Incremental computation of scalar implicatures in the Maze task

John Duff, Pranav Anand, and Amanda Rysling
UC Santa Cruz Linguistics

XPrag X @ Paris ❖ 21 September 2023

jduff@ucsc.edu

Incremental uncertainty about meaning

Decisions are made quickly, sensitive to heuristics and context, and costly to revise.

(context prompts early generation of non-default meaning)

The farm owners discussed the cotton. 🚩

(incremental meaning decisions are revised at cost)

The farm owners discussed the cotton. 🚩

The fabric… 🚩

(self-paced reading; Foraker & Murphy, 2012)

Fitting pragmatic meaning into the picture

(context prompts early generation of enriched meaning)

Some of the executives were fired.

∃ ∧ ~∀ 

(incremental enrichments are revised… at cost?)

Some of the executives were fired.

∃ ∧ ~∀ 

In fact, they all were.

∃ ~~∧ ~∀~~ 

?

In this talk: The waters are murky!

Review existing work on implicature generation and consequences.

Attempt to extend existing findings in SPR experiment.

➡ Some downstream effects of implicatures do not replicate.

➡ Weak evidence for context-specific generation and cancellation.

Compare performance in a Maze task.

➡ Patterns somewhat more organized, but context-insensitive.

➡ Deeper implicature consideration facilitated cancellation.

Roadmap

1. Introduction
2. **Existing work**
3. Materials
4. E1: Self-paced reading
5. E2: A-Maze reading
6. Discussion and conclusions

The costs and consequences of implicature

- **Implicature generation is costly.**
  - Slowdowns in reading at triggers. (Breheny et al. 2006, Bergen & Grodner 2012)
  - Slowdowns in picture verification. (Bott & Noveck 2004, Bott et al. 2012)
  - Generation less likely under other cognitive load.

    (De Neys & Schaeken 2007, Marty & Chemla 2013)

- **Implicature generation has consequences.**
  - In supportive contexts, implicature consistent continuations are read faster.

    (Breheny et al. 2006, Bergen & Grodner 2012)

Bergen & Grodner (2012): Design

S1

I carefully inspected the new shipment of jewelry.

I helped unload the new shipment of jewelry.

KNOWLEDGEABLE SPEAKER

NEUTRAL SPEAKER

S2

Some of _____ the gold watches were fakes.

Only some of

IMPLICATURE

ENTAILMENT

S3

The rest were real, but the company is still planning to sue.

In fact, they all were, so the company is planning to sue.

AFFIRMATION

CANCELLATION

Bergen & Grodner (2012): Results

Some of the gold watches were fakes.

S2: For IMPL and not ENTAIL, in RTs at and after *some*, KNOW > NEUT.

GENERATION COST

The rest were real, but the company…

AFFIRM: For IMPL and not ENTAIL, in RTs after subj, NEUT > KNOW.

LATER EVIDENCE OF GENERATION

CORRELATED!

In fact, they all were, so the company…

CANCEL: No difference between NEUT and KNOW.

NO EVIDENCE OF REANALYSIS

A worry: Costly belief state reanalysis?

KNOW + IMPL + CANCEL

I carefully inspected…

Some were fake…

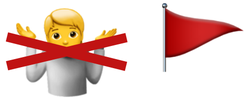
In fact, they all were…

NEUT + IMPL + CANCEL

I helped unload…

Some were fake…

In fact, they all were…



Not convincing proof of cost-free cancellation.

Roadmap

1. Introduction
2. Existing work
3. **Materials**
4. E1: Self-paced reading
5. E2: A-Maze reading
6. Discussion and conclusions

Materials

S1

Petra wrote an article about the company's response to the scandal.

Petra heard a bit about the company's response to the scandal.

KNOWLEDGEABLE PROTAGONIST

NEUTRAL PROTAGONIST

S2

She realized that ~~some of~~ only some of the marketing executives were fired.

IMPLICATURE

ENTAILMENT

S3

The rest suffered a huge pay cut, which seemed fair.

In fact, they all were, which seemed fair.

AFFIRMATION

CANCELLATION

Factivity and scalar implicature: Assumptions

KNOW

Petra wrote an article…

Pro: 🤓

She realized that some were fired…

Pro: ∃ ∧ ~∀

Spk: ∃ ∧ ~∀

NEUT

Petra heard a bit…

Pro: 🤷‍♂️

She realized that some were fired…

Pro: ∃

Spk: ∃

Norming

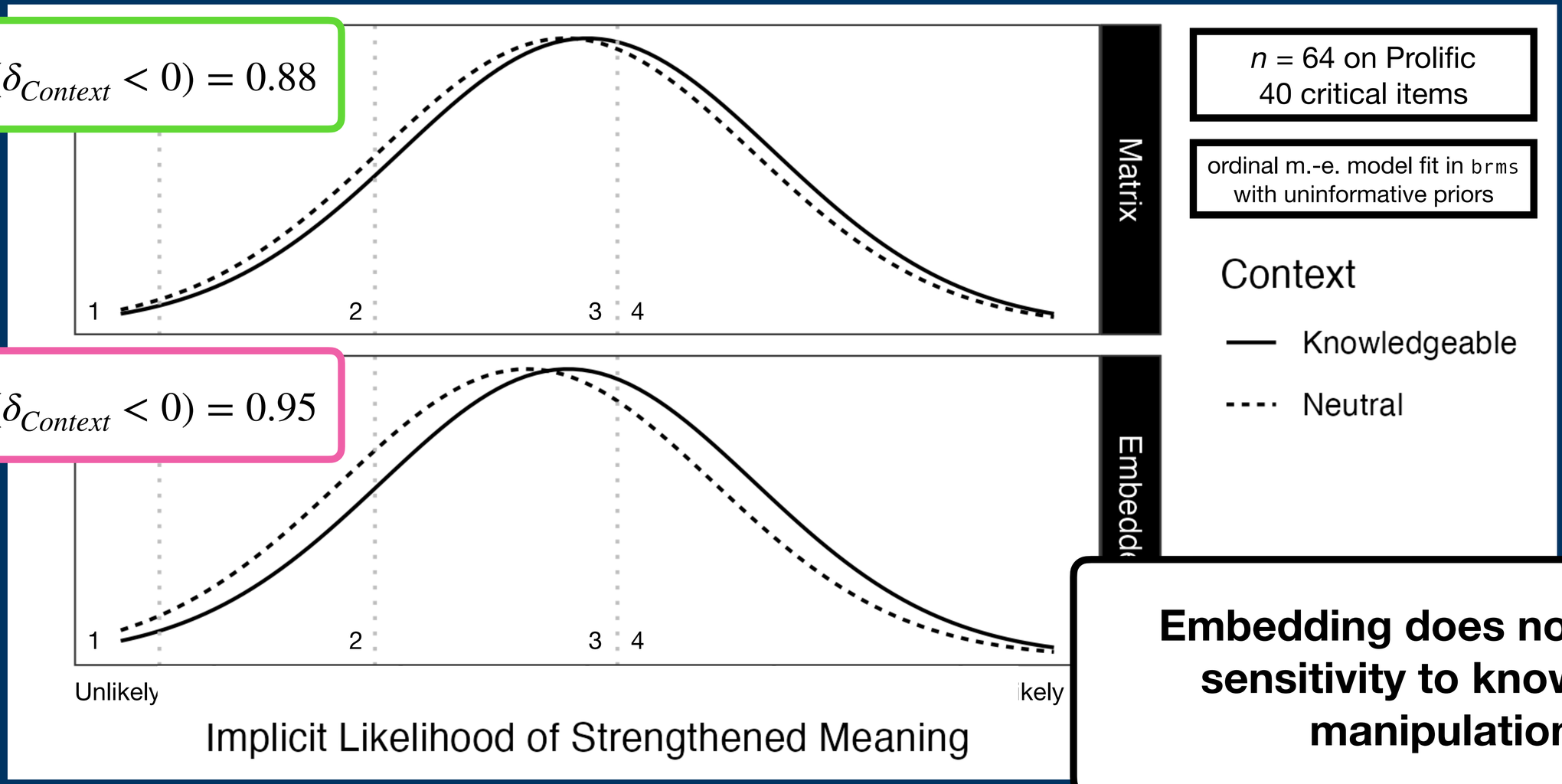
I {wrote an article, heard a bit} … Some of the executives were fired.

Petra {wrote an article, heard a bit} … She realized that some of the executives were fired.

How likely is it that at least one marketing executive kept their job?

$P(\delta_{Context} < 0) = 0.88$

$P(\delta_{Context} < 0) = 0.95$



Embedding does not reduce sensitivity to knowledge manipulation.

Roadmap

1. Introduction
2. Existing work
3. Materials
4. **E1: Self-paced reading**
5. E2: A-Maze reading
6. Discussion and conclusions

Materials

S1

Petra wrote an article about the company's response to the scandal.

Petra heard a bit about the company's response to the scandal.

KNOWLEDGEABLE PROTAGONIST

NEUTRAL PROTAGONIST

S2 She realized that some of / only some of the marketing executives were fired.

IMPLICATURE

ENTAILMENT

S3

The rest suffered a huge pay cut, which seemed fair.

In fact, they all were, which seemed fair.

