

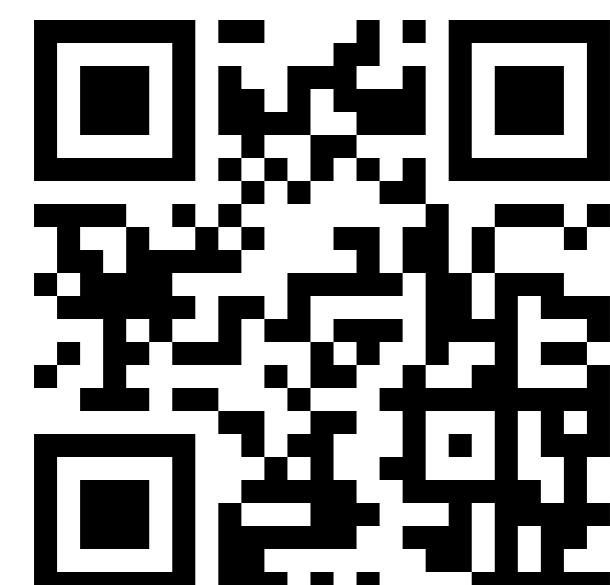
Introduction

- Benchmark data (e.g., [3, 8]) are an important tool for developing theories and evaluating model predictions.

Our Work (in Progress)

- We collect self-paced reading benchmark data for a battery of postulated effects in **German** (10 phenomena).
- In parallel, we also collect eye-tracking data on the same materials.
- So far (December 2025):
 - 950 Prolific participants have been tested with SPR. 63 were excluded due to low accuracy on comprehension questions.
 - 195 in-lab participants have been tested with eye tracking. 1 was excluded due to low accuracy on comprehension questions.
- We show the results so far, next to predictions based on qualitative theories, GPT-2 surprisal [1, 4, 11, 16], and lossy-context GPT-2 surprisal [2, 3, 6, 16].
- Model comparisons using Pareto-smoothed importance sampling [19] assess out-of-sample predictiveness.

Pre-Registration Protocol (SPR)



osf.io/wprag9

- Predictor performance ranking (SPR): **1. Lossy-context surprisal, 2. Surprisal, 3. Theory** (but no reliable differences in eye tracking)

Predictions, Results, and Model Comparison

Experimental Designs

GPSD (2x2): Garden Paths From Subject-vs.-Direct-Object Ambiguity
Ambiguous/Unambiguous \times S-O/O-S — closely replicating [13]

GPSI (2x2): Garden Paths From Subject-vs.-Indirect-Object Ambiguity
Ambiguous/Unambiguous \times Active/Passive — loosely replicating [14]

GPCA (2x2): Garden Paths From Coordination Ambiguity
NP-/VP-Coordination \times AP-/PP-Modifier — closely replicating [10]

GPMI (2x2): Garden Paths From Modifier-vs.-Indirect-Object Ambiguity
Modifier/No-Modifier \times Ambiguous/Unambiguous — closely replicating [9]

AGAT (2x2): Agreement Attraction in Grammatical Sentences
Singular-/Plural-Controller \times Match/Mismatch — closely replicating [5]

LOCO (2x2): Local Coherence
Coherent/Incoherent \times Intervener/No-Intervener — closely replicating [15]

