adaptation' with referential task of behavioral experiments.

EXPERIMENT

WEB-BASED EXPERIMENT: PREFERENCE RATINGS

Wie fragen Sie?

Brauchst du den großen blauen Aufkleber?

Brauchst du den blauen großen Aufkleber?

☐ Keine Beschreibung passt.

Weiter mit der Leertaste.

GLOSSES

The question below visual contexts as part of the cover story in the current experiment (see. Fig. 1)

a. Wie fragen Sie?
how ask you
'How do you ask?'

MORE GLOSSES

...and questions on both sides of the slider for rating

a. Brauchst du den großen blauen Aufkleber?
need you the big blue sticker
‘ Do you need the big blue sticker?’

b. Brauchst du den blauen großen Aufkleber?
need you the blue big sticker
‘ Do you need the blue big sticker?’

MIXED DESIGN

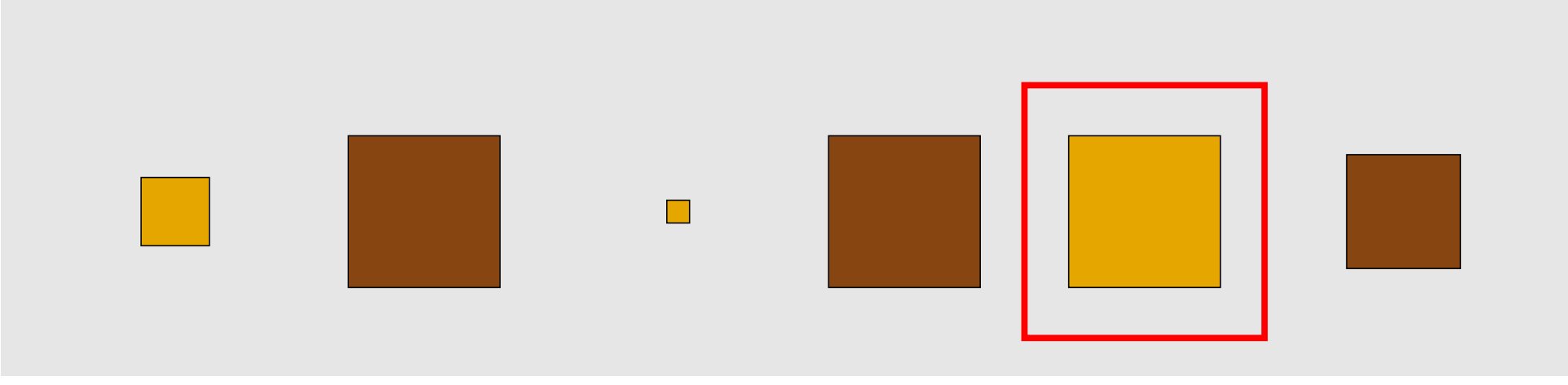
- Within factors:
  1. COMBINATION of adjectives from different semantic classes

     (2 levels: *dimension & color/shape* versus *color & shape*)
  2. RELEVANCE of the corresponding properties for reference resolution

     (3 levels: *first*, *second* or *both* properties relevant)
- Between factor
  3. SIZE DISTRIBUTION of objects

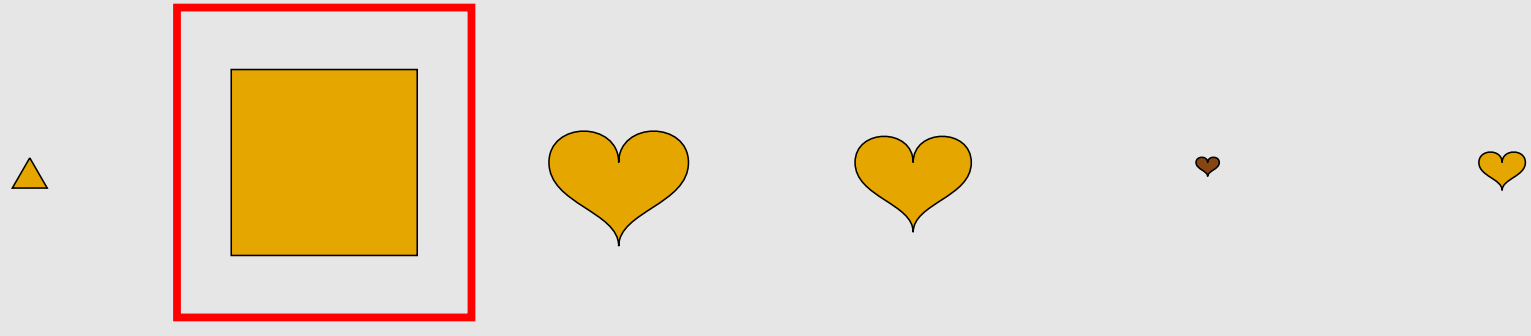
     (2 levels: *sharp* vs. *blurred*; controlling 'contextual subjectivity')

brdc – blurred

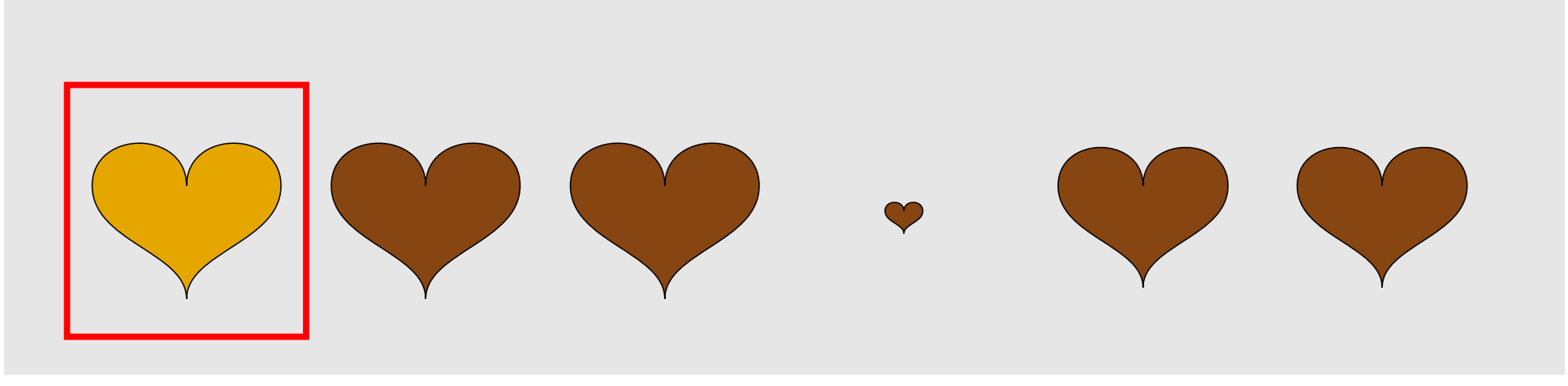


Scripts for generating these items are modified on the scirpt Generating Dot Arrays for Psycholinguistic Experiments by Shane Steinert-Threlkeld using pycairo 1.15.10

frdc – blurred



srdc – blurred



MIXED DESIGN

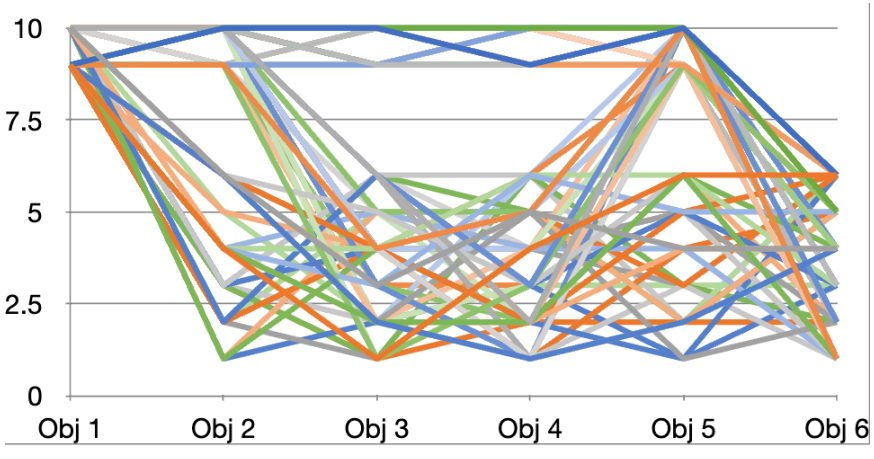
- Within factors:
  1. COMBINATION of adjectives from different semantic classes

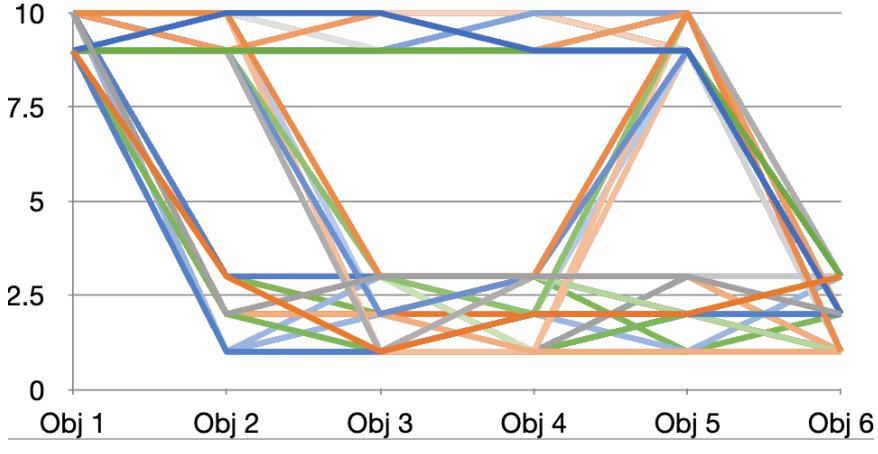
     (2 levels: *dimension & color/shape* versus *color & shape*)
  2. RELEVANCE of the corresponding properties for reference resolution

     (3 levels: *first*, *second* or *both* properties relevant)
- Between factor
  3. SIZE DISTRIBUTION of objects

     (2 levels: *sharp* vs. *blurred*; controlling 'contextual subjectivity')

Large objects of sizes 9 or 10 (in a somewhat arbitrary unit).