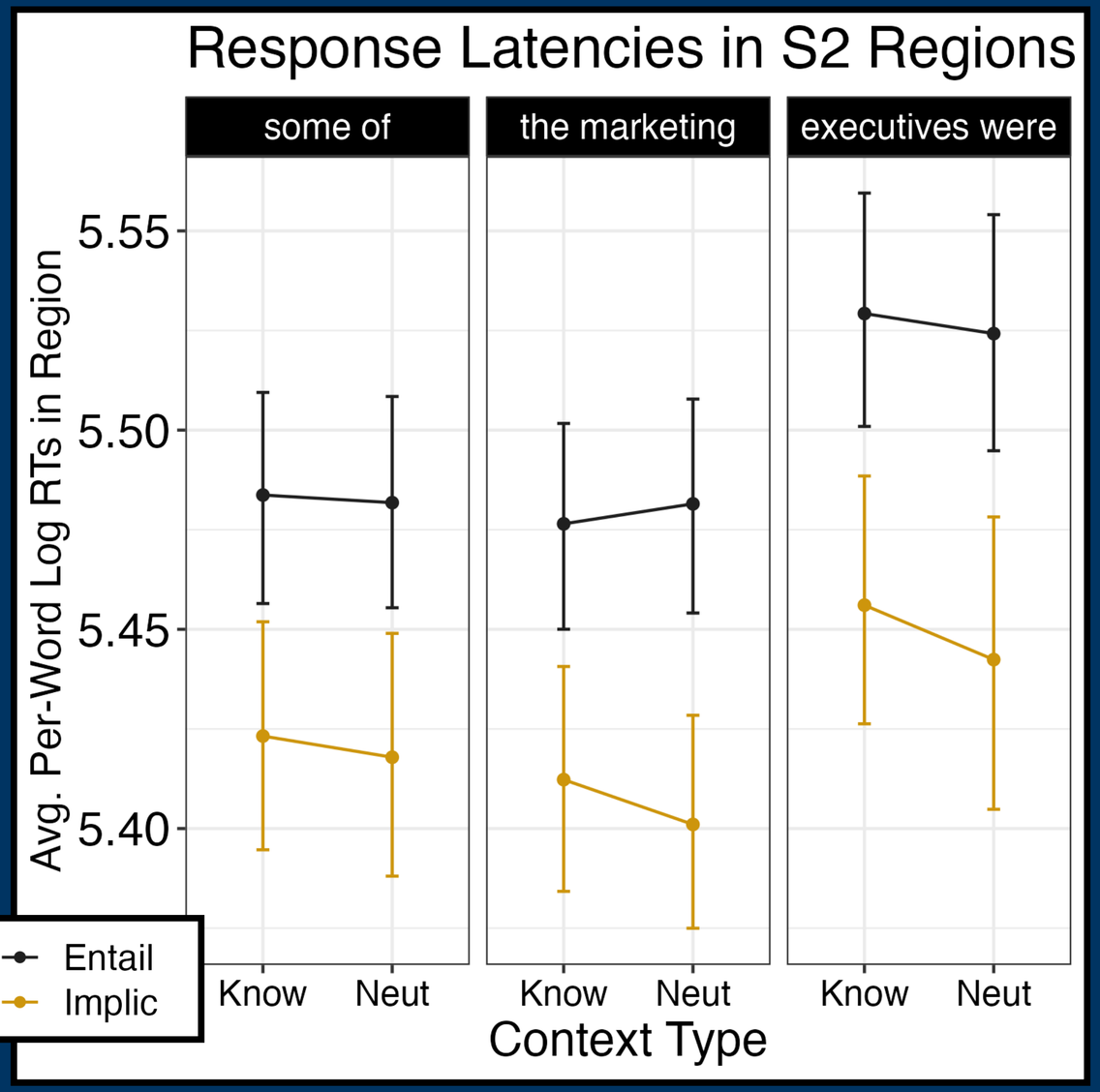
AFFIRMATION

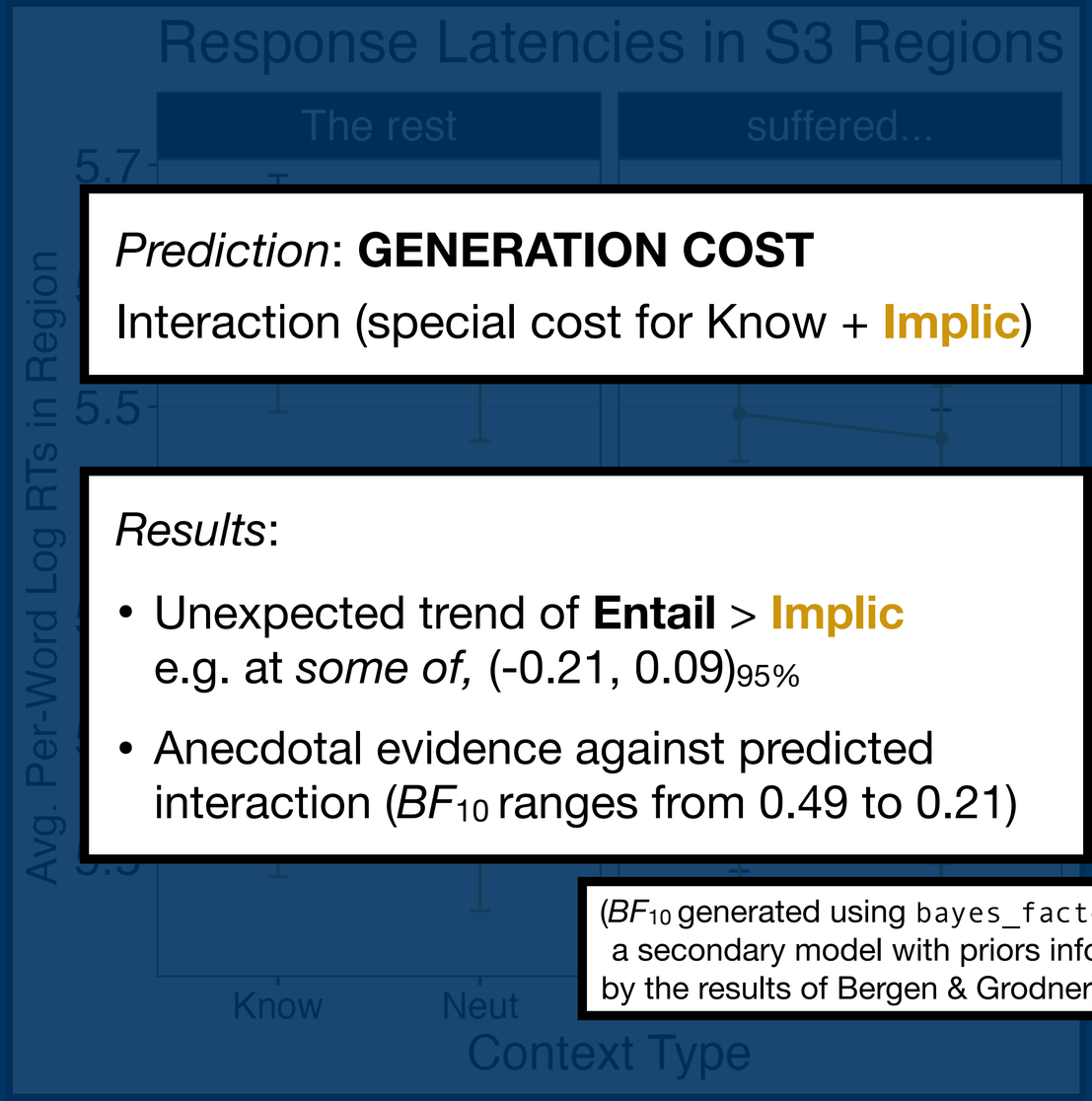
CANCELLATION

E1: Self-paced reading

n = 80 on Prolific
40 critical items + 70 fillers

linear m.-e. model fit in `brms`
to log RTs, uninf. priors





Prediction: **GENERATION COST**

Interaction (special cost for Know + **Implic**)

Results:

- Unexpected trend of **Entail** > **Implic**
  e.g. at *some of,* $(-0.21, 0.09)_{95\%}$
- Anecdotal evidence against predicted interaction (BF_{10} ranges from 0.49 to 0.21)

(BF_{10} generated using `bayes_factor` and a secondary model with priors informed by the results of Bergen & Grodner, 2012)

E1: Self-paced reading

n = 80 on Prolific
40 critical items + 70 fillers

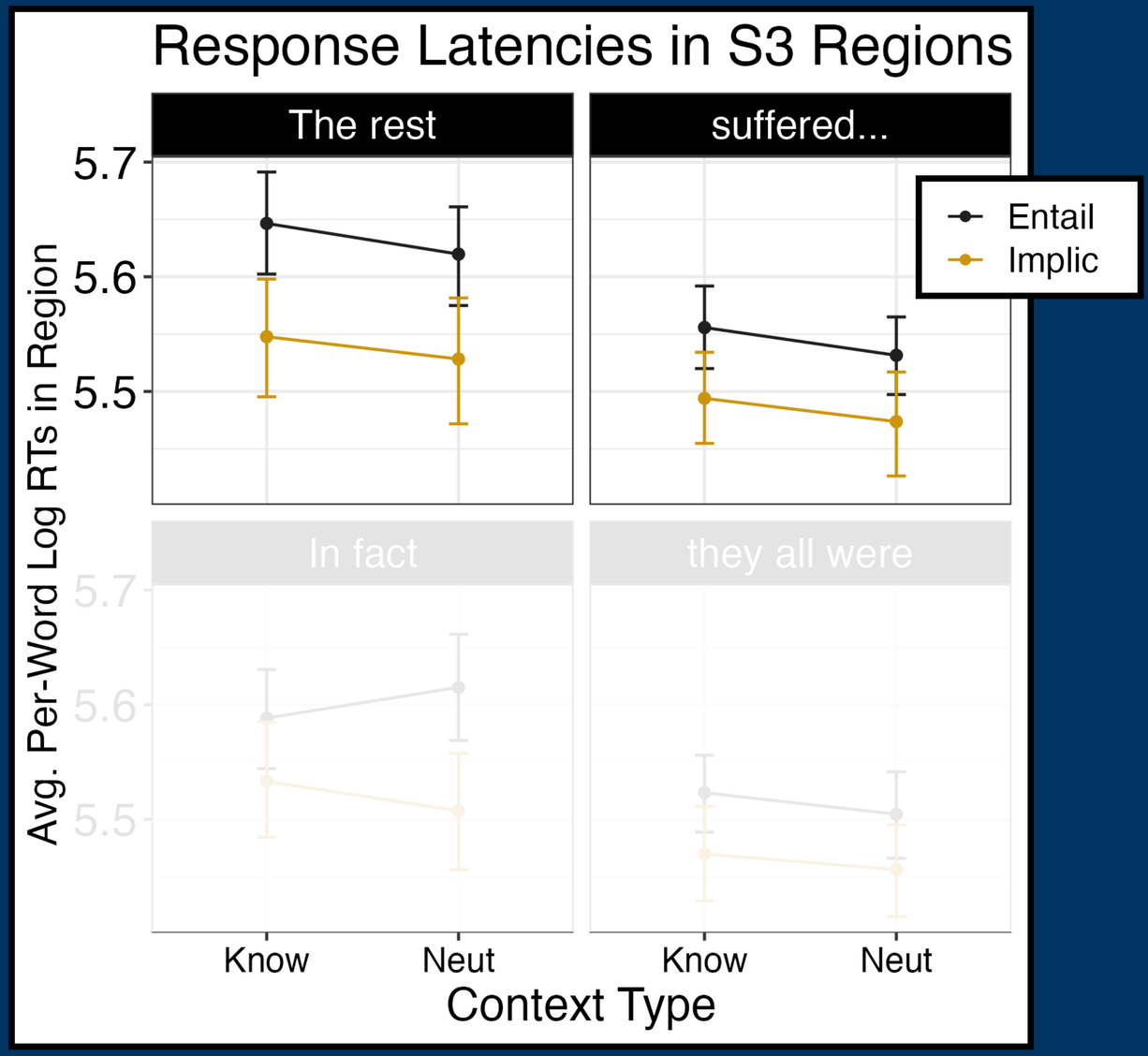
linear m.-e. model fit in `brms`
to log RTs, uninf. priors

Prediction: **EVIDENCE OF GENERATION**

Interaction (facilitation for Know + **Implic**)

Results:

- **Entail** > **Implic** again trending
- Moderate evidence against predicted interaction (BF_{10} ranges from 0.26 to 0.05)



E1: Self-paced reading

n = 80 on Prolific
40 critical items + 70 fillers

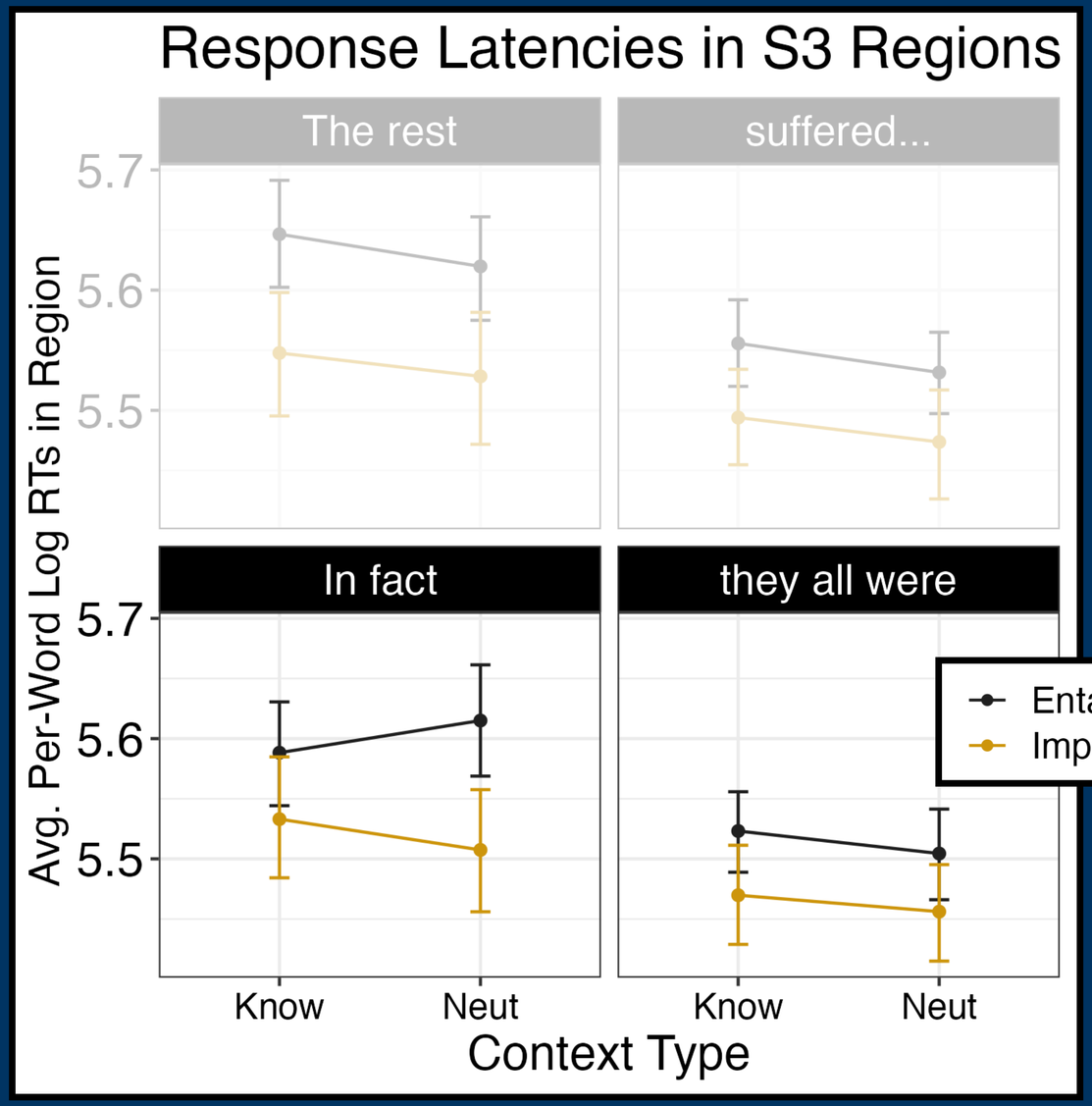
linear m.-e. model fit in `brms`
to log RTs, uninf. priors