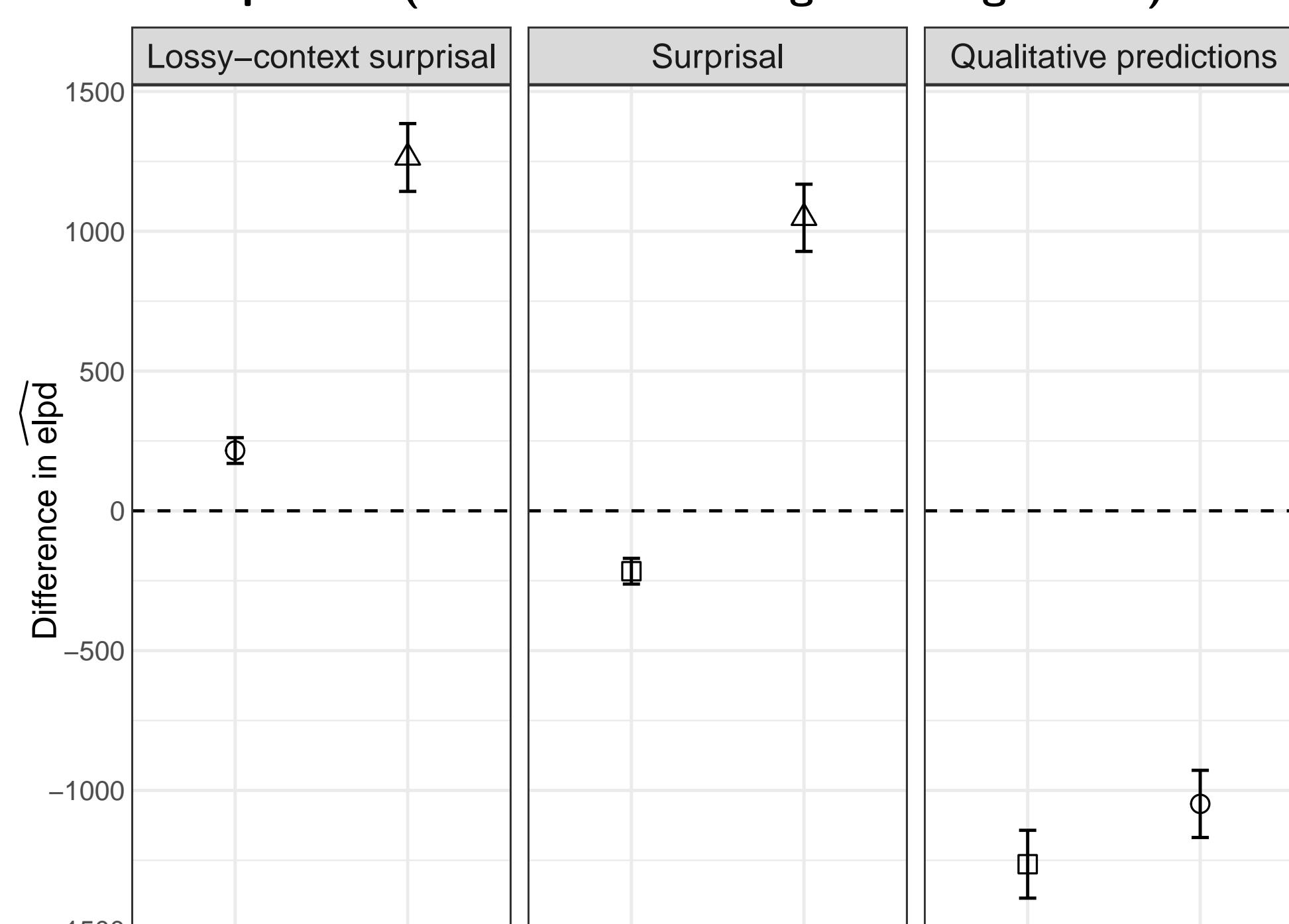
SBIN (2x2): Similarity-Based Interference
Subject-Cue [Yes/No] \times Animacy-Cue [Yes/No] — closely replicating [17]

RCSO (2x2): Subject vs. Object Relative Clauses
Subject/Object \times Double-/Single-Embedding — German adaptation of [7]