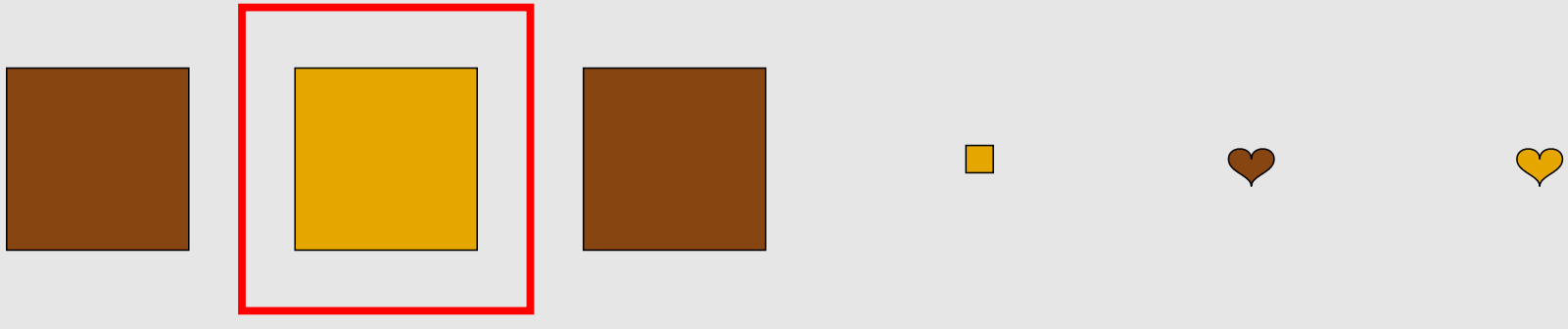
Blurred SIZE DISTRIBUTION

- Small objects of sizes: $[1, 6]$

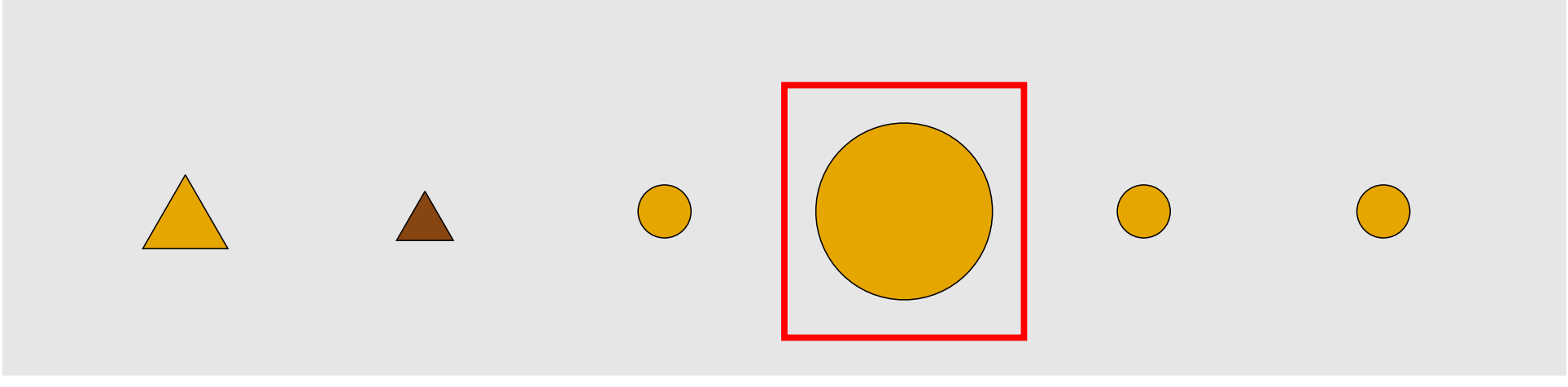
Sharp SIZE DISTRIBUTION

- Small objects of sizes: $[1, 3]$

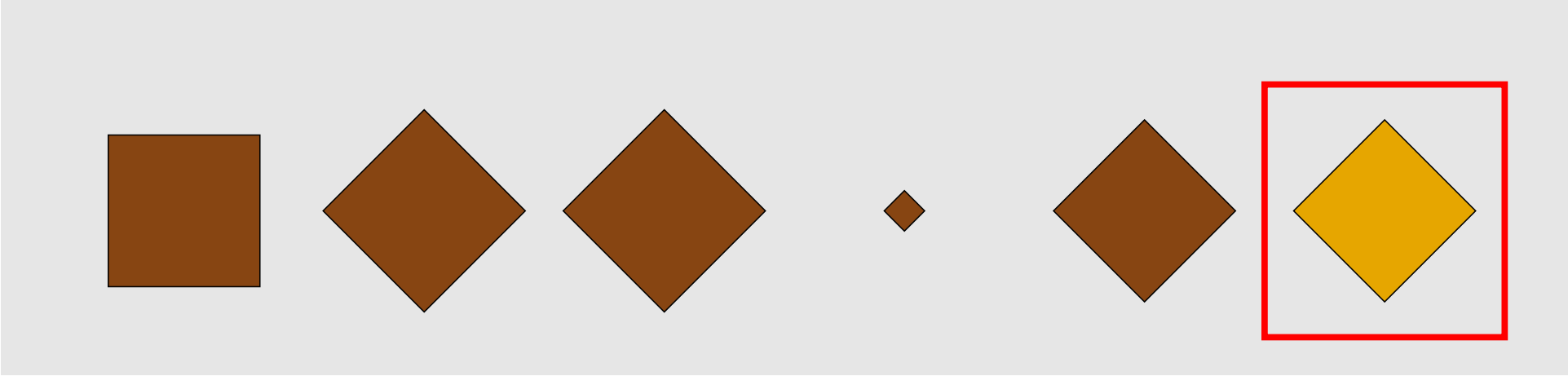
brdc – sharp



frdc – sharp



srdc – sharp



120 PARTICIPANTS

recruited via prolific.co; 180 trials, 81 experimental items

PREDICTIONS

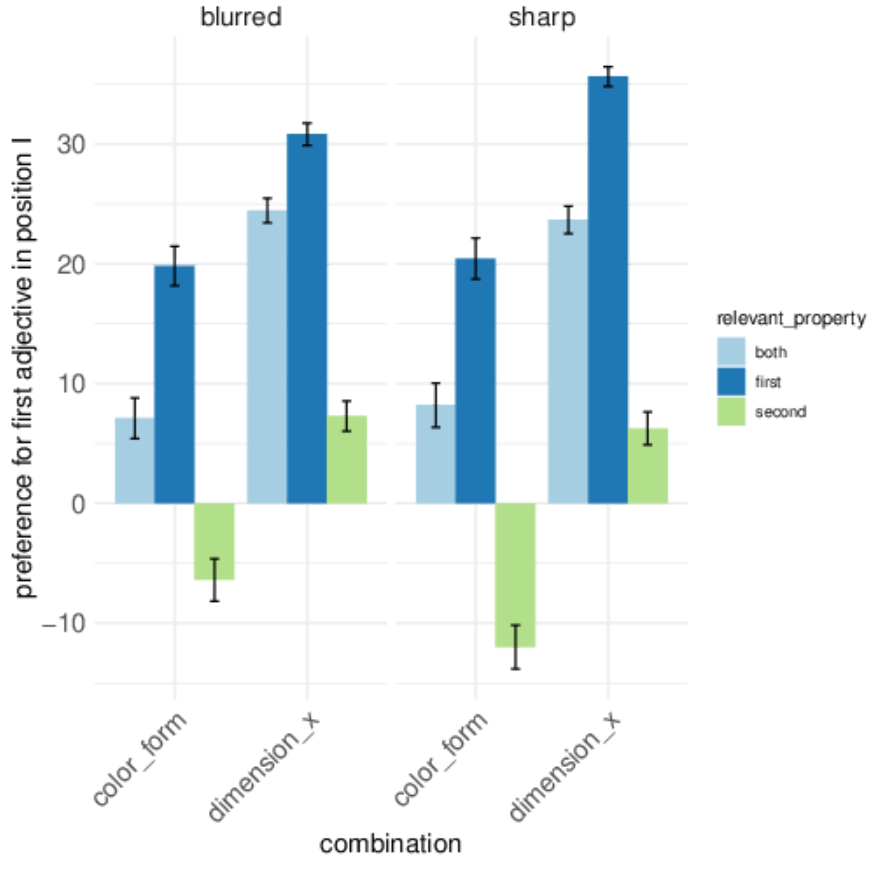
Preference for...

- size-first orders if they are present in the combination (SUBJECTIVITY)
- orders with contextually more discriminatory adjectives first (DISCRIMINATORY STRENGTH)

Reduced preference for...