Prediction: **COSTLY CANCELLATION**

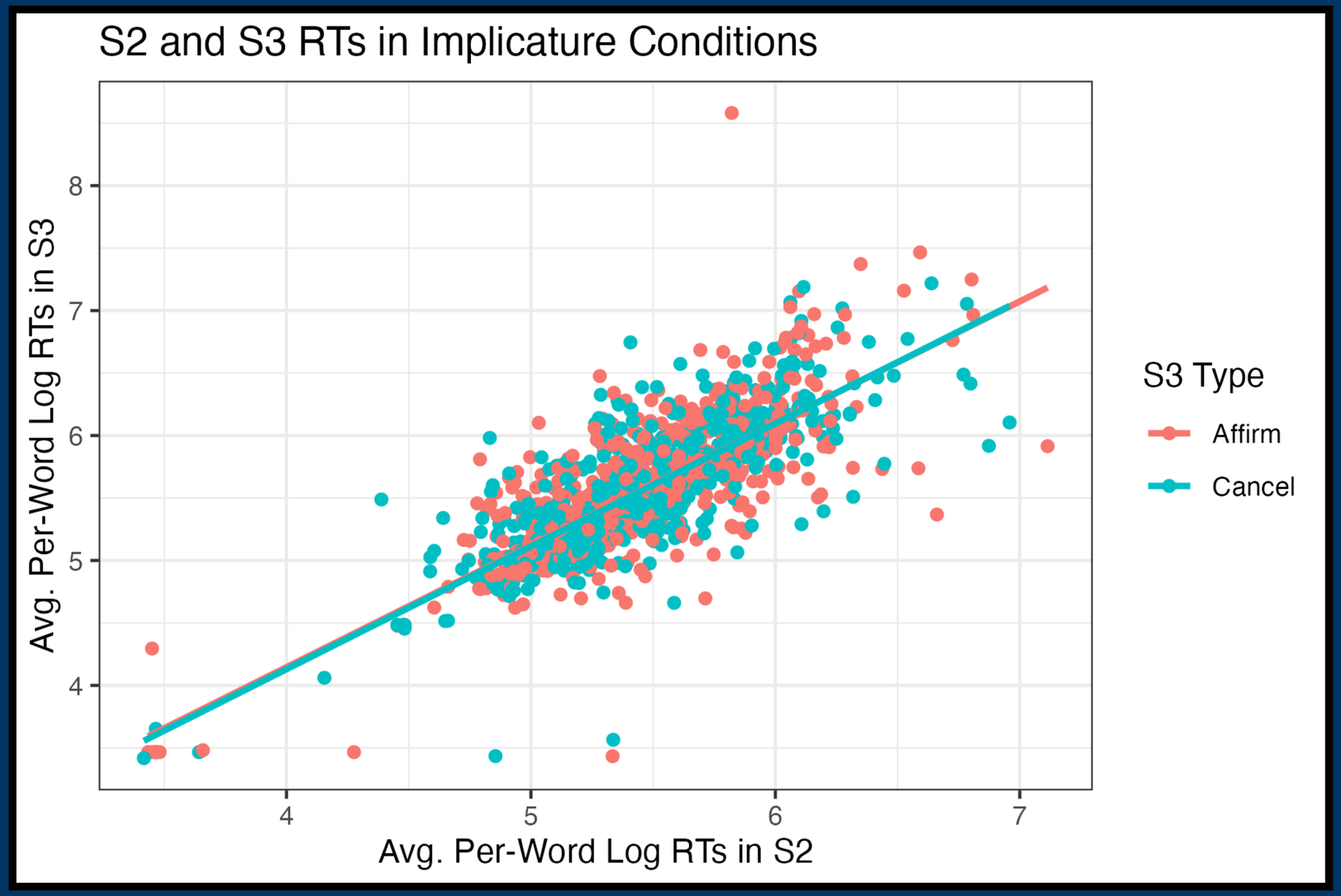
Interaction (special cost for Know + **Implic**)

Results:

- **Entail** > **Implic** credible, $(-0.13, 0.03)_{95\%}$
- Anecdotal evidence for predicted interaction at *in fact* ($BF_{10} = 1.25$)
  - $P(\delta_{\text{Context} \mid \text{Implic}} < 0) = 0.85$



E1: Within-trial correlations

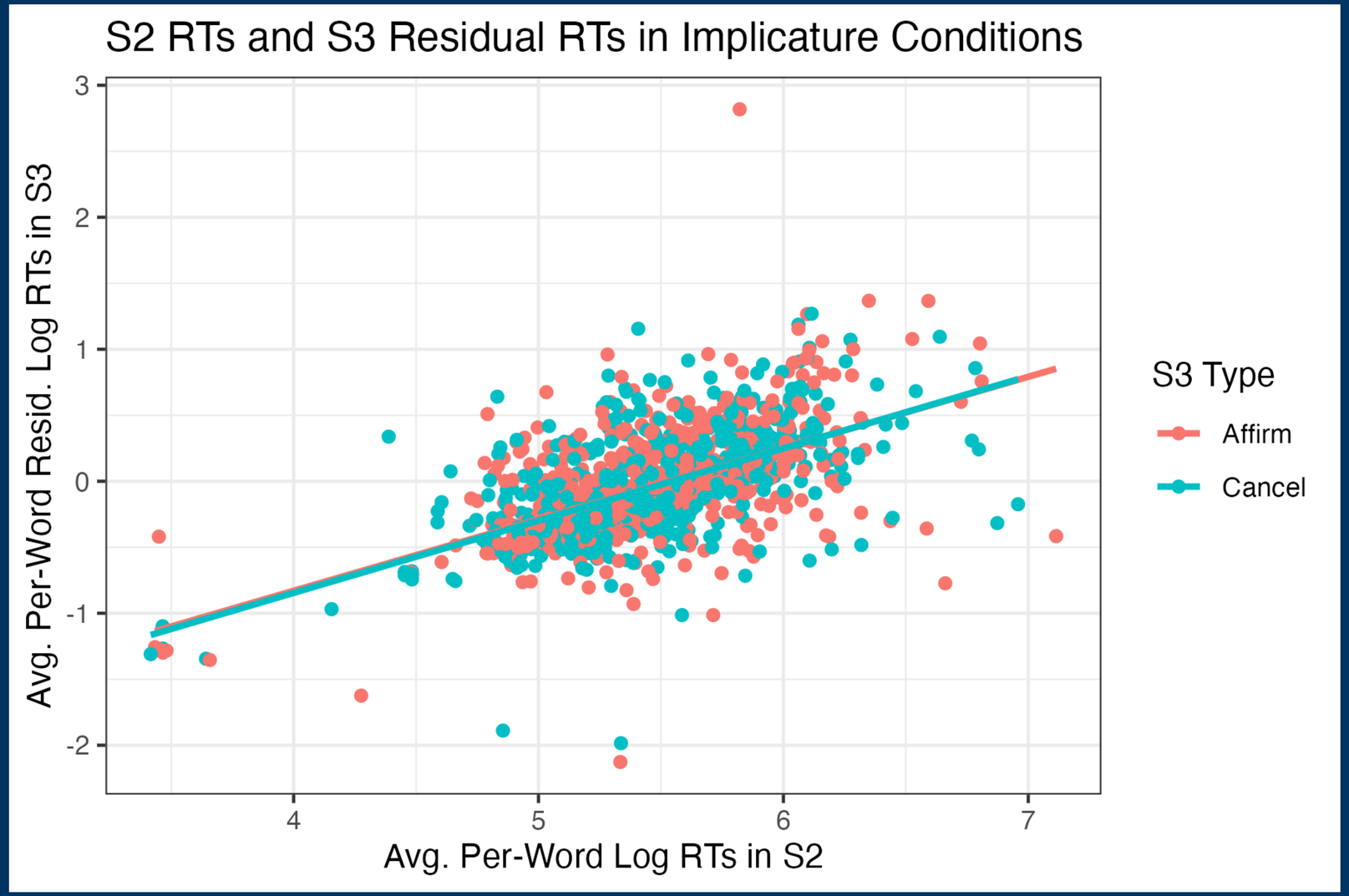


Obvious positive relationship (slower trials are slower consistently).

Step 1: Fit S3 ~ S2 model on a control condition (affirmed Entailments).

Step 2: Generate predictions for S3 in critical conditions and calculate residuals.

E1: Within-trial correlations



Predictions: More time in S2
→ less time in affirmative S3
→ more time in cancellation S3

Result: Slower S2s yield slower S3s regardless of condition.

S2: $(0.15, 0.25)_{95\%}$
S2 x S3Type: $(-0.02, 0.04)_{95\%}$

Fails to replicate Bergen & Grodner's key correlation.

Struggling with an implicature predicts continued difficulty?

E1: Discussion

GENERATION COST?

No concurrent evidence.* (*unless you subset to earlier trials)

✅

EVIDENCE OF GENERATION?

No concurrent evidence for pre-activation of the complement set.

❌

COSTLY CANCELLATION?

Some novel but weak evidence for cancellation costs. (*especially in earlier trials)

✅

Roadmap

1. Introduction
2. Existing work
3. Materials
4. E1: Self-paced reading
5. **E2: A-Maze reading**
6. Discussion and conclusions

The (A-)Maze task

The A-Maze task (Boyce et al. 2020):

localizes	brown	ten	may	introduce	pear	effects
potatoes	costs	but	hip	riverbeds	task	closest

Duff et al. 2020/21:

Eager lexical commitments in the Maze: ambiguities are resolved earlier than normal.

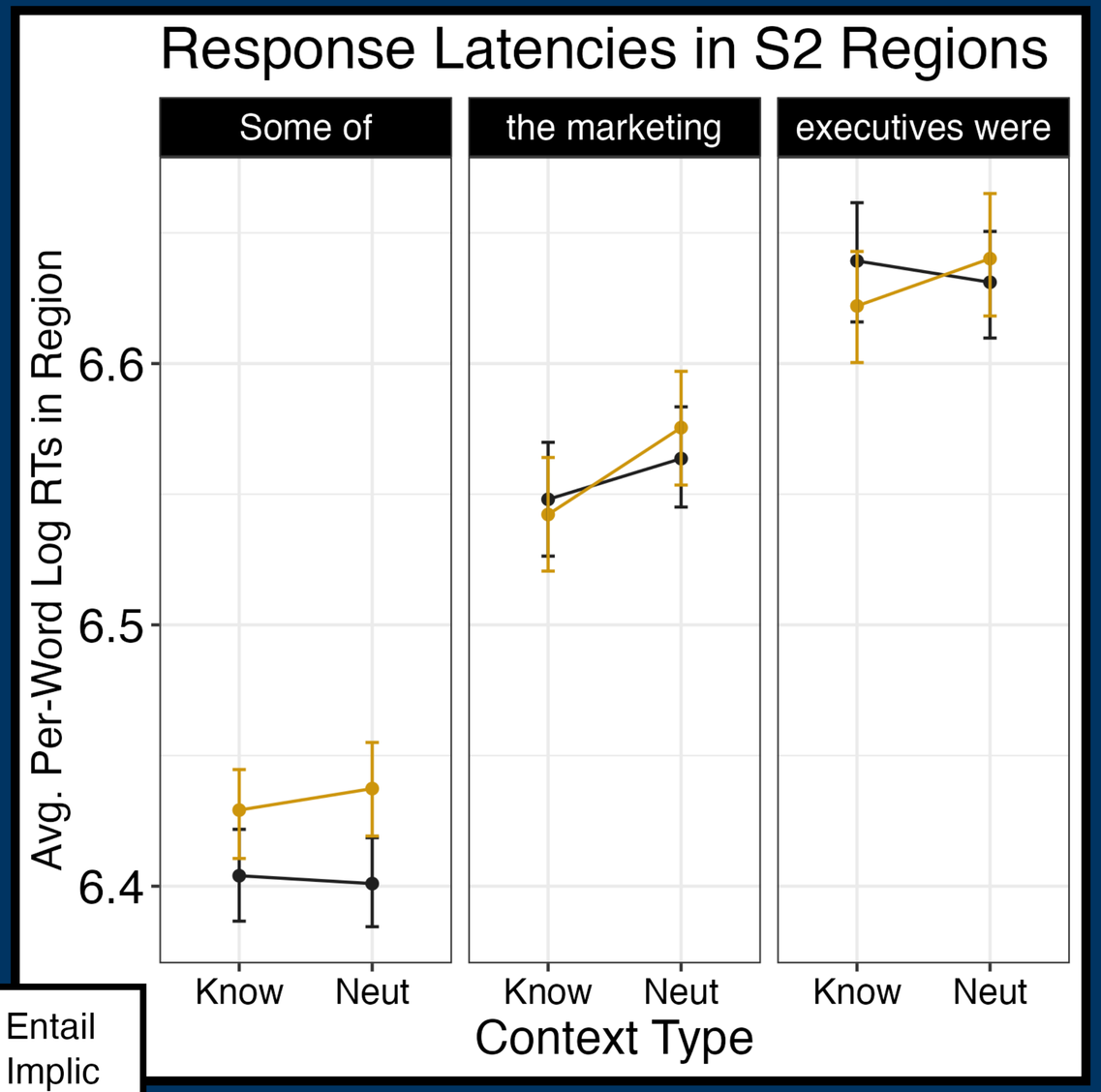
- Is the Maze appropriate for more complex pragmatic comprehension?
- Does the Maze also motivate earlier, firmer implicature generation?