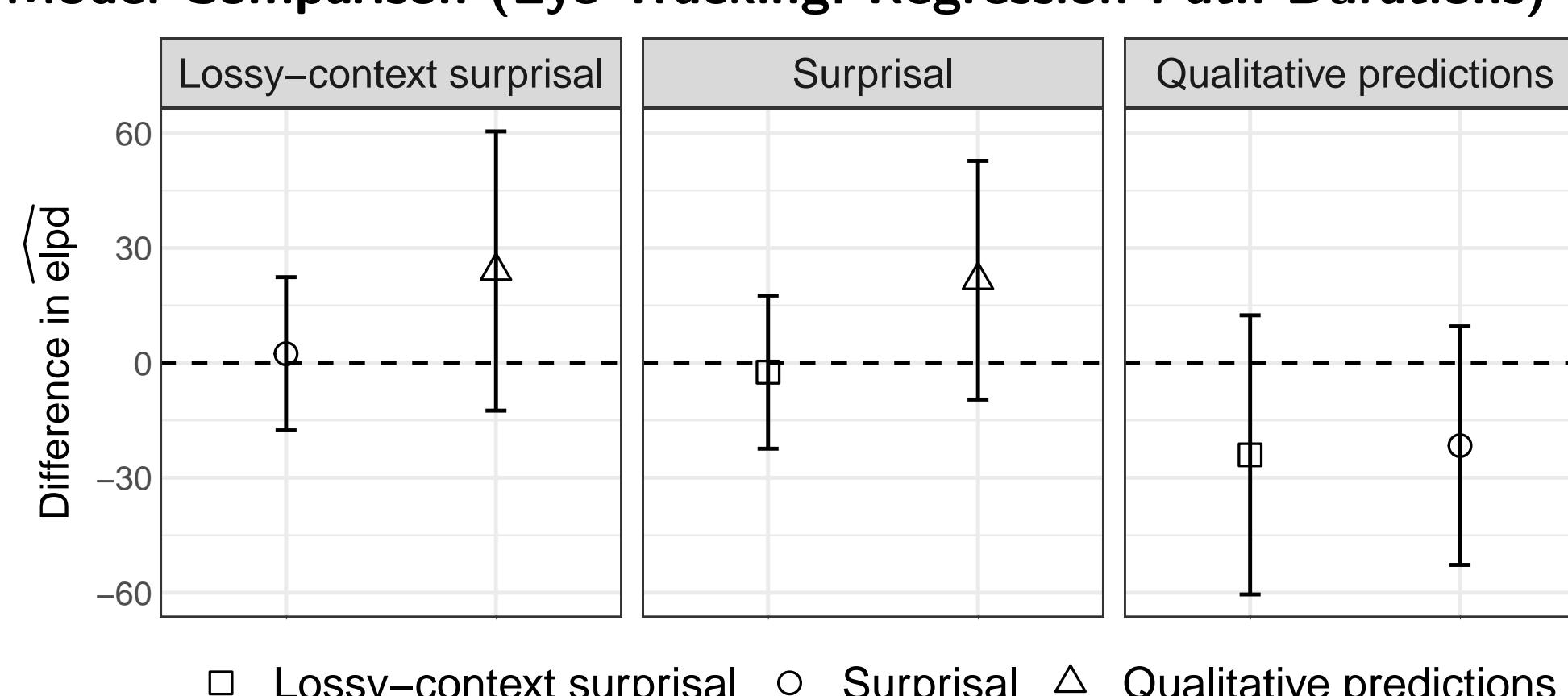
SYAA (3x1): Syntax-Based Attachment Ambiguity
High-/Low-/Ambiguous-Attachment — closely replicating [12]

SEAA (3x1): Semantics-Based Attachment Ambiguity
High-/Low-/Ambiguous-Attachment — German adaptation of [18]