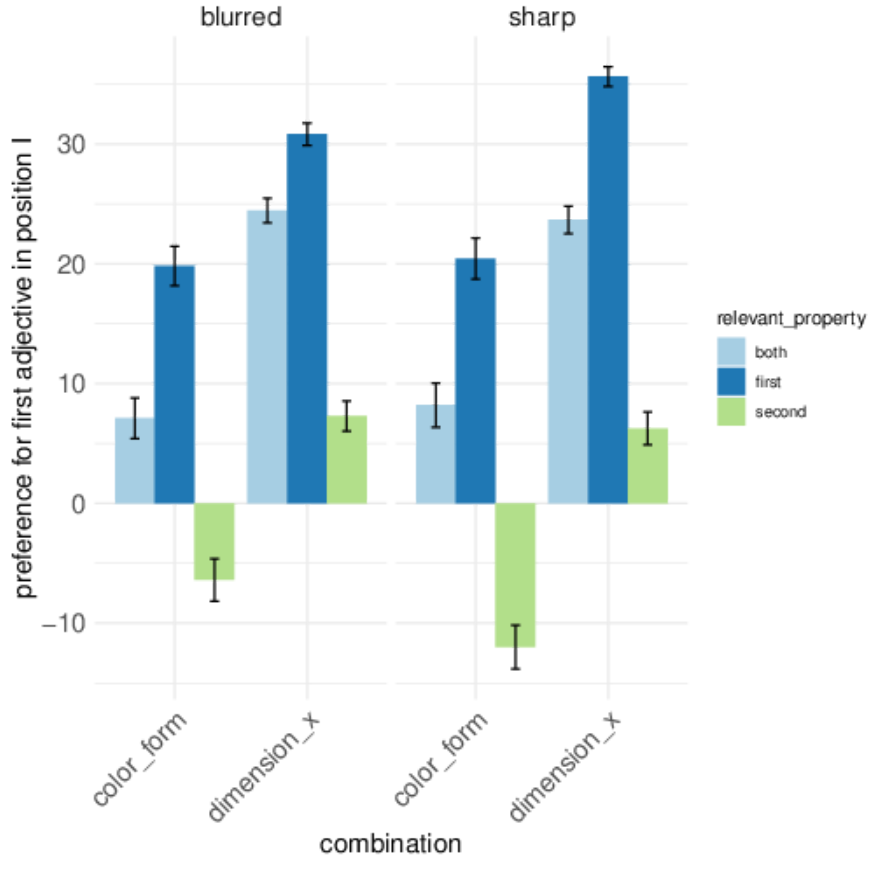
- *big*-first orders in *sharp* distributions, where size is effectively less-subjective (derived from SUBJECTIVITY but potentially in conflict with DISCRIMINATORY STRENGTH)

RESULTS



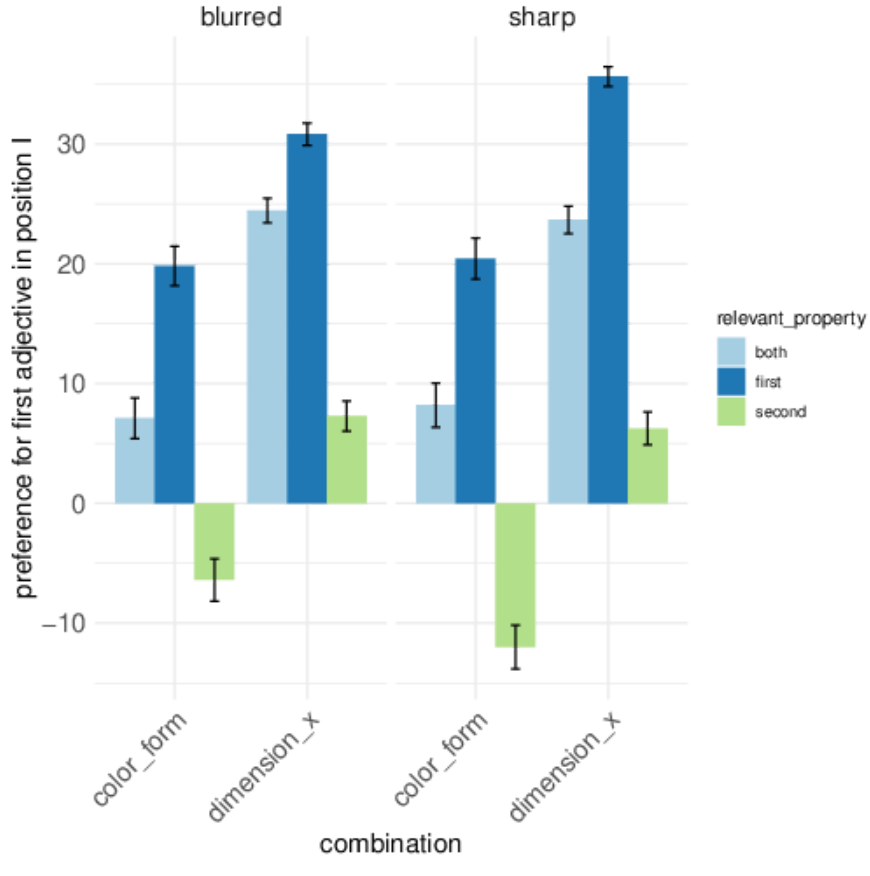
- Preference for orders with contextually more discriminatory adjectives first (as expected)

RESULTS



- Preference for size-first orders (as expected)

RESULTS



- In *big-relevant* contexts with *sharp* DISTRIBUTION, preference for size-first orders **increased – contrary to prediction from SUBJECTIVITY**

DISCUSSION

- Effects of SUBJECTIVITY and DISCRIMINATORY STRENGTH replicated

  $\Rightarrow$ Both factors contribute to ordering preferences
- Interaction indicates that preference for subjective-first ordering is increased if these adjectives are contextually more informative.

  $\Rightarrow$ Challenges explanation of SUBJECTIVITY based on low communicative efficiency

COMPUTATIONAL MODEL

PHENOMENA WE WANT TO MODEL

- SUBJECTIVITY
  (suggests listener perspective)
- DISCRIMINATORY STRENGTH
  (suggests speaker perspective)

RATIONAL SPEECH ACT (RSA) FRAMEWORK (FRANK & GOODMAN, 2012)

- Literal listener infers the likelihood of an intended referent r given utterances u:
  - by applying the literal meaning of u on r: $[\![u]\!](r)$
  - by combining prior expectations about r: $P(r)$

$$L_0(r|u) \propto [\![u]\!](r) \cdot P(r)$$

- Pragmatic speaker infers the likelihood of an intended utterance u given referents r:
  - by maximizing informativity: $\log L_0(r|u)$
  - by minimizing production cost: $C(u)$
  - governed by rationality parameter: α

$$S_1(u|r) \propto \exp(\alpha \cdot (\log L_0(r|u) - C(u)))$$

LIMITATION OF APPLYING RSA TO OUR EMPIRCAL DATA AND SOME PREVIOUS MODELS

- Insensitive to **overinformativeness** (e.g. in size-relevant context)
  If two utterances are both true, speaker will always prefer the shorter one due to the cost term
  $\Rightarrow$ a continous semantic (cf. Degen et al. 2020)
- Insensitive to **order**
  Their order in computation does not matter due to the law of commutativity (i.e. $[\![u_A]\!](r) \cdot [\![u_B]\!](r) = [\![u_B]\!](r) \cdot [\![u_A]\!](r)$)
  $\Rightarrow$ sequential context update & context-dependent semantics (cf. Simonic, 2018; Scontras et al., 2019; Franke et al., 2019)
- Insensitive to **unfinished, word-level interpretation**
  Both speaker and listener are based on completed utterances.
  $\Rightarrow$ Incrementality (cf. Cohn-Gordon et al., 2019; Waldon & Degen, 2021; Yu, Waldon & Degen, 2023)

MEANING FUNCTIONS

For color adjectives:

$$[\![blue]\!] = \lambda x. \begin{cases} \epsilon & \text{if x is blue,} \\ 1 - \epsilon & \text{if x is not blue} \end{cases}$$

where ϵ relates to the model parameter: $color\_semvalue$ (see below)

MEANING FUNCTIONS

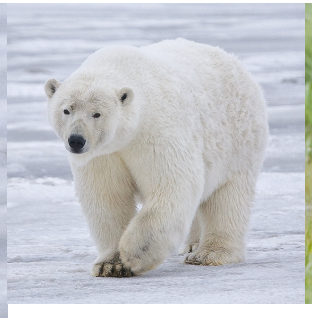
For size adjectives: k%-semantics (Schmidt et al. 2009, Cremers et al. 2022)

Tall iff its height exceeds that **threshold** of k% of the objects in the context $supp$

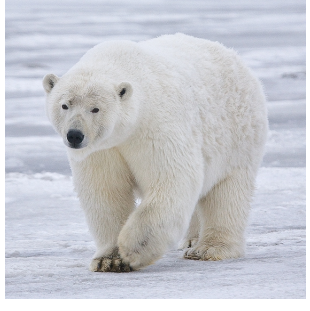
$$[\![big]\!]^{supp} = \lambda x.size(x) > max(supp) - k * (max(supp) - min(supp))$$

Deviation from the ground truth: **"Perceptual Blur"** according to the Weber-Fechner (wf) law (similar as implemented in van Tiel et al. 2021)

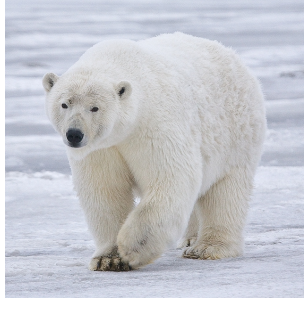
SEQUENTIAL CONTEXT UPDATE


SEQUENTIAL CONTEXT UPDATE: BEAR

SEQUENTIAL CONTEXT UPDATE: WHITE BEAR



SEQUENTIAL CONTEXT UPDATE: BIG WHITE BEAR

IDEAS FOR AN INCREMENTAL RSA MODEL

(Schlotterbeck & Wang 2023)

- Incremental listener interprets from $\overleftarrow{\text{right to left}}$

  (i.e. restricts potential referents sequentially, in accordance with preferred interpretation).
- Incremental speaker maximizes informativity at each word from $\overrightarrow{\text{left to right}}$.