Forster, Guerrera & Elliot (2009), Boyce, Futrell & Levy (2020), Duff, Brasoveanu & Rysling (Posters @ AMLaP 2020, CUNY 2021)

E2: A-Maze

n = 71 on Prolific
40 critical items + 70 fillers

linear m.-e. model fit in `brms`
to log RTs, uninf. priors



Prediction: **GENERATION COST**

Interaction (special cost for Know + **Implic**)

Results:

- Trending main effect of **Implic** > **Entail**
  e.g. at *some of,* $(-0.21, 0.09)_{95\%}$
- Moderate evidence against predicted interaction ($BF_{10} = 0.11$)
- But anecdotal evidence **for** general difficulty for **Implic** ($BF_{10} = 2.07$)

E2: A-Maze

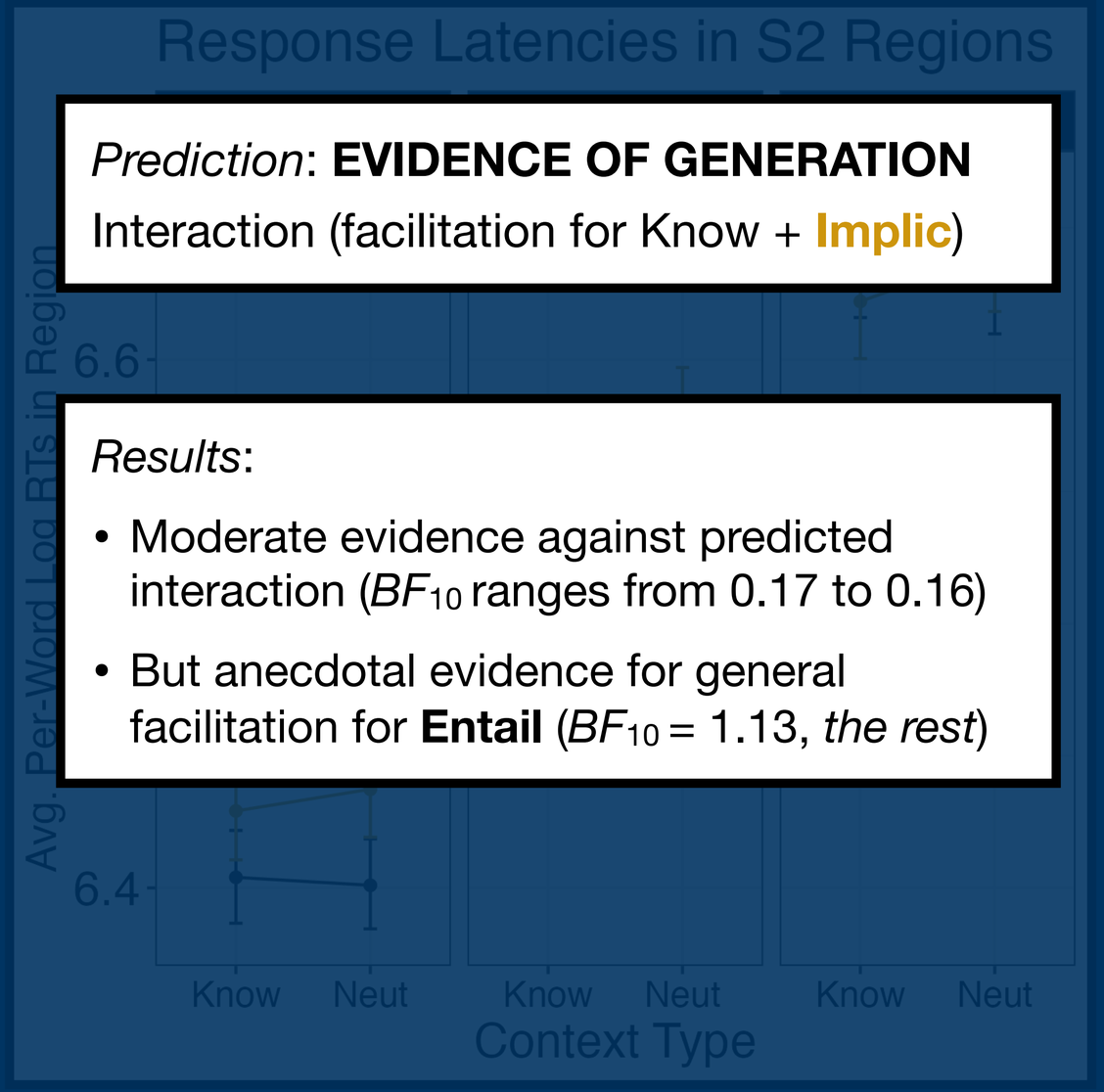
n = 71 on Prolific
40 critical items + 70 fillers

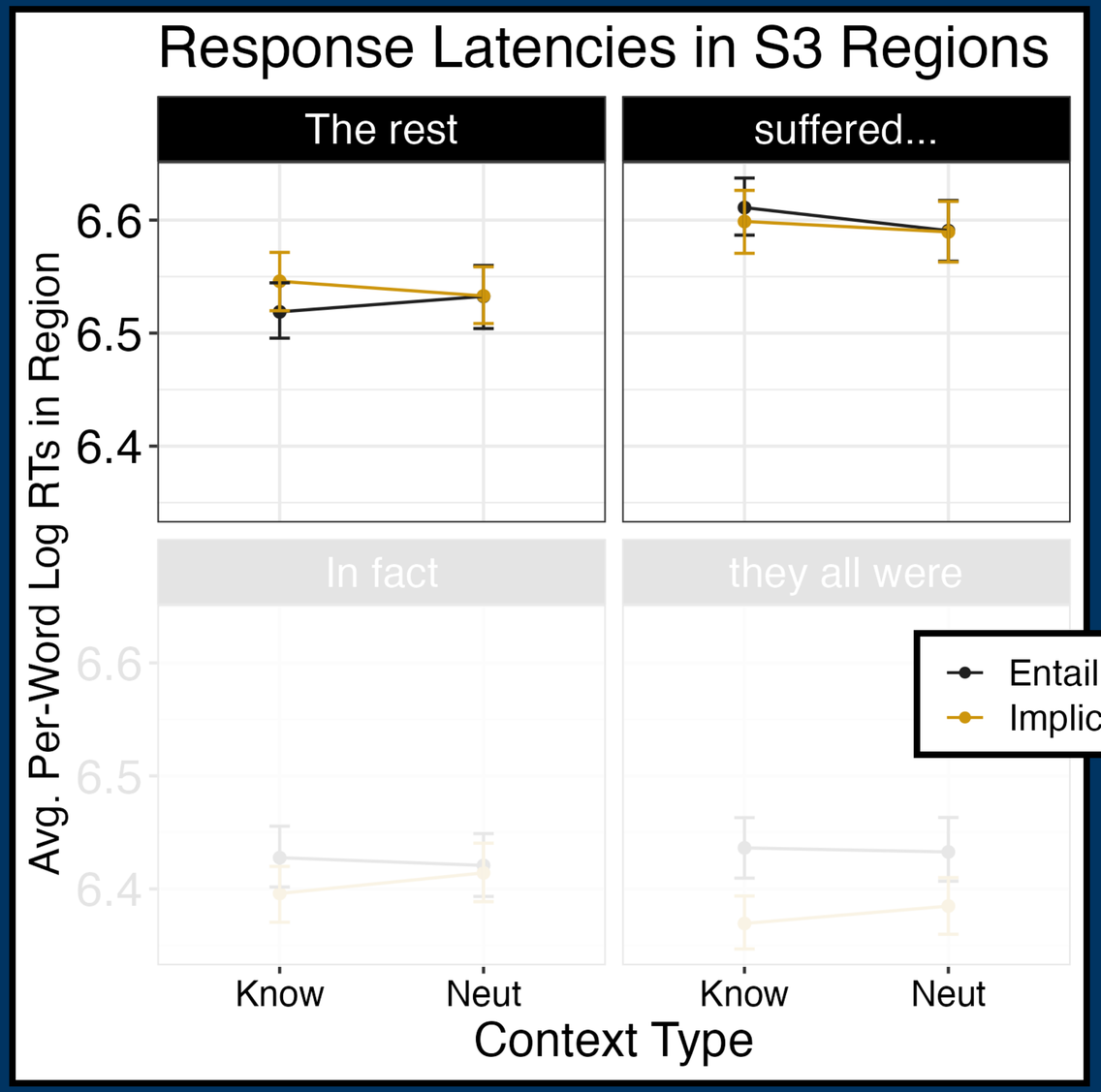
linear m.-e. model fit in `brms`
to log RTs, uninf. priors

Prediction: **EVIDENCE OF GENERATION**
Interaction (facilitation for Know + **Implic**)

Results:

- Moderate evidence against predicted interaction (BF_{10} ranges from 0.17 to 0.16)
- But anecdotal evidence for general facilitation for **Entail** (BF_{10} = 1.13, *the rest*)





E2: A-Maze

n = 71 on Prolific
40 critical items + 70 fillers

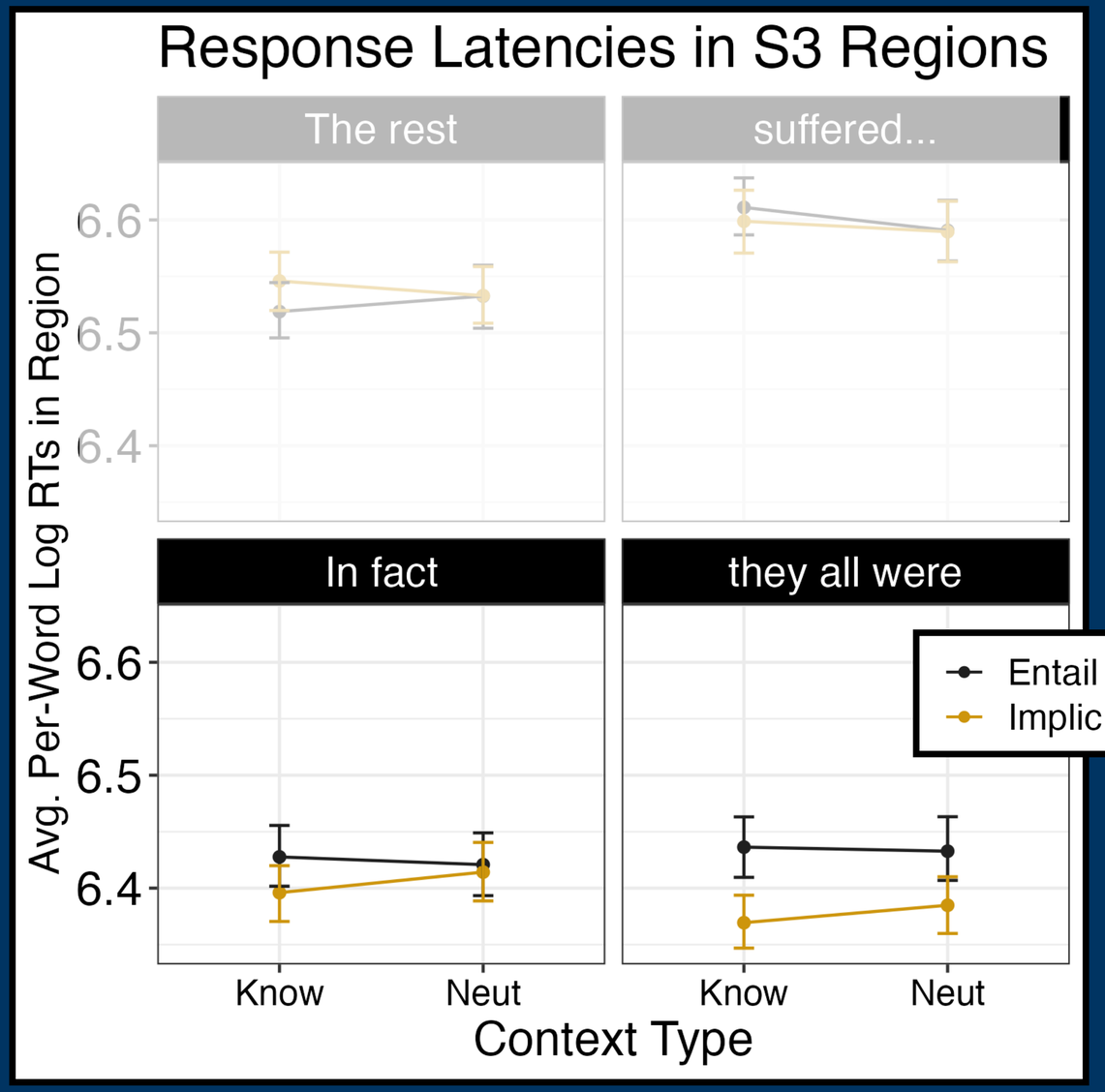
linear m.-e. model fit in `brms`
to log RTs, uninf. priors