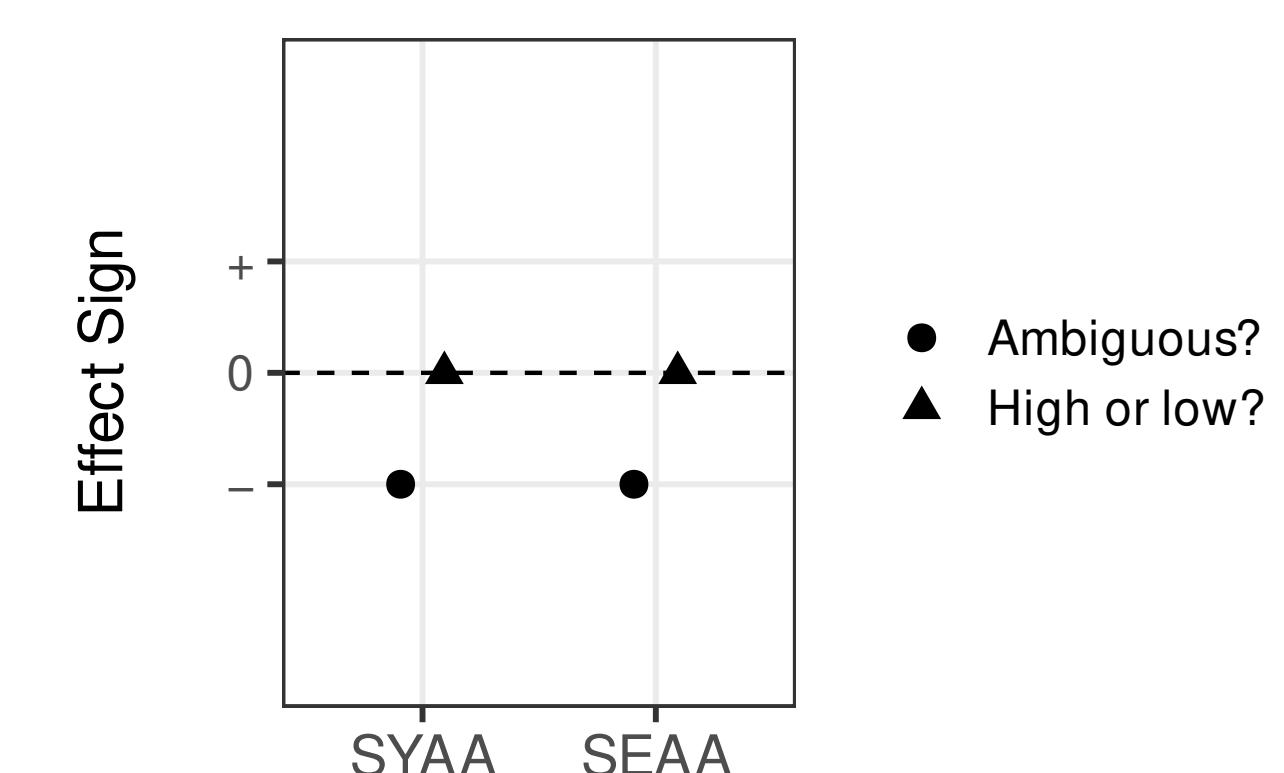
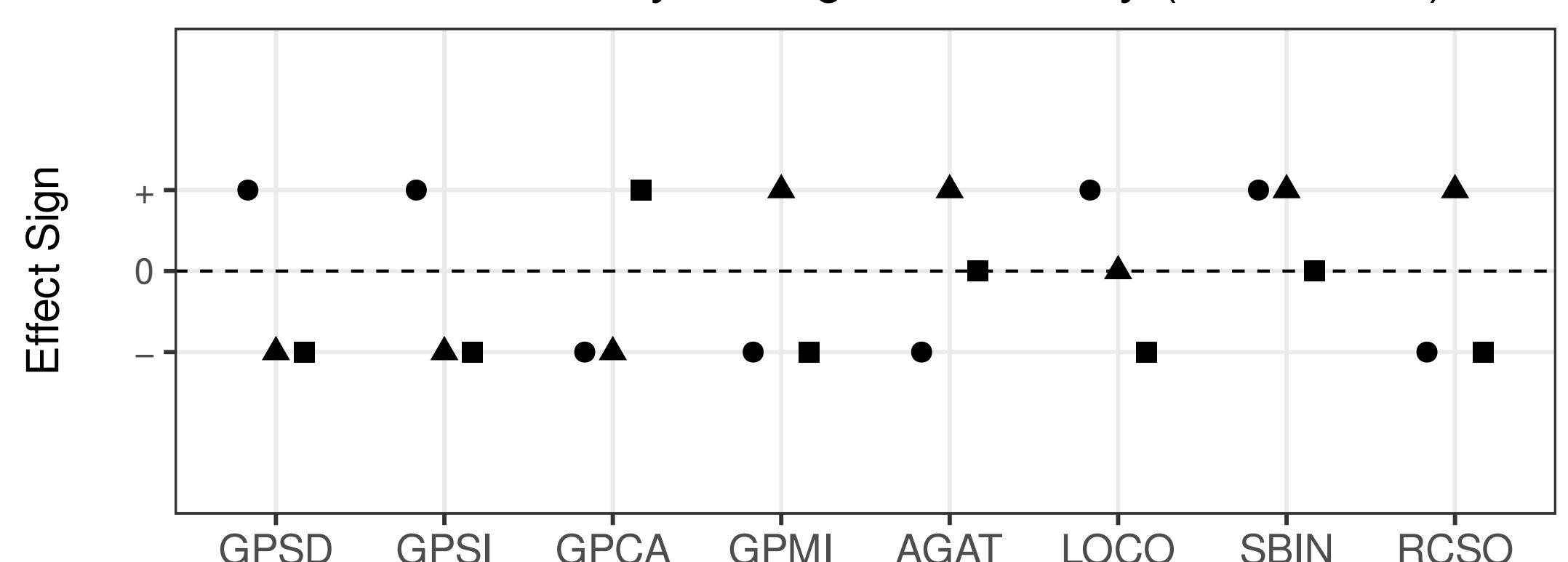
Model Comparison (Self-Paced Reading: Reading Times)