ILLUSTRATION: TWO WORDS EXAMPLE (I = 2)

$$S(\emptyset, i = 2)$$

$$S(\emptyset, i = i - 1 = 1)$$

$$S(w_1, i = 1)$$

where word w comes from an extremely simple grammar:

$$B \rightarrow aA$$

$$A \rightarrow bB$$

$$A \rightarrow \emptyset$$

$$L(w_1, i = 1) \propto [\![w_1]\!]^{supp}(r) \cdot P(w_1)$$

where $supp$ refers to the support of P(r) (i.e. the current context)

$$S(w_1, i = 1) \propto \exp(\alpha \cdot (\log L(r|w_1) - C(w_1)))$$

$$S(w_1, i = 1 + 1 = 2)$$

$$S(w_2\ w_1, i = 2)$$

Again, w_2 comes from a LM

$$L(w_2\ w_1, i = 2)$$

$$L(w_1, i = 2 - 1 = 1)$$

$$L(w_1, i = 1) \propto [\![w_1]\!]^{supp}(r) \cdot P(w_1)$$

$$L(w_2\ w_1, i = 1 + 1 = 2)$$

$$L(w_2\ w_1, i=2) \propto [\![w_1]\!]^{supp(L(w_1, i=1))}(r) \cdot L(w_1, i=1)$$

$$S(w_2\ w_1, i = 2) \propto \exp(\alpha \cdot (\log L(w_2\ w_1, i = 2) - S(w_1, i = 1)))$$

Notice $C(w)$ and $P(w)$ are interchangeable (for proof, see Scontras et al. 2021, Appendix III)

GENERALISATION FOR FULL INCREMENTALITY

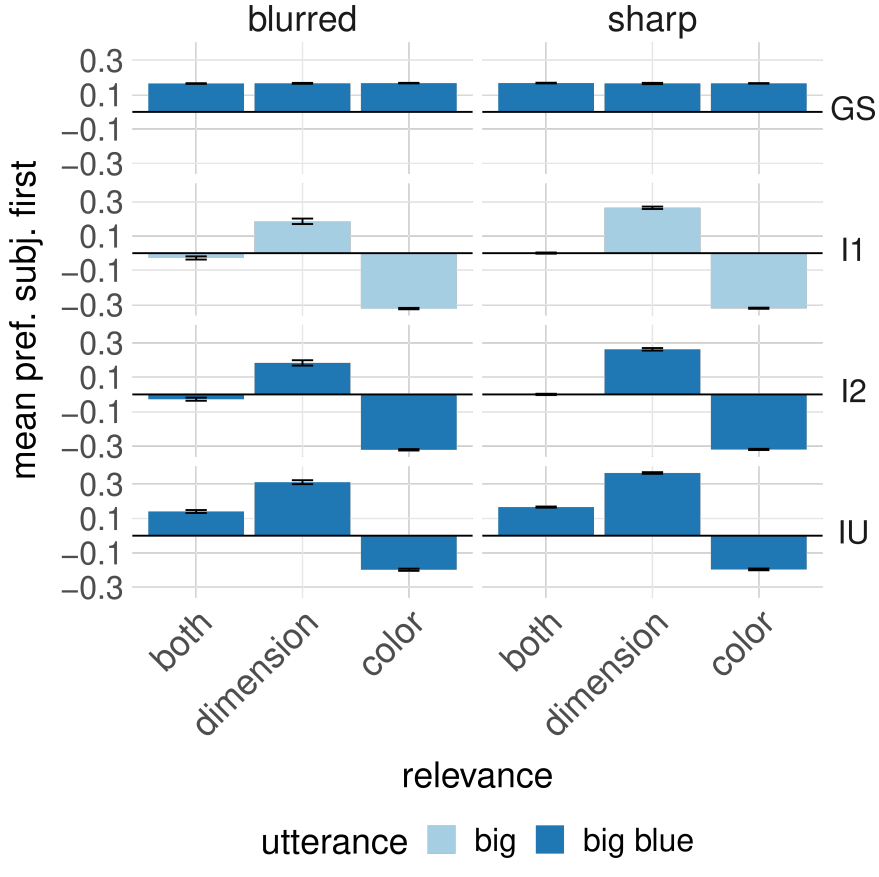
(1)	Incremental Listener	$L_0^{inc}(r\|w_{i,n})$	$\propto [\![w_i]\!]^{\text{supp}(L_0^{inc}(\cdot\|w_{i+1,n}))}(r) \cdot L_0^{inc}(r\|w_{i+1,n})$
(2)		$L_0^{inc}(r\|w_n)$	$\propto [\![w_n]\!]^{\text{supp}(P)}(r) \cdot P(r)$
(3)	Global Speaker	$S_1(w_{1,n}\|r)$	$\propto \mathbb{U}(w_{1,n}; r) \cdot P(w_{1,n})$
(4)	Incremental Sequence	$S_1^{inc}(w_{1,n}\|r)$	$\propto \mathbb{U}(w_{1,n}; r) \cdot P_{Lang}(w_n\|w_{1,n-1}) \cdot S_1^{inc}(w_{1,n-1}\|r)$
(5)	Speaker	$S_1^{inc}(w_1\|r)$	$\propto \mathbb{U}(w_1; r) \cdot P_{Lang}(w_1\|\varnothing)$
(6)	Incremental Utterance Speaker	$S_1^{inc\_utt}(w_{1,n}\|r)$	$\propto \exp(\alpha \cdot (\log(S_1^{inc}(w_{1,n}\|r)))) \cdot P(w_{1,n})$
(7)	Utility	$\mathbb{U}(\overrightarrow{w}; r)$	$= \exp(\beta \cdot (\log(L_0^{inc}(r\|\overrightarrow{w})) - \text{c}(\overrightarrow{w})))$

In all definitions above:

- r stands for a referent;
- w_i, $w_{i,n}$ and $\overrightarrow{w}$ stand for the i-th word in a sequence, a sequence of $n - i$ words and any sequence of one or more words, respectively;
- $supp(\cdot)$ denotes the support of a probability distribution;
- P denotes prior probabilities over referents and utterances;
- P_{Lang} assigns prior probabilities to potential next words;
- and, finally, α and β are rationality parameters that govern the soft-max functions defined in rows (6) and (7), respectively.

In addition we used a bias in the prior $P(w_{1,n})$ of $S_1^{inc\_utt}$.

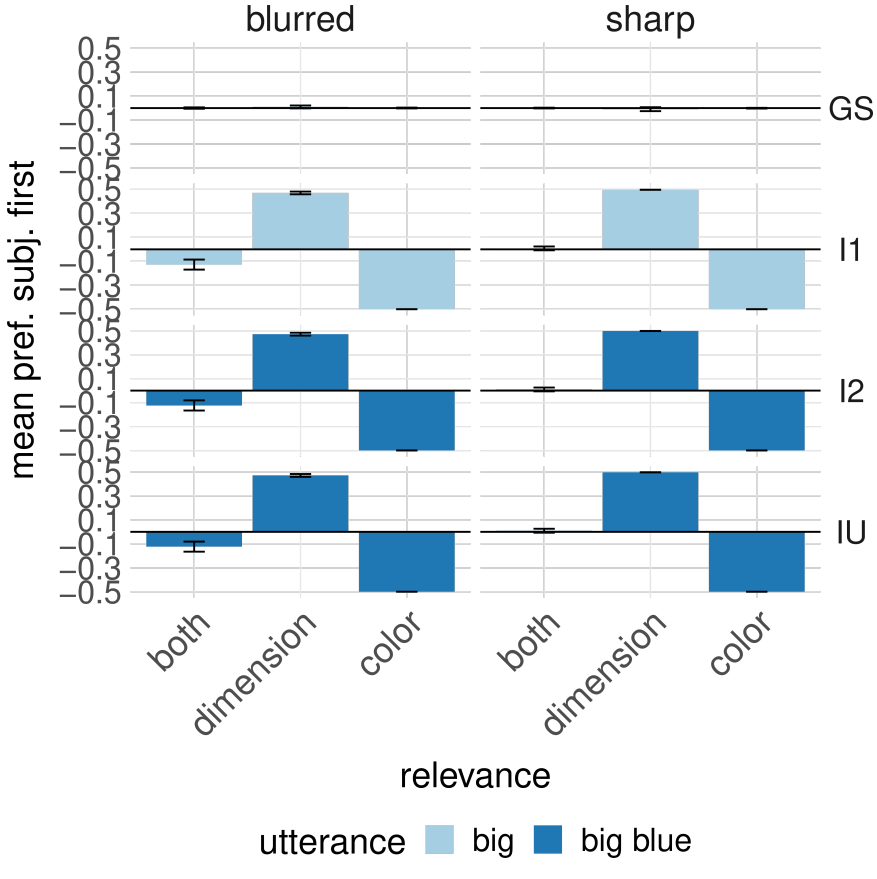
QUALITATIVE RESULTS



- Qualitative effects captured with biased incremental speaker

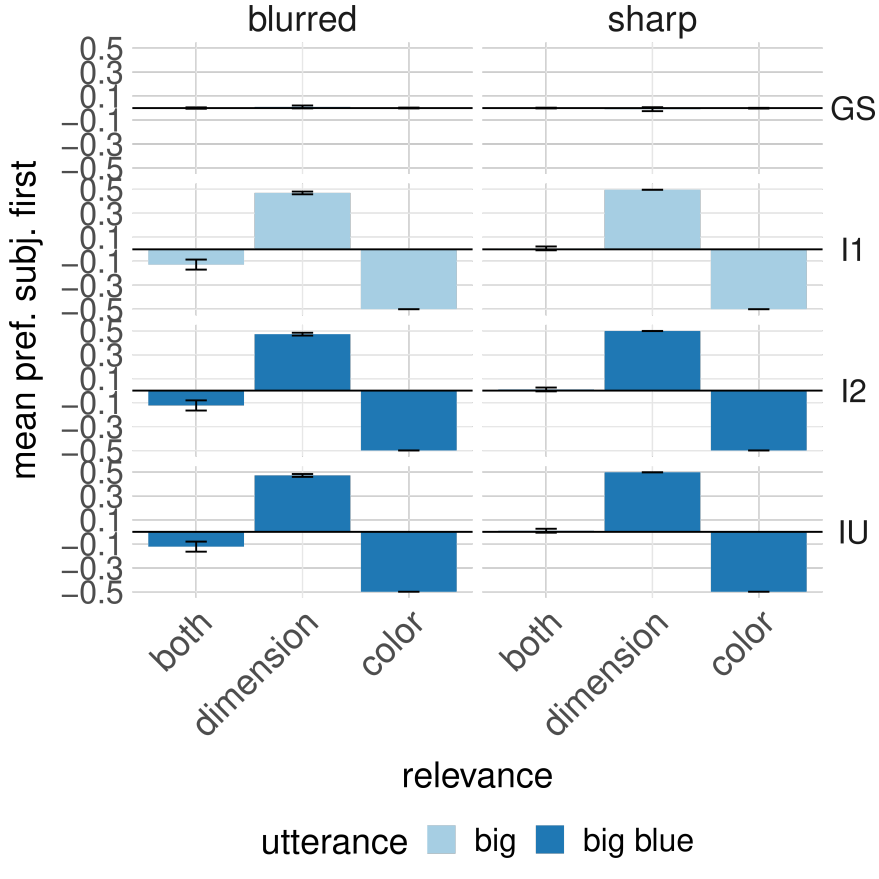
We got there in three steps...