Prediction: **COSTLY CANCELLATION**

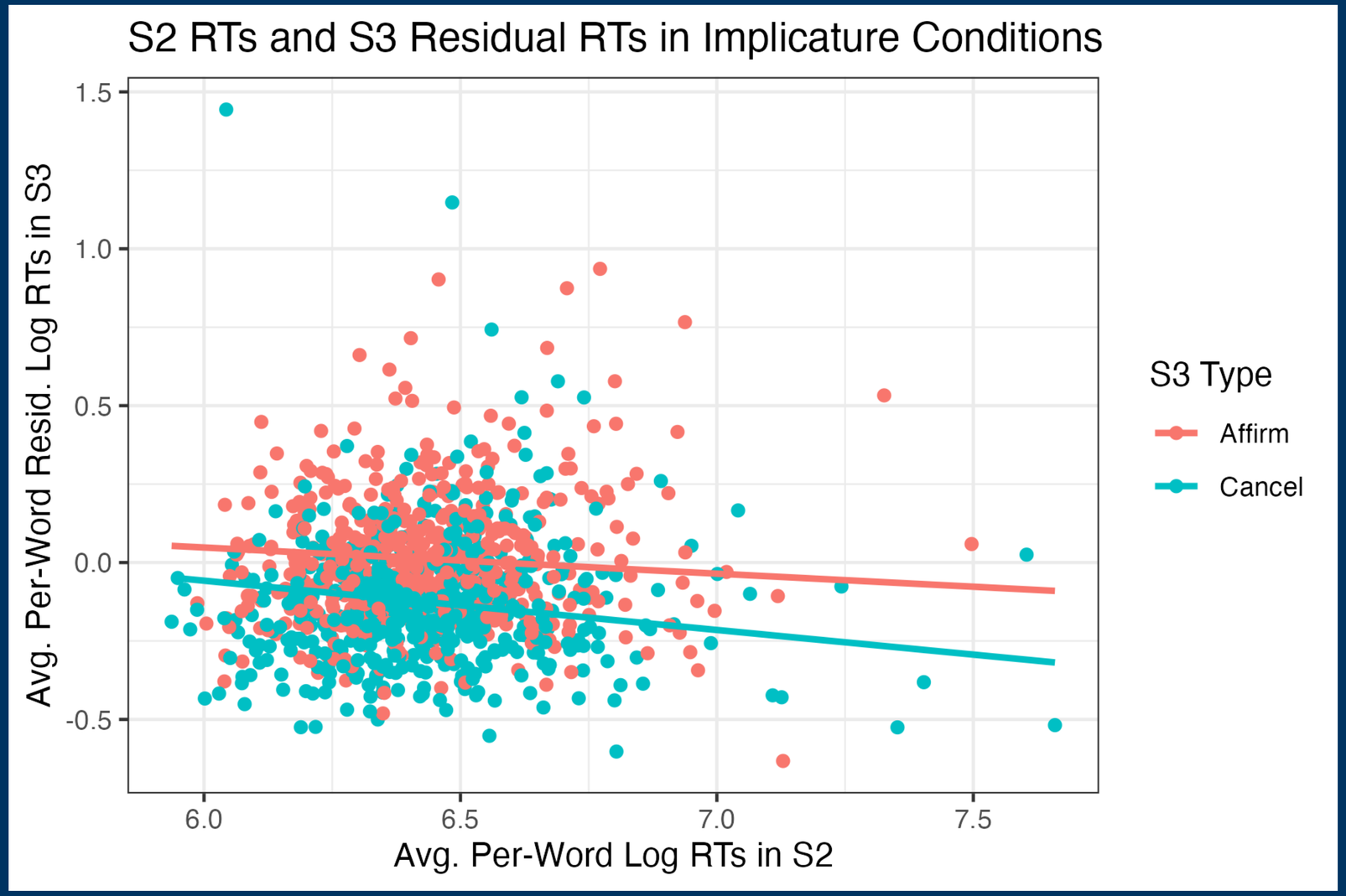
Interaction (special cost for Know + **Implic**)

Results:

- Very strong evidence against predicted interaction (BF_{10} ranges from 0.003 to 0.03)
- But anecdotal evidence for general difficulty with **Entail** (BF_{10} = 2.14, *they all…*)



E2: Within-trial correlations



Predictions: More time in S2
→ less time in affirmative S3
→ more time in cancellation S3

Result: Slower S2s yield faster S3s, even (especially?) for cancellation.

S2: $(-0.18, -0.06)_{95\%}$
S2 X S3Type: $(-0.09, 0.02)_{95\%}$

Opposite results from SPR.

Lingering at an implicature trigger predicts general facility, even with later cancellation.

E2: Discussion

GENERATION COST?

General cost for implicatures vs. *only*, but no context effects. ✅

EVIDENCE OF GENERATION?

Only may facilitate the complement set more than implicatures, no context effects. ✅

COSTLY CANCELLATION?

Contradictions of *only* yield larger slowdowns than implic. cancellation, no context effects. ✅

CAREFUL IMPLICATURES CANCEL EASILY?

Slowdowns at *some* are associated with ease, not difficulty, of cancellation. ❓

Roadmap

1. Introduction
2. Existing work
3. Materials
4. E1: Self-paced reading
5. E2: A-Maze reading
6. **Discussion and conclusions**

Task sensitivity in pragmatic processing

- Long experiments with many repetitions generate high power but can dampen desired effects.
- SPR studies are especially noisy and likely to cause fatigue or adaptation.
- The Maze can offer more precise measurement, but may encourage context-free processing.
    - Use with caution when investigating pragmatic sentence processing.

Insights into implicature generation

- Considering an implicature comes with a cost…
  - Compared to implicature-inconsistent contexts (in early SPR trials)
  - Compared to entailments (in Maze)
- The size of this cost on any given trial may have different explanations:
  - In SPR, costs seem to diagnose pragmatic difficulty that continues.
  - In the Maze, costs seem to diagnose pragmatic care and attention.

Insights into implicature cancellation

- Mixed evidence for costly cancellation.
  - Some weak evidence for predicted costs in SPR.
  - In the Maze data, no ability to prove cancellation costs.
    - If present, they are smaller than contradiction costs.
    - Unexpected negative correlation: cancellation may be facilitated by careful construction of the implicature.

Thanks!

Special thanks to RAs Sebastian Bissiri and Kasey La, plus Alexander Göbel and Matt Wagers for helpful discussion.

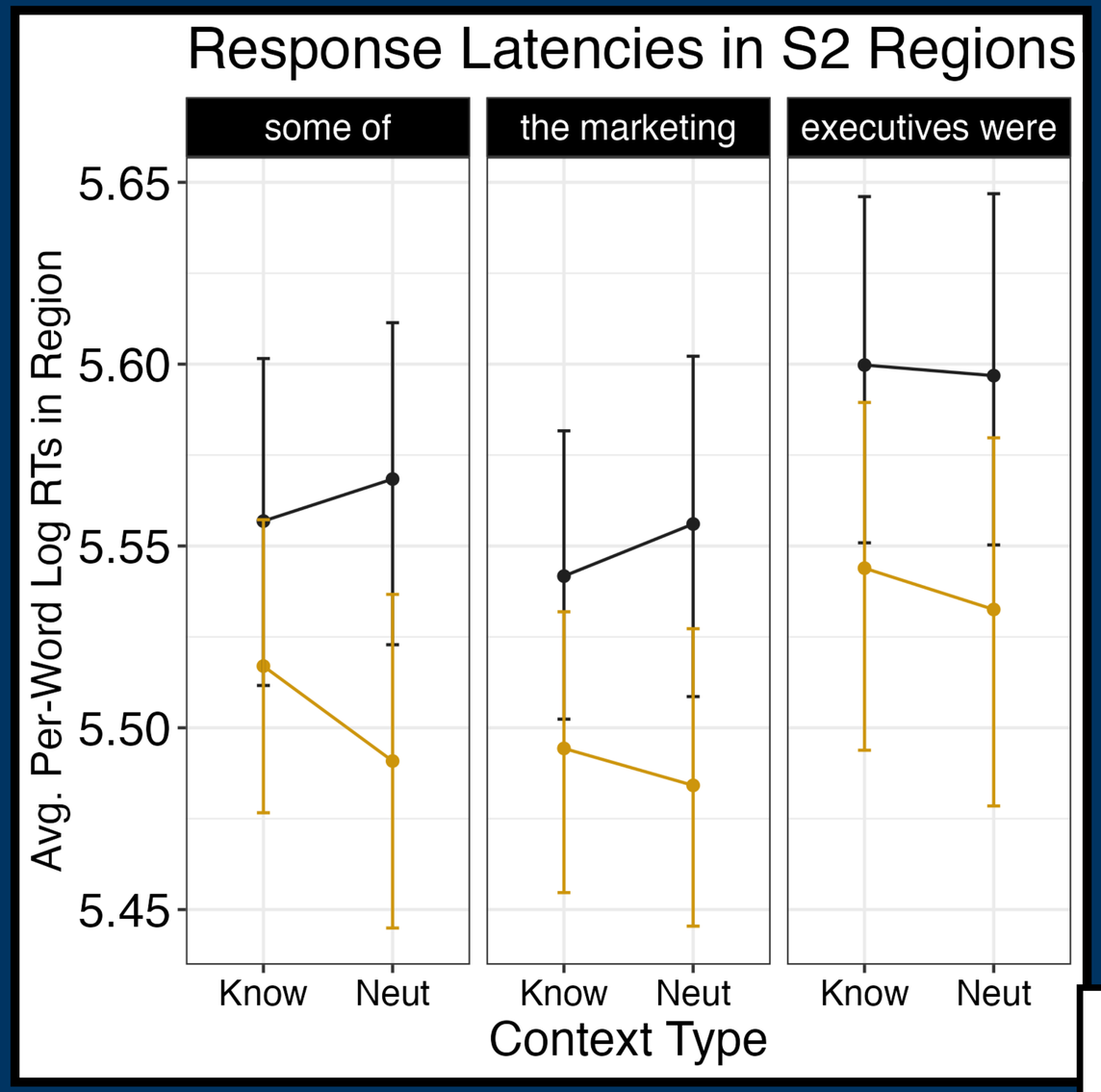
Happy to share more information and answer questions about:

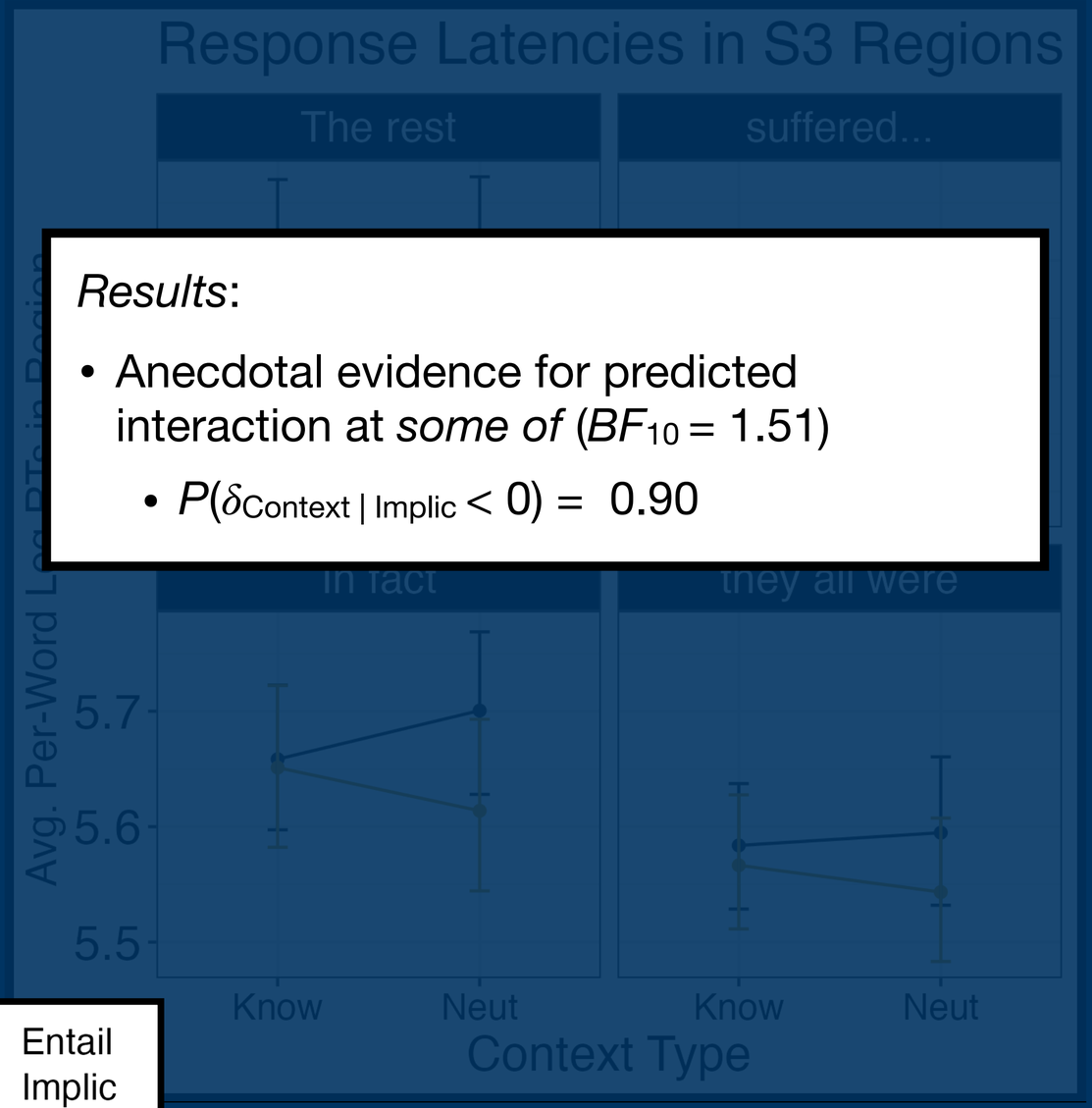
- Items, procedure and modeling details for these experiments
- Related SPR/Maze experiments on lexical and distributive ambiguities and online generation of causal discourse inferences