QUALITATIVE RESULTS



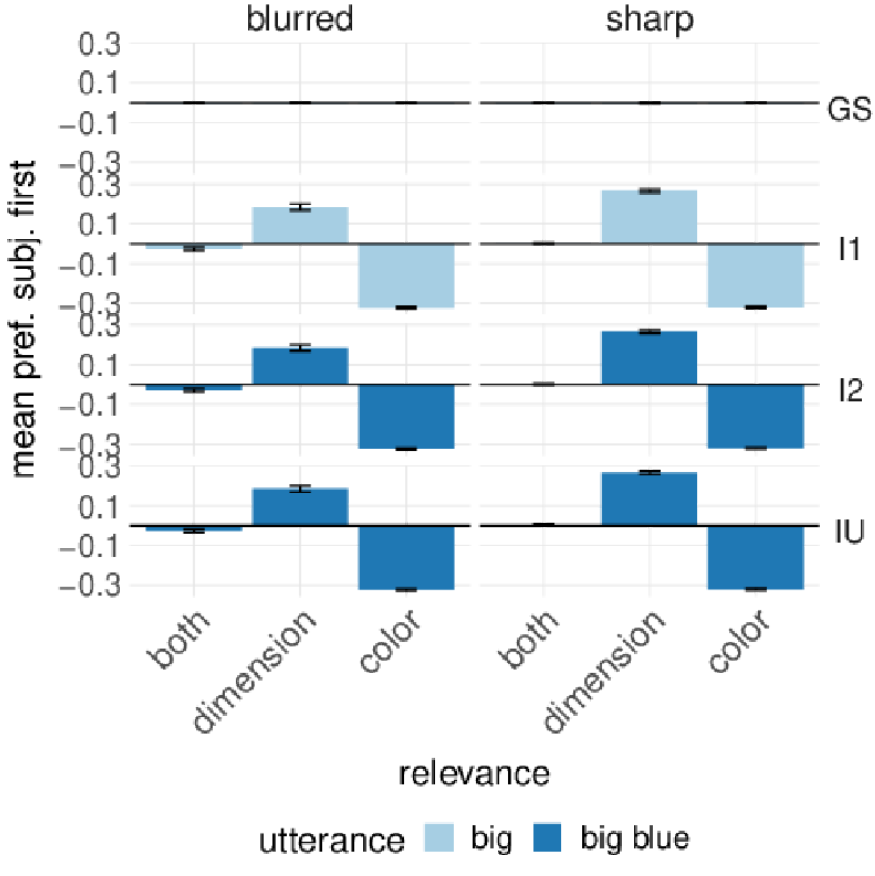
- Interaction might be there

QUALITATIVE RESULTS



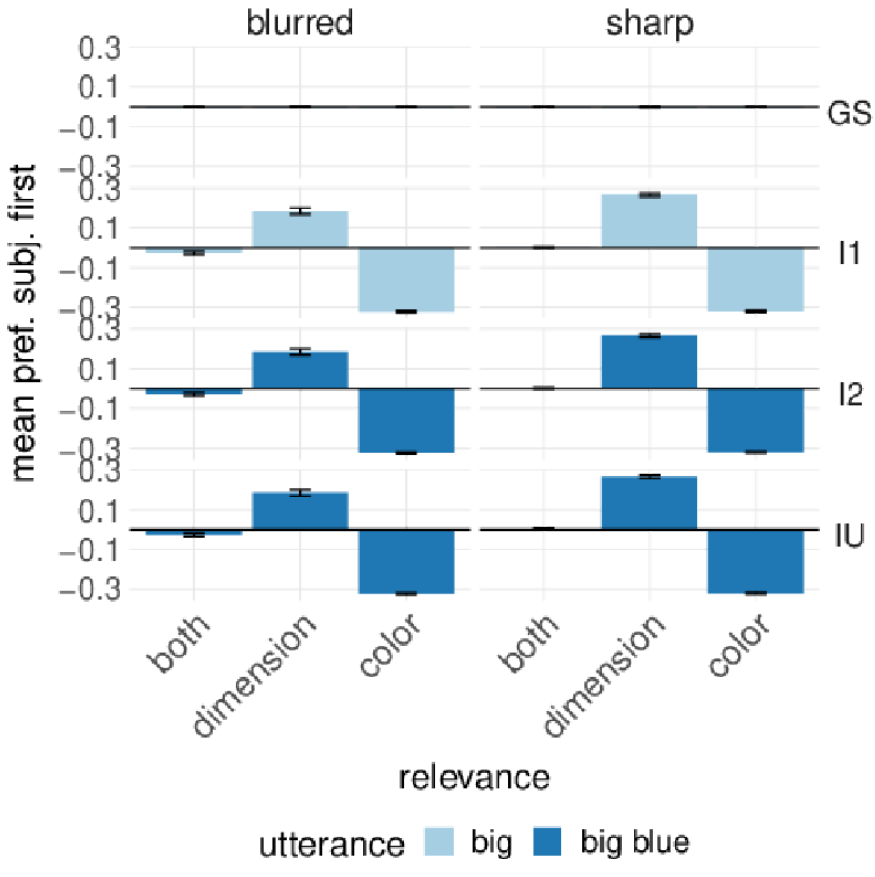
$\Rightarrow$ decrease rationality parameter α to move away from ceiling

QUALITATIVE RESULTS



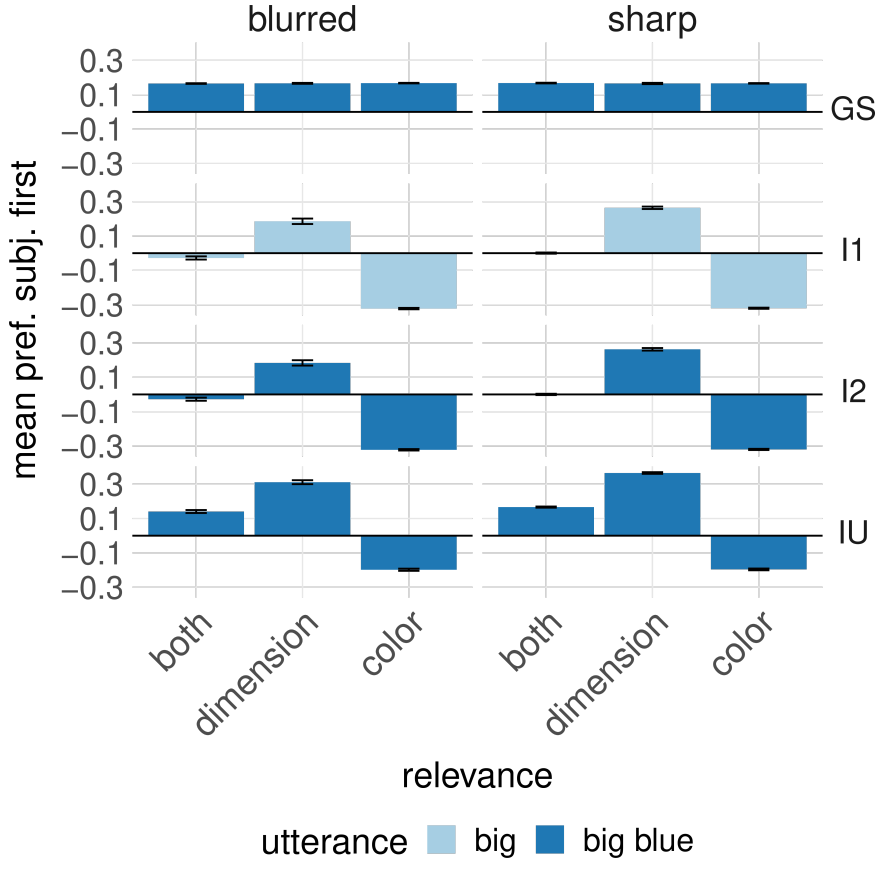
- Better, but no preference for subjective-first orders

QUALITATIVE RESULTS ON *DIMENSION_X DATA*



$\Rightarrow$ introduce *bias*

QUALITATIVE RESULTS



- Qualitative effects captured with biased incremental speaker

DISCUSSION

- Qualitative effects are more or less captured with biased incremental speaker, but there are still some deviances.

  $\Rightarrow$ Can deviances be overcome?