Appendix 1: Early time window for E1

E1: Results from first four exposures





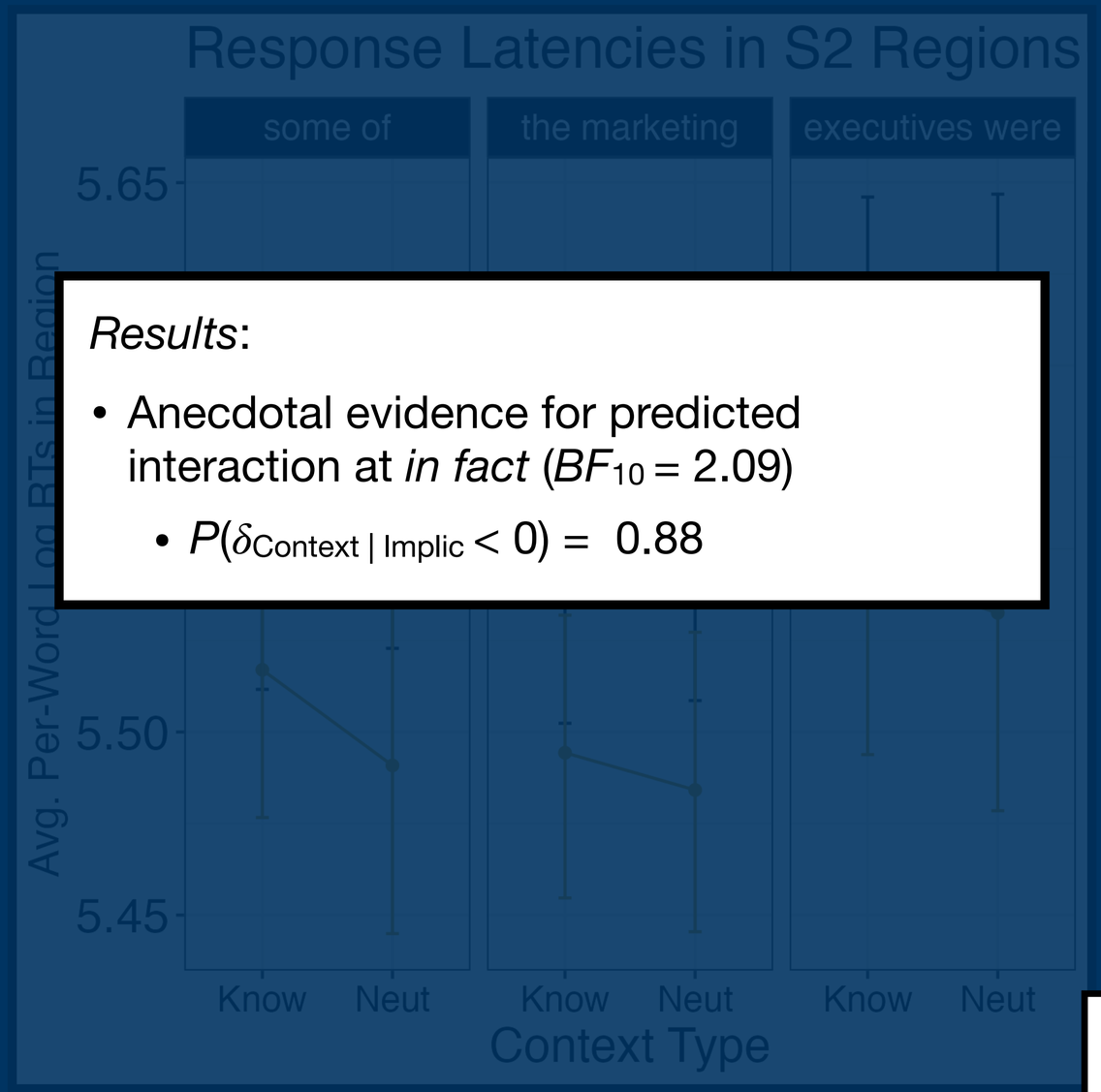
Results:

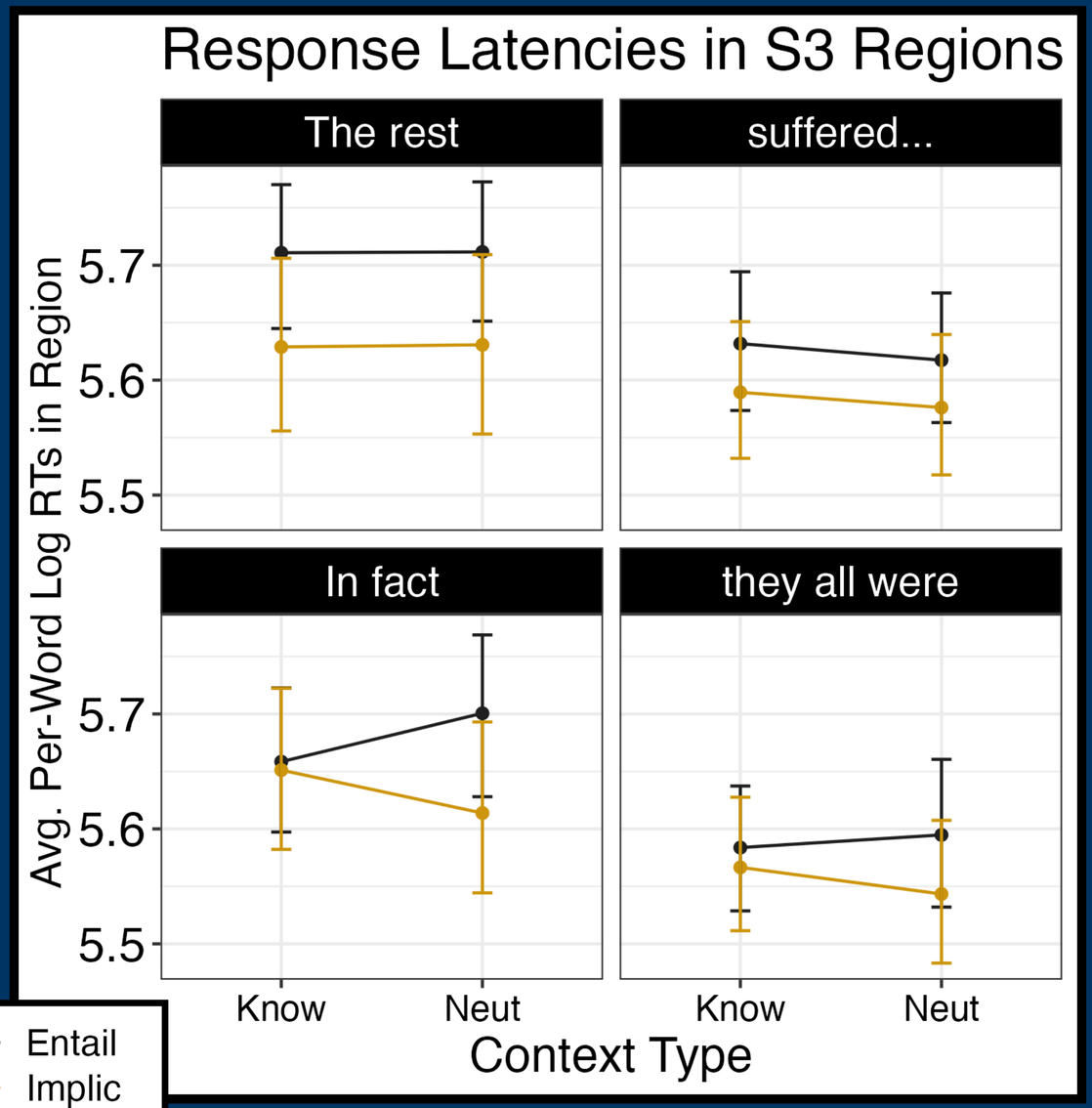
- Anecdotal evidence for predicted interaction at *some of* ($BF_{10} = 1.51$)
  - $P(\delta_{\text{Context} \mid \text{Implic}} < 0) = 0.90$

E1: Results from first four exposures

Results:

- Anecdotal evidence for predicted interaction at *in fact* ($BF_{10} = 2.09$)
  - $P(\delta_{\text{Context} \mid \text{Implic}} < 0) = 0.88$





Appendix 2: Other sample items

(1-KNOW) Albert carefully inspected the new shipment of jewelry.

(1-NEUT) Albert helped unload the new shipment of jewelry.

(2) He noticed that (only) some of the gold watches were fakes.

(3-AFF) The rest were real, but the company is still planning to sue.

(3-CAN) In fact, they all were, so the company is planning to sue.

(1-KNOW) At his client's request, Wilbur meticulously compiled the investment report.

(1-NEUT) At his client's request, Wilbur skimmed the investment report.

(2) He noticed that (only) some of the real estate investments lost money.

(3-AFF) The others were successful, in spite of the recent economic downturn.

(3-CAN) In fact, they all did, because of the recent economic downturn.

(1-KNOW) Jonathan designed the packaging for the company's new environmentally-friendly soaps.

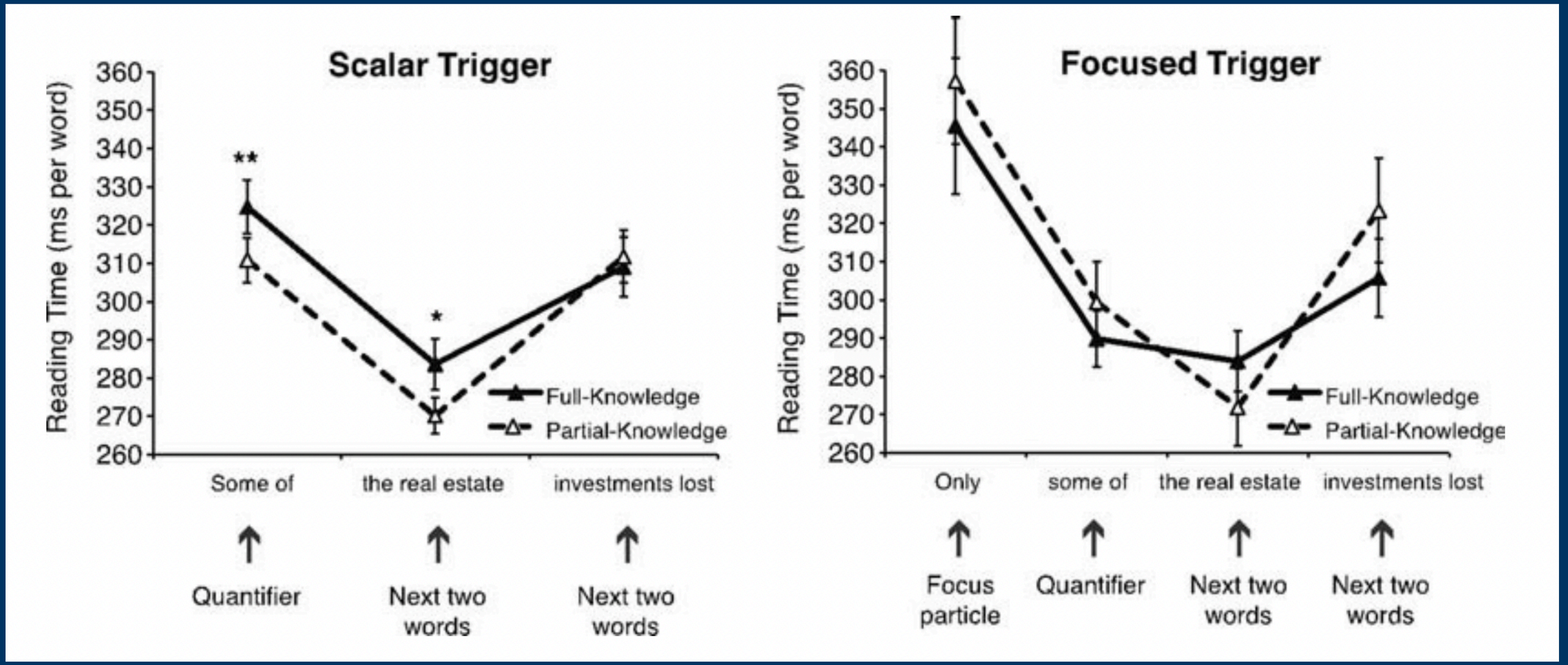
(1-NEUT) Jonathan read a profile of the company's new environmentally-friendly soaps.

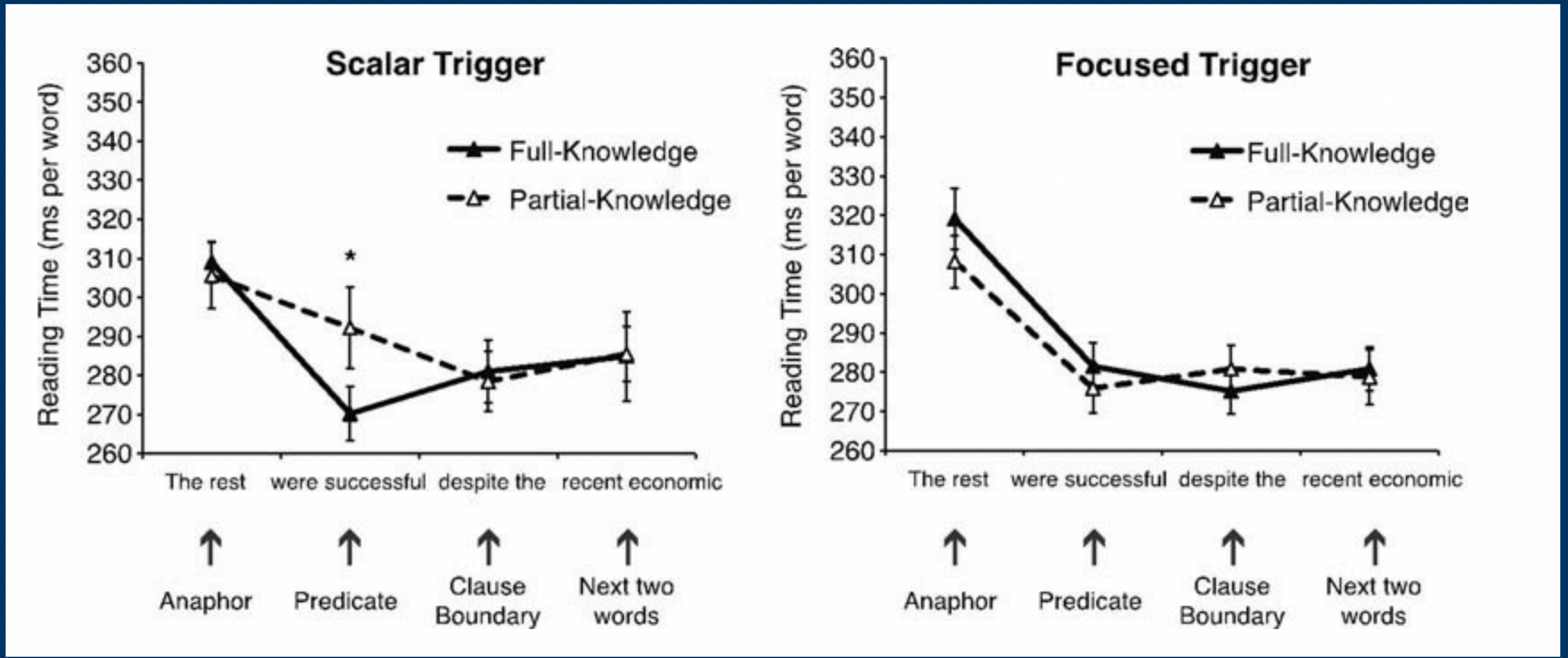
(2) He was glad that (only) some of the bottles were recyclable.

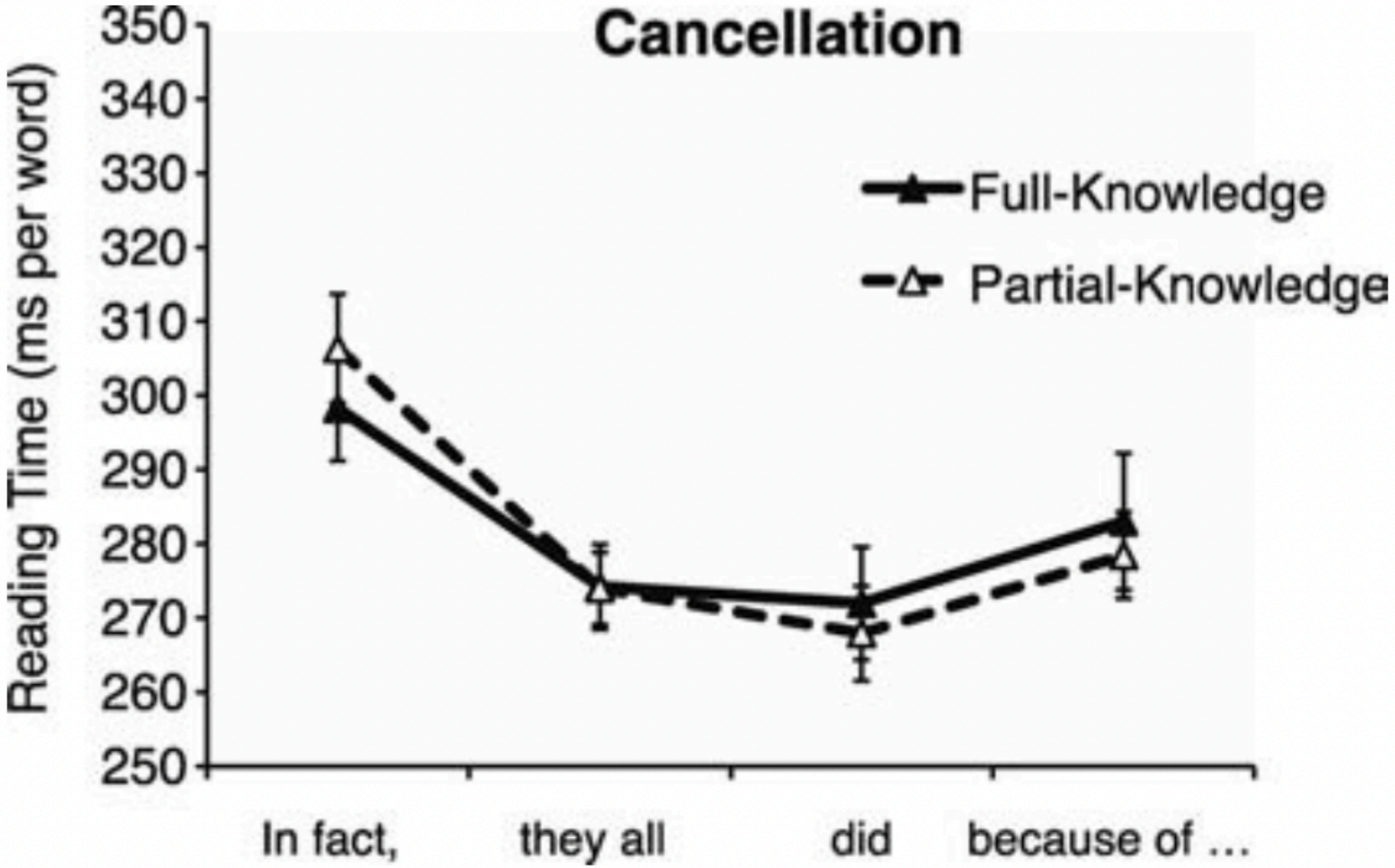
(3-AFF) The rest were compostable, but consumers didn't seem to care.

(3-CAN) In fact, they all were, but consumers didn't seem to care.

Appendix 3: Bergen & Grodner (2012)


Scalar Trigger
Reading Time (ms per word)
360
350
340
330
320
310
300
290
280
270
260
**
*
Full-Knowledge
Partial-Knowledge
Some of
the real estate
investments lost
Quantifier
Next two words
Next two words
Focused Trigger
Reading Time (ms per word)
360
350
340
330
320
310
300
290
280
270
260
Full-Knowledge
Partial-Knowledge
Only
some of
the real estate
investments lost
Focus particle
Quantifier
Next two words
Next two words


Scalar Trigger
Reading Time (ms per word)
360
350
340
330
320
310
300
290
280
270
260
Full-Knowledge
Partial-Knowledge
*
The rest
were successful
despite the
recent economic
Anaphor
Predicate
Clause Boundary
Next two words
Focused Trigger
Reading Time (ms per word)
360
350
340
330
320
310
300
290
280
270
260
Full-Knowledge
Partial-Knowledge
The rest
were successful
despite the
recent economic
Anaphor
Predicate
Clause Boundary
Next two words


Cancellation
Reading Time (ms per word)
350
340
330
320
310
300
290
280
270
260
250
Full-Knowledge
Partial-Knowledge
In fact,
they all
did
because of …

Appendix 4: Polysemy in the Maze

(Duff, Brasoveanu & Rysling @ CUNY 2021)

Underspecification

Unfortunately, the newspaper was destroyed…

after it lost its advertising profits.

(M1) (M2)

Reportedly, the jam displeased Tom…

after it doubled his morning commute.

(M1) (M2)

Claim: Full commitment to a particular meaning of a polyseme is delayed.

Why?

Utility: Because it's efficient when possible: prevents costly reanalysis.

Necessity: Because the processor cannot resolve polysemes without context.

What happens when underspecification wouldn't be useful?

Eye-tracking-while-reading: Frazier & Rayner (1990), Frisson & Pickering (1999), Pickering & Frisson (2001), Foraker & Murphy (2012), Brocher, Koenig, Mauner, & Foraker (2018)

Enter the Maze

(Obviously, the referee had…)

WELFARE DROPPED

(~40%)

The A-Maze (Boyce et al. 2020) encourages eager interpretation.

- Representing semantic context necessary to pick the correct target

↳ **Underspecification is no longer useful.**

If underspecification is utility-based / necessary then we won't / will see it in the Maze.

Reanalysis costs for homonymy and polysemy.

More reanalysis costs for homonymy.