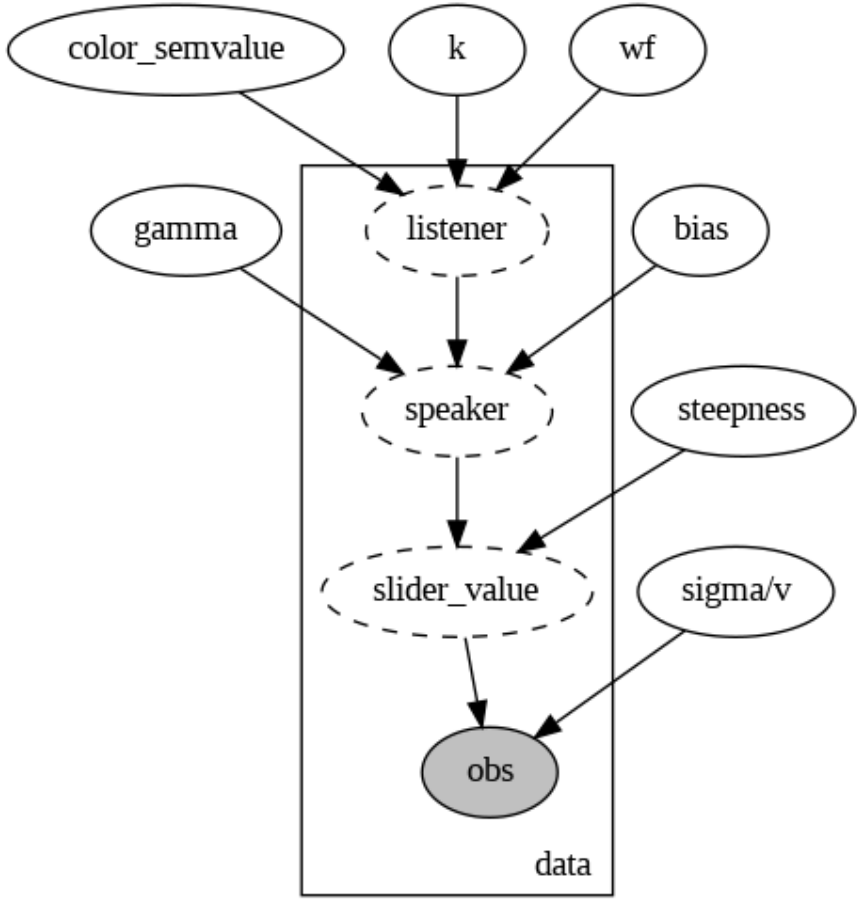
  $\Rightarrow$ Is incremental listener not needed?
- **We apply quantitative analysis to address these questions!**

MODEL IMPLEMENTATION

MODEL SPECIFICATIONS

- $color\_semvalue, k, wf \sim Uniform(0, 1)$ relate to effects from the listener perspective
- $gamma$ (i.e. rationality), $bias$ (i.e. P(w)) $\sim HalfNormal(5)$ relate to effects from the speaker perspective
- $steepness \sim Uniform(0, 1)$ relates to link function mapping predicted likelihood onto slider value
- $sigma$ (i.e. standard deviation) $\sim Uniform(0, 0.10)$ if sample distribution is $TruncatedNormal$
- v (i.e. concentration) $\sim Uniform(0, 2)$ if sample distribution is $Beta$

Sample distribution introduces model residuals.

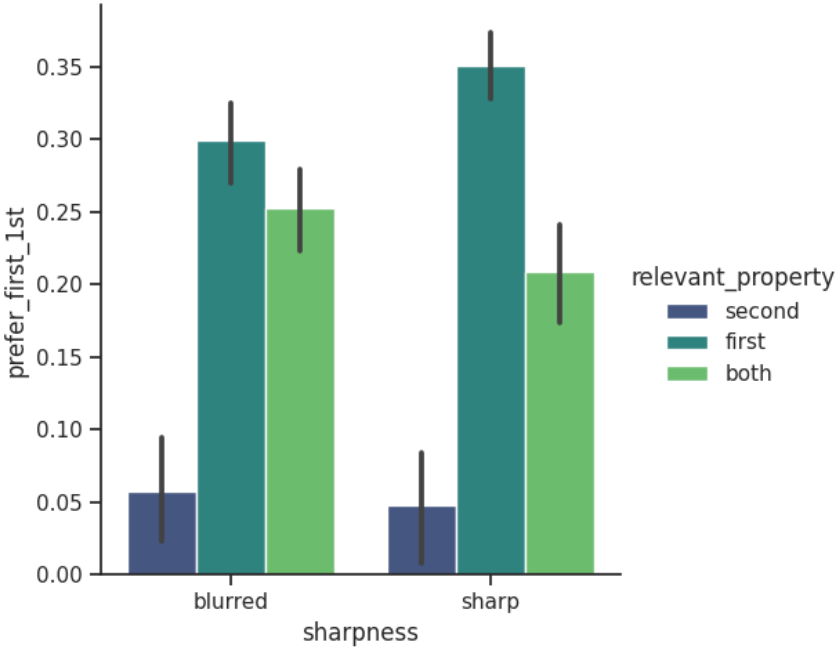


INFERENCE

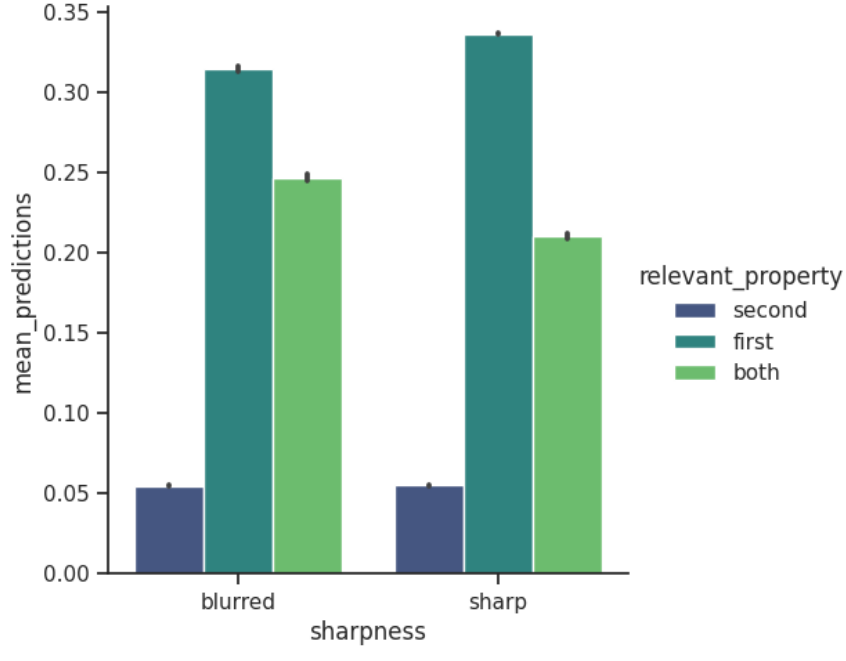
We use No U-Turn Sampler (NUTS) (numpyro 0.12.1, Phan et al. 2019, Bingham et al. 2019) to perform MCMC inference:

- Warum up size: 5000
- Sample size: 30000

RESULTS: EFFECTS CAPTURED

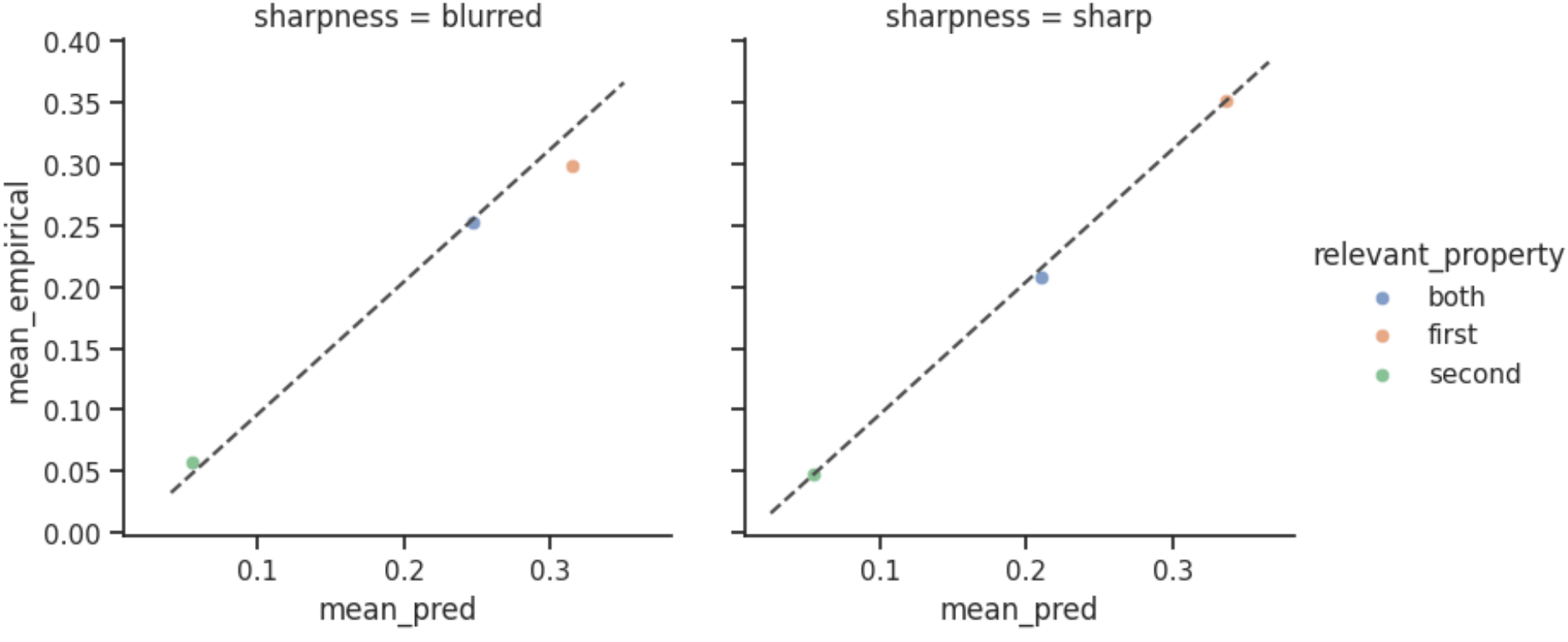


Empirical data

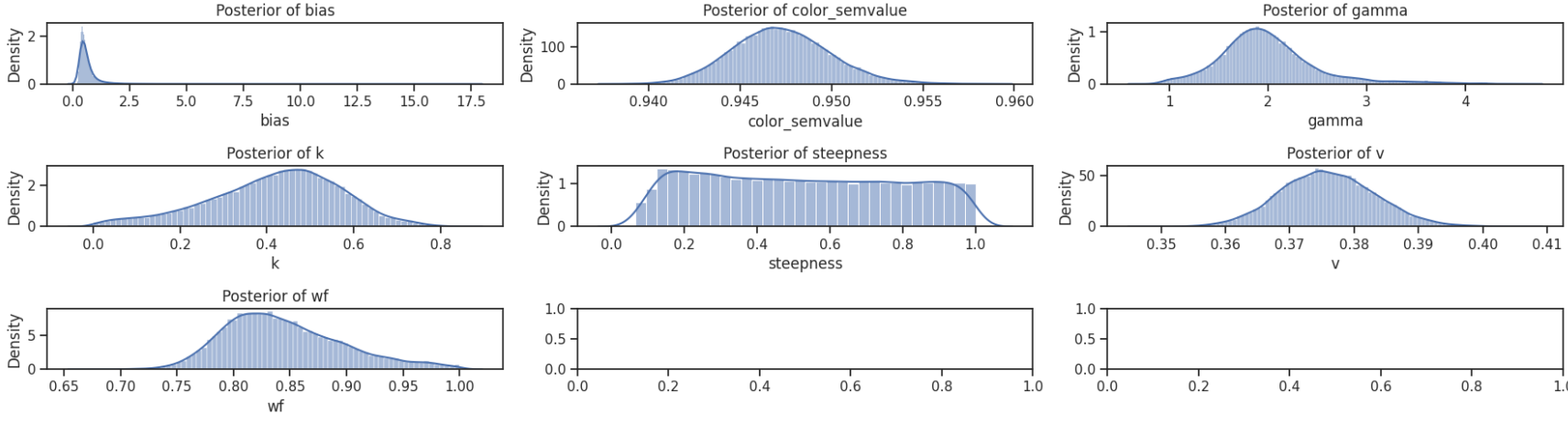


Model predicitons

CORRELATION PLOT AGAINST EMPIRICAL DATA



RESULTS: POSTERIOR DISTRIBUTIONS OF PARAMETERS



Maximum a posteriori (MAP) and 90% highest density interval (HDI) for each parameters:

bias: MAP = 0.70, HDI = [0.22, 1.07]

color semantic value: MAP = 0.95, HDI = [0.94, 0.95]

v: MAP = 0.38, HDI = [0.36, 0.39]

steepness: MAP = 0.52, HDI = [0.12, 0.93]

wf: MAP = 0.84, HDI = [0.76, 0.93]

alpha(gamma): MAP = 1.99, HDI = [1.16, 2.71]

k: MAP = 0.42, HDI = [0.16, 0.67]

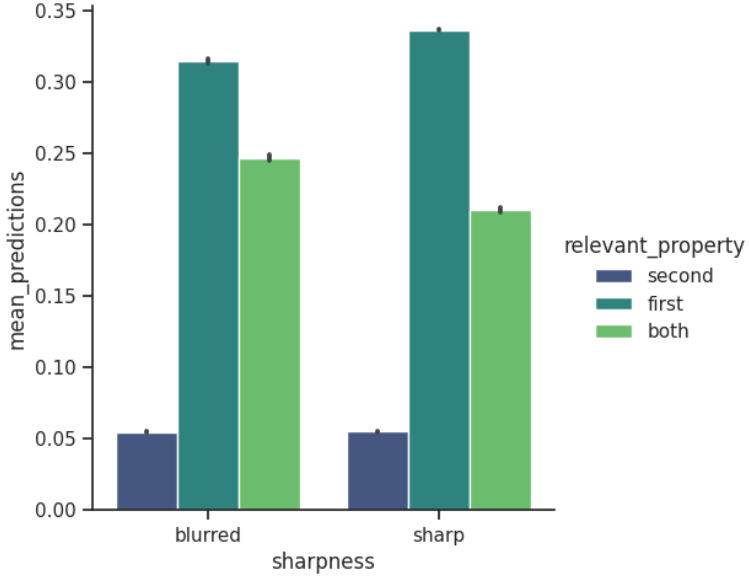
MODEL VARIANTS

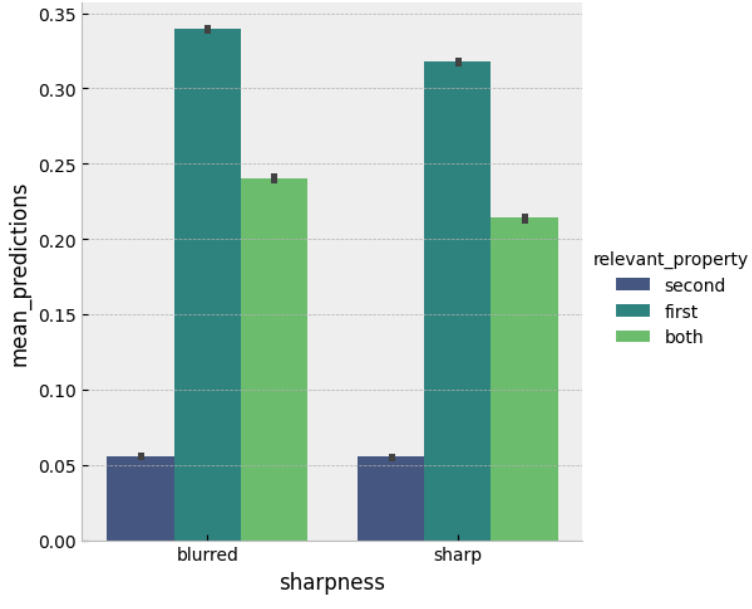
We use a Grid-Search like method to explore model variants:

Search space:

- **Speaker:**
  - incremental
  - global
- **Threshold methods of size Adj:**
  - context dependent (sampling based)
  - context-independent (support based)
- **Sample distribution:**
  - $TruncatedNormal$
  - $Beta$
- **Link function:**
  - logit
  - logistic
  - linear
  - identity

COMPARING PREDICTIONS OF MODELS WITH DIFFERENT SPEAKER





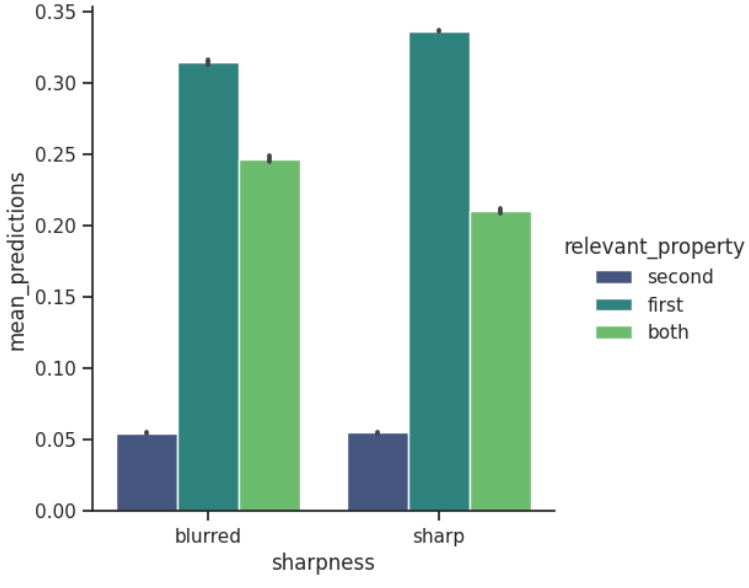
- Incremental speaker
- Non-incremental speaker

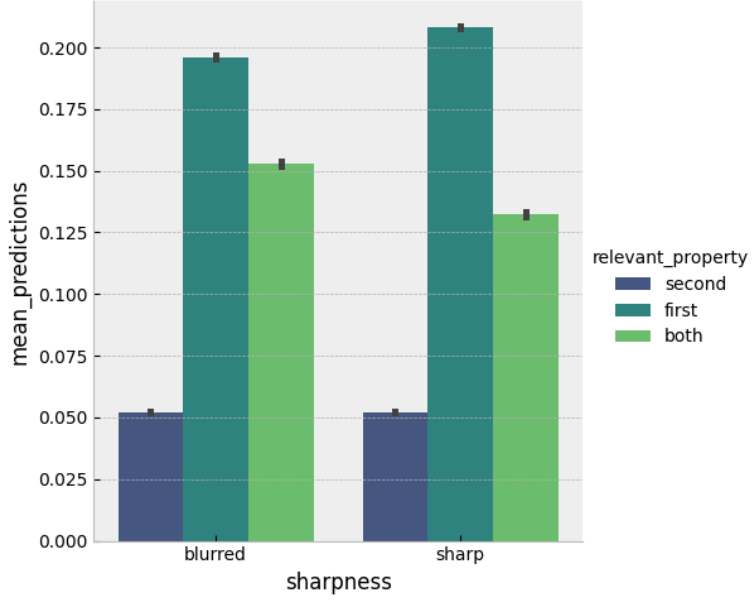
The non-incremental speaker predicts an interaction effect that contradicts the direction observed in empirical data.

MODEL COMPARISON

	Corr. Coeff.	L2	LPD
baseline: Global Speaker + f. Semantic	0.289	0.140	-6.991
Global Speaker + c. Semantic	0.290	0.140	-6.991
Inc. Speaker + c. Semantic	**0.299***	**0.140***	**-6.989***
Inc. Speaker + f. Semantic	0.299	0.140	-6.989

COMPARING PREDICTIONS OF MODELS WITH DIFFERENT SAMPLE DISTRIBUTION





- $TruncatedNormal$ distribution with logit link function
- $Beta$ distribution with logit link function

The range of predictions made by $Beta$ distribuion does not align with empircal data.