Forster, Guerrera, & Elliot (2009), Boyce, Futrell, & Levy (2020)

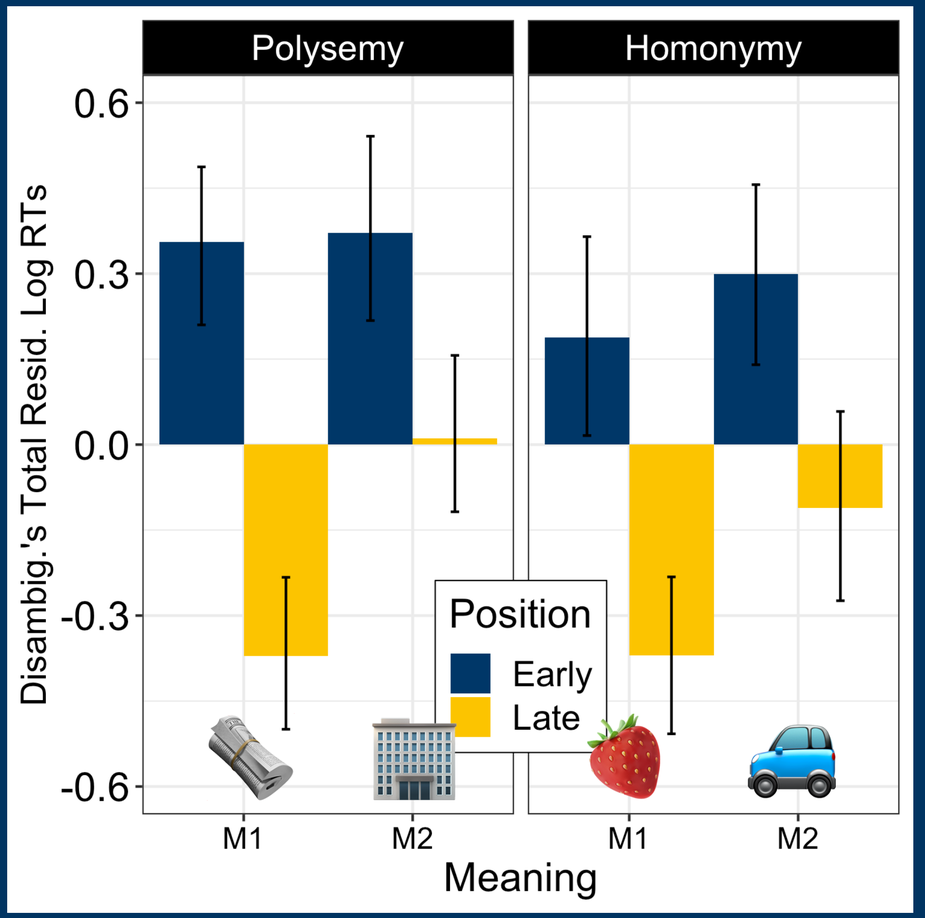
64 Latin-squared items (32 POL, 32 HOM); 128 fillers; *n* = 24 UCSC + 24 Prolific

E1: No underspecification in the Maze

M1, EARLY	Unfortunately, after it was soaked with rain the **newspaper** was destroyed.
M1, LATE	Unfortunately, the **newspaper** was destroyed after it was soaked with rain.
M2, EARLY	Unfortunately, after it lost its advertising profits the **newspaper** was destroyed.
M2, LATE	Unfortunately, the **newspaper** was destroyed after it lost its advertising profits.

- POSITION: **LATE** read faster, presumably due to cataphora in **EARLY**
- POSITION X MEANING: Reduced for M2, apparent reanalysis costs
- No POS X POL/HOM (X M): no difference in reanalysis for POL V. HOM
- Replicated in error rates (not shown): No POL/HOM difference

↳ **No evidence for necessary underspecification in the Maze.**



Log RTs residualized over position and length, summed, analyzed via LMER fit in STAN, fixed effects treatment-coded. Effects reported if 95% credible interval excludes 0.

64 Latin-squared items (32 POL, 32 HOM); 128 fillers; *n* = 24 UCSC + 24 Prolific

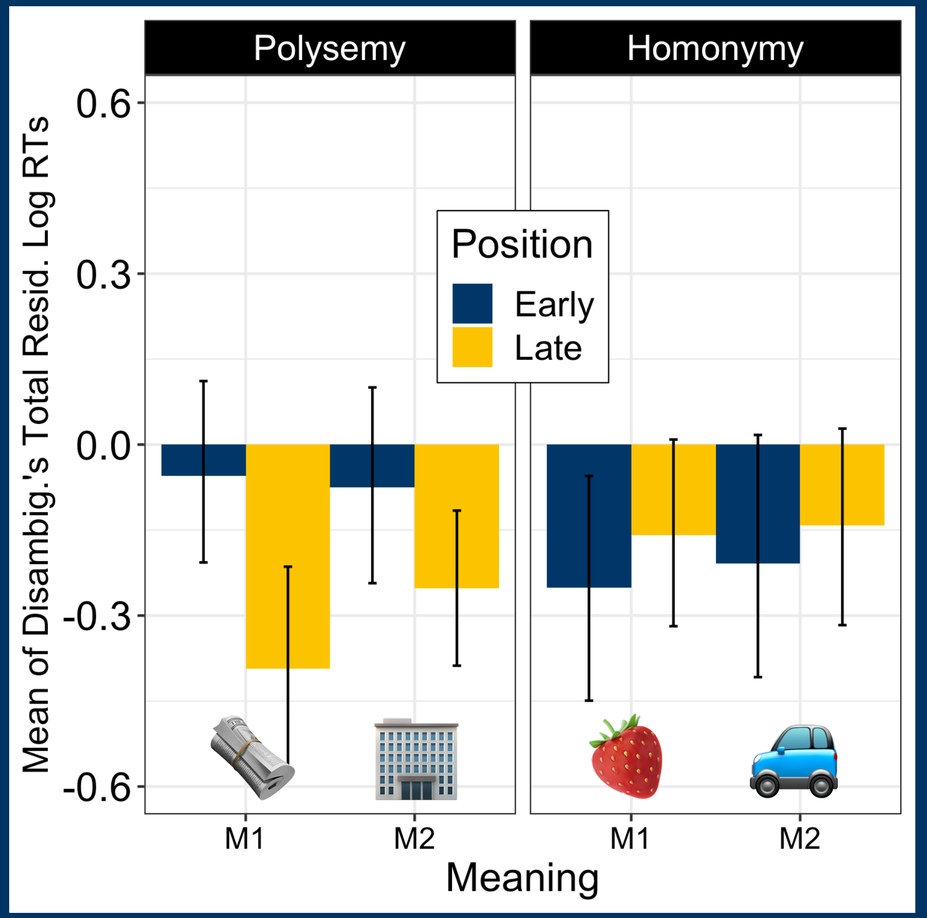
E2: Underspecification in SPR

M1, EARLY	Unfortunately, after it was soaked with rain the **newspaper** was destroyed.
M1, LATE	Unfortunately, the **newspaper** was destroyed after it was soaked with rain.
M2, EARLY	Unfortunately, after it lost its advertising profits the **newspaper** was destroyed.
M2, LATE	Unfortunately, the **newspaper** was destroyed after it lost its advertising profits.

SPR replication to ensure the Maze results are due to the task.

- POSITION: **LATE** read faster, again due to cataphora in **EARLY**
- POSITION X POL/HOM: Crossover for HOM, **extra reanalysis costs**

↳ **E1 results can be attributed to a Maze-specific task effect.**



Log RTs residualized over position and length, summed, analyzed via LMER fit in STAN, fixed effects treatment-coded. Effects reported if 95% credible interval excludes 0.

Upshots

Underspecification effects in polysemy are mediated by task demands.

- ↳ **Underspecification is optional and apparently strategic.**
- ↳ **Open questions remain: what makes it possible?**

The Maze task modulates strategies of incremental interpretation.

- ↳ **Shouldn’t be used as a 1:1 replacement for eyetracking or SPR.**
- ↳ **BUT: a powerful tool for clarifying the source of behavior.***

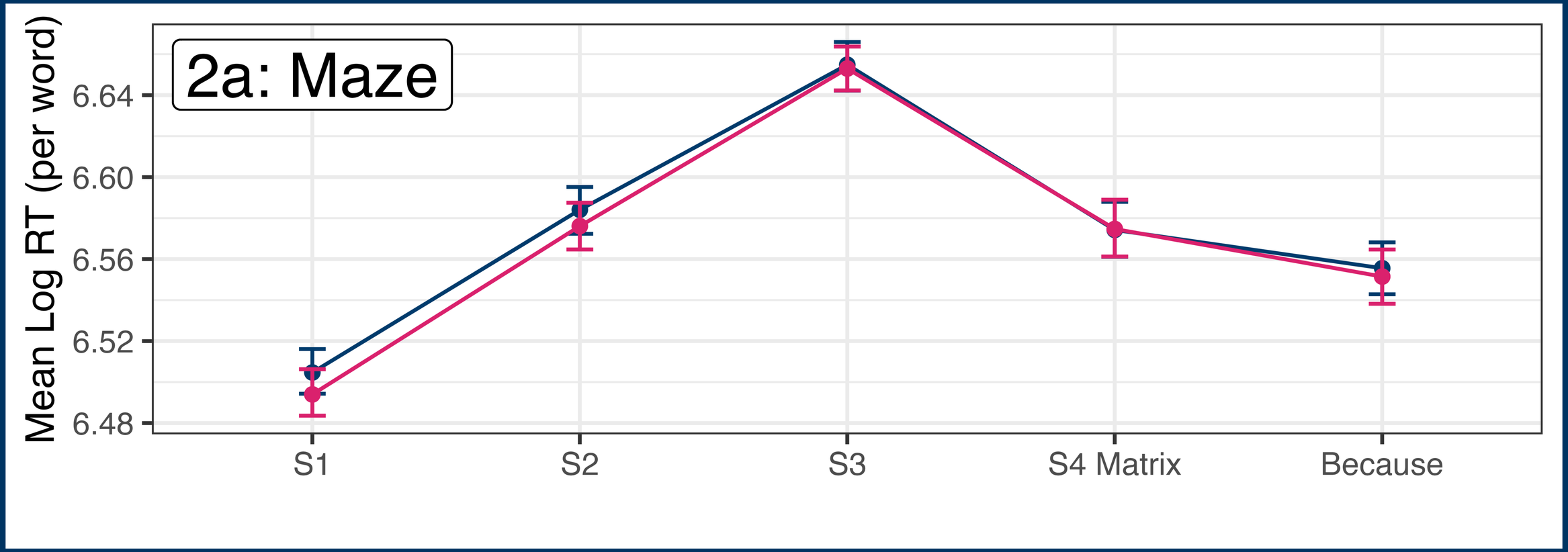
* e.g. Sloggett, Van Handel, Sasaki, Duff, Rich, Orth, Anand, & Rysling (2020 CUNY Poster)

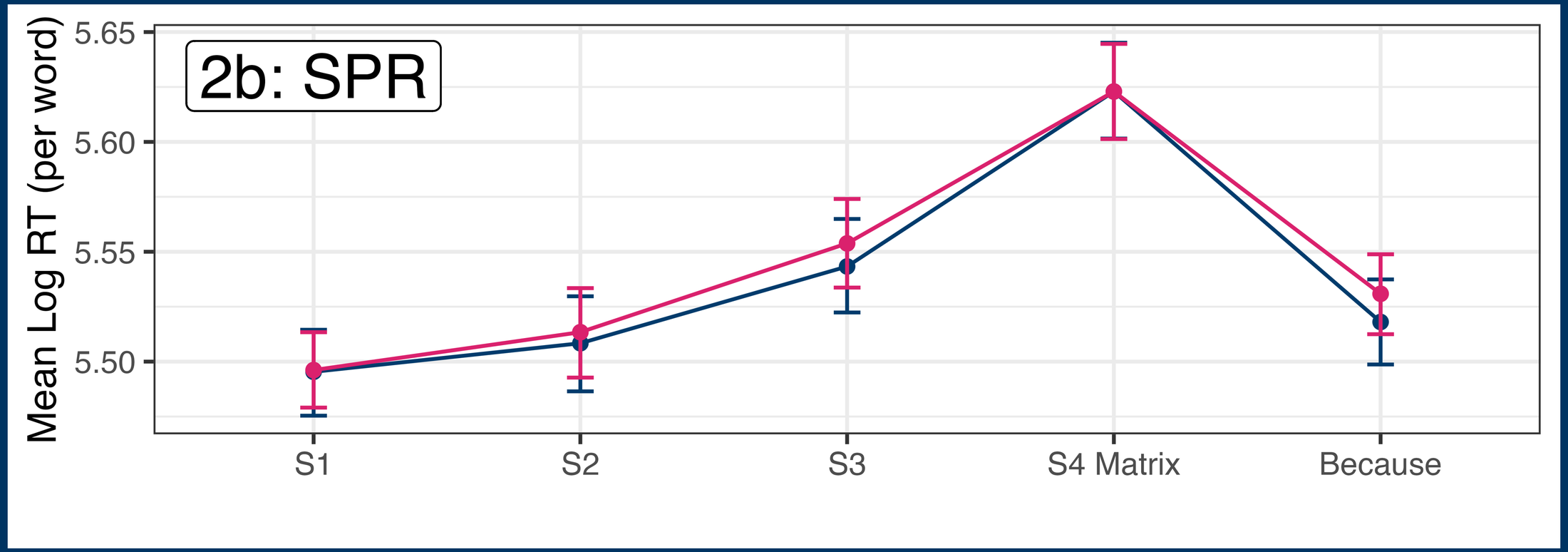
Appendix 5: Causal inferences in the Maze

(Duff, Anand & Rysling @ AMLaP 2023)

Sally lives in a small city, where recently there was a citywide election for a new mayor with several candidates, and she had to decide among them on her mail-in ballot.

Knowledgeable	She spent some time reading everything she could about the candidates before mailing in her ballot.
Ignorant	She didn't have any time to read anything about the candidates before mailing in her ballot.
S1	In the end, she voted for Pat Mirabella.
S2	He has the most progressive platform in the race.
S3	He's from a very socio-economically diverse area...
S4	She voted for him because his name was first on the ballot.


2a: Maze
Mean Log RT (per word)
6.64
6.60
6.56
6.52
6.48
S1
S2
S3
S4 Matrix
Because


2b: SPR
Mean Log RT (per word)
5.65
5.60
5.55
5.50
S1
S2
S3
S4 Matrix
Because