HISTOGRAM OF EMPIRICAL SLIDER VALUES

COMPARISON OF LINK FUNCTIONS AND SAMPLE DISTRIBUTIONS

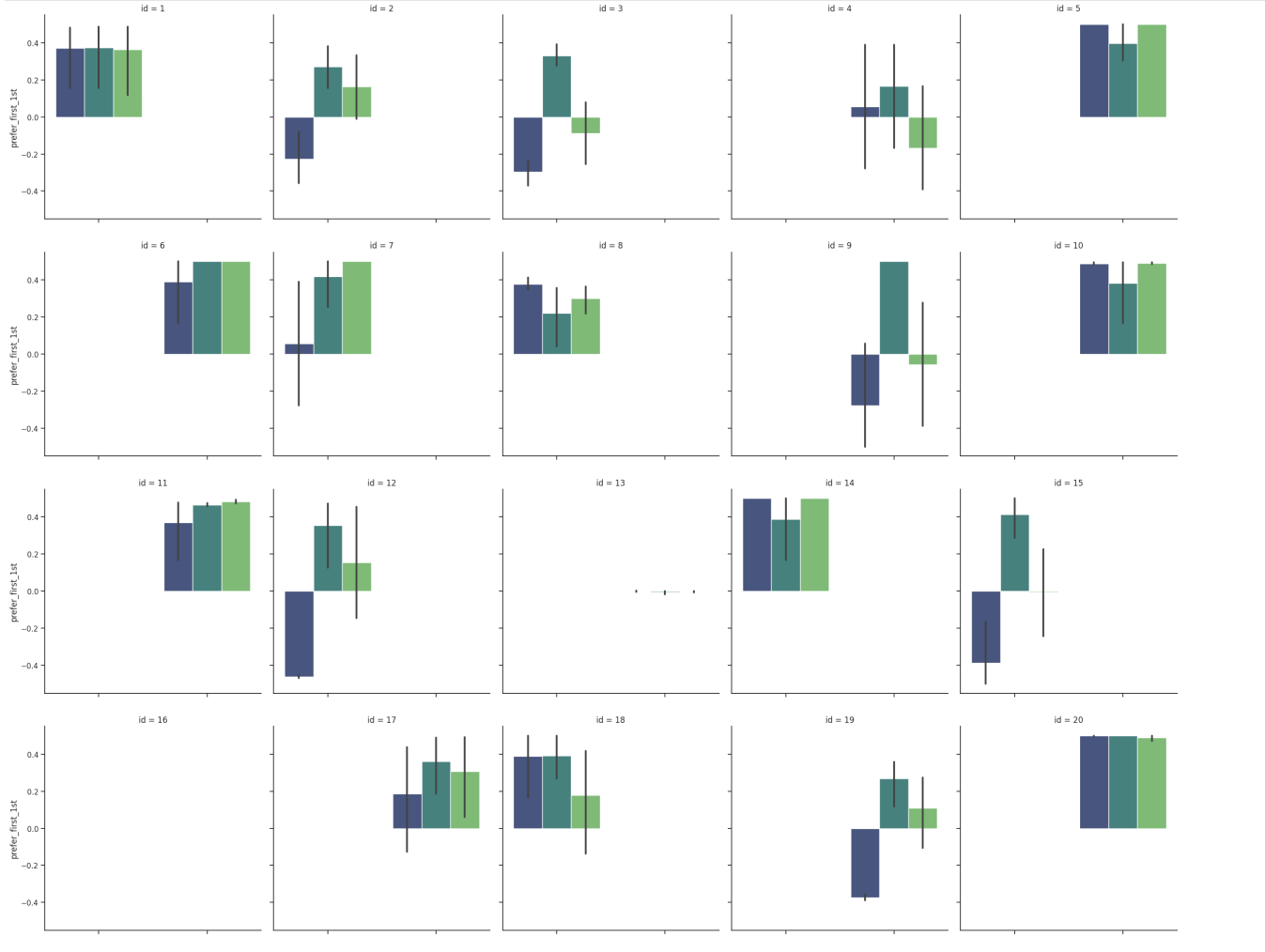
	Normal			Beta		
	Corr. Coeff.	L2	LPD	Corr. Coeff.	L2	LPD
it	0.299*	0.132	-15.540*	0.298*	0.140*	-6.9
istic	0.298	0.132	-15.544	0.005	0.186	-7.0
ear	0.299*	0.132	-15.542	0.295	0.145	-7.0
ntity	0.298	0.132	-15.542	0.298*	0.140*	-6.9

DISCUSSION

- **Incremental speaker model** can capture effects
- Unlike qualitative results, we observe effects from listener perspective **contrary** to empirical data.

  $\Rightarrow$ expected from **theory predicitons and experimental design**.
- $TruncatedNormal$ distribution provides the **right range**, but does not align with the pattern of empircal data.
- $Beta$ distribution may better describe the **true underlying distribution** of empirical data, but does not provide the right range.
- **Minor** differences exist among various link functions.
- **By participants** random effects structure may needed to account for **individual variances**.

AGGREGATED MEAN BY CONDITIONS AND PARTICIPANTS



GENERAL DISCUSSION

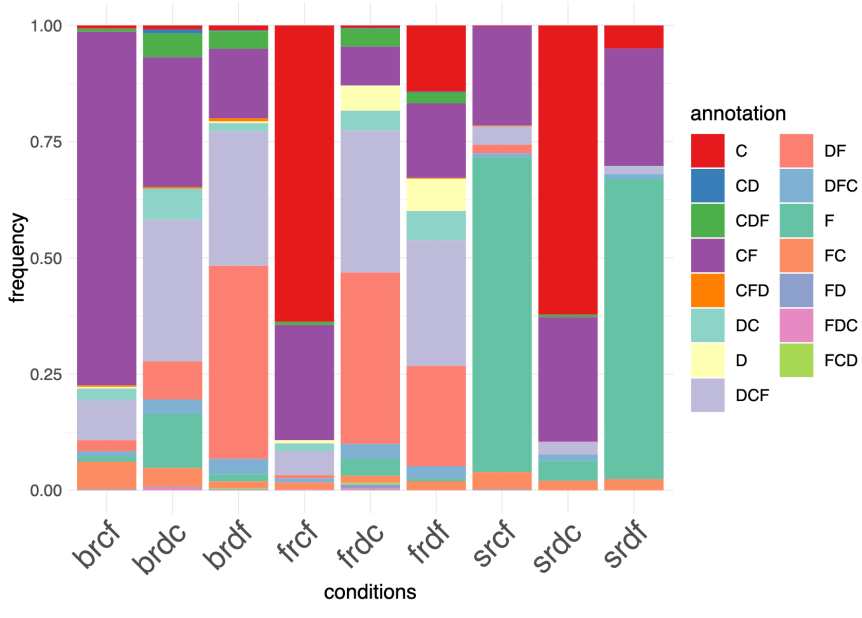
We provided results from a preference rating experiment in visual referential context, which reveal:

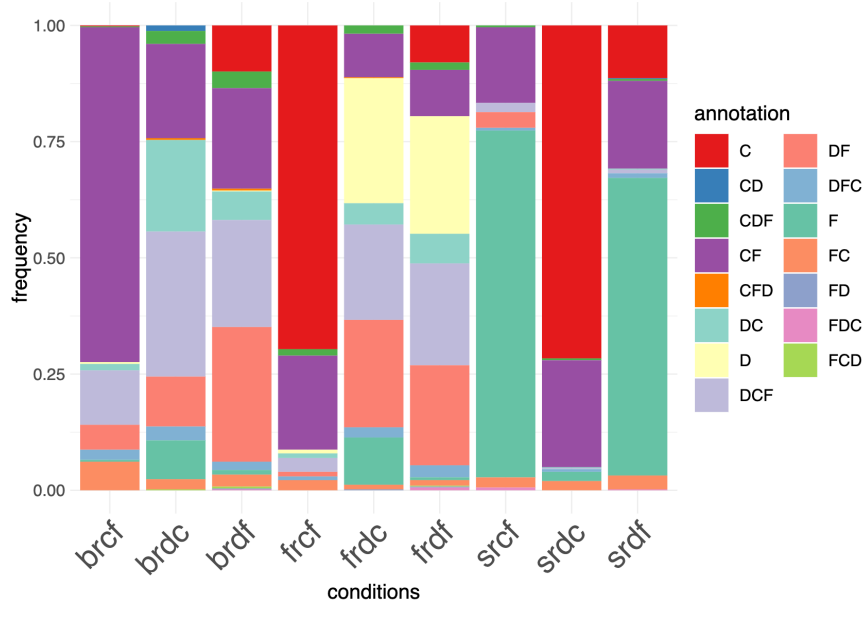
- **Both** SUBJECTIVITY and DISCRIMINATORY STRENGTH contribute to adjective ordering preference as stated in literature.
- A **complex** interaction effect necessitates further explanations beyond just relying on communication efficiency, but also some other psycholinguistic factors like salience or availability.

We also presented **a fully incremental RSA model**, and implemented it for quantitative analysis on our empirical data, which reveals:

- **Incremental speaker** can capture aggregated means.
- **Global speaker** predicts effects from listener perspective **contrary** to empirical data, but in line with **subjectivity**.

ANOTHER EXPERIMENT: A FREE PRODUCTION TASK





blurred sharp

To better understand:

- Interaction effect
- Overinformativeness

Without explicit link function

Preliminary GLMER analysis reveals:

- Similiar to slider values, SHARPNESS increases the preference for subjective-first ordering in SIZE-RELEVANT context
- However, it also increases the preference for color-first ordering in COLOR-RELEVANT context
  Furthermore, it decreases the likelihood of overinformative usage of *big* in SIZE-RELEVANT context

  $\Rightarrow$ Generative models needed!
- Are gradable dimension adjectives useful because they communicate extreme values?

NEXT STEPS

- We want to build a hierarchical model structure to incorporate random effects.
- We aim to generate random contexts, similar to those used in previous simulation-based accounts of subjectivity. Our goal is to further investigate how sequential context update in our model affect the listener's perspective.
- We aim to model data from a free production task using the same core model.
- We want to take steps towards a general model of interpretation based on truly incremental semantics (e.g. Bott & Sternefeld, 2017).

OPEN QUESTIONS

- Are gradable dimension adjectives useful in referential communication because they communicate **extreme** values?

THANK YOU!