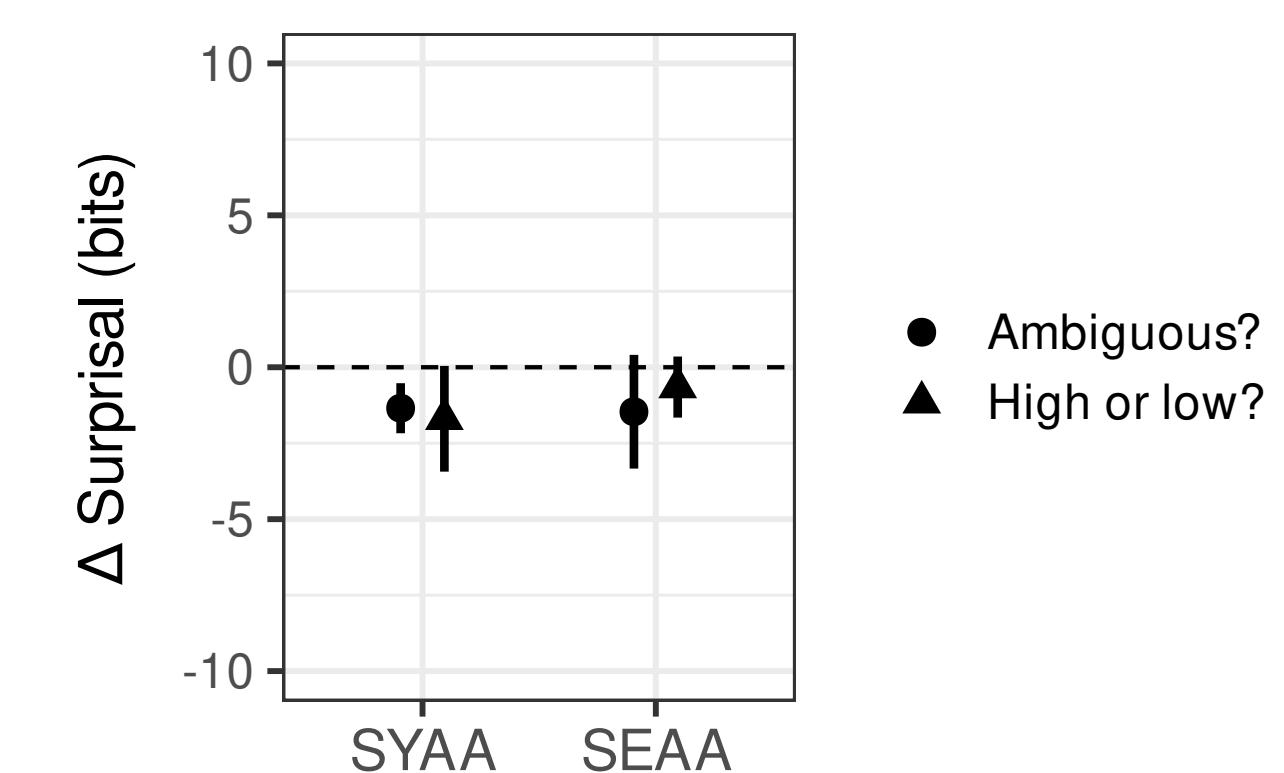
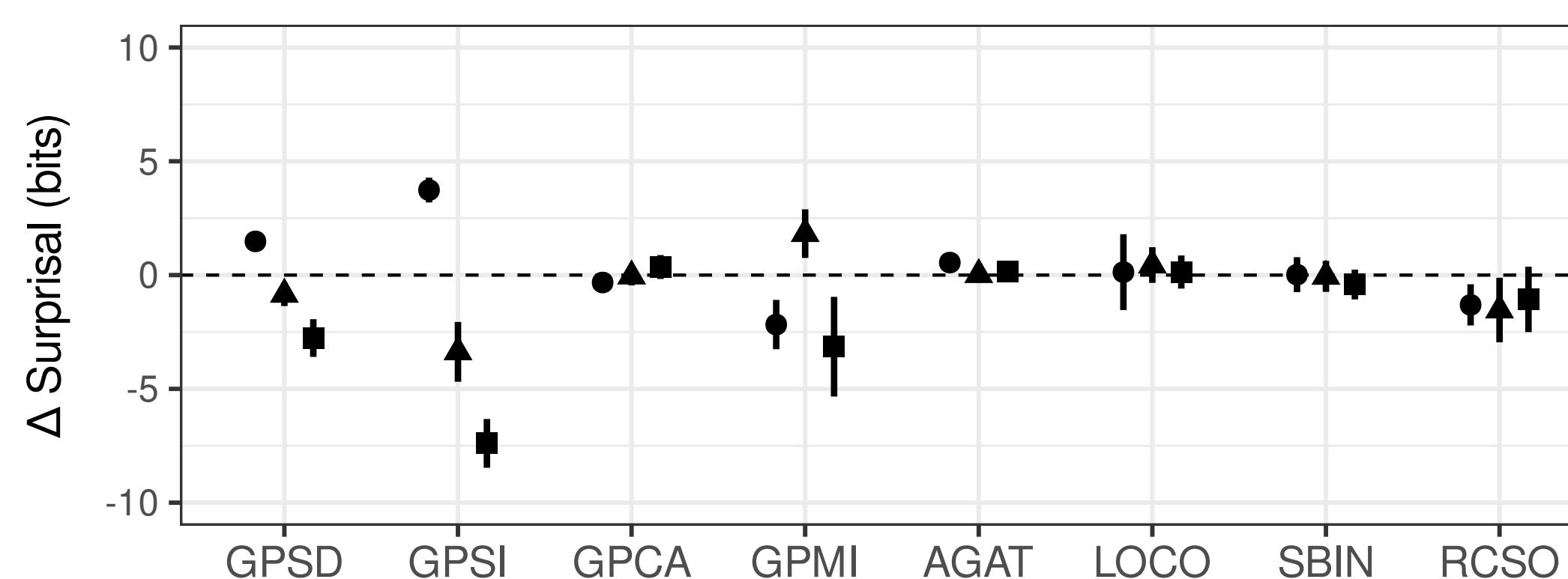
Model Comparison (Eye Tracking: Regression Path Durations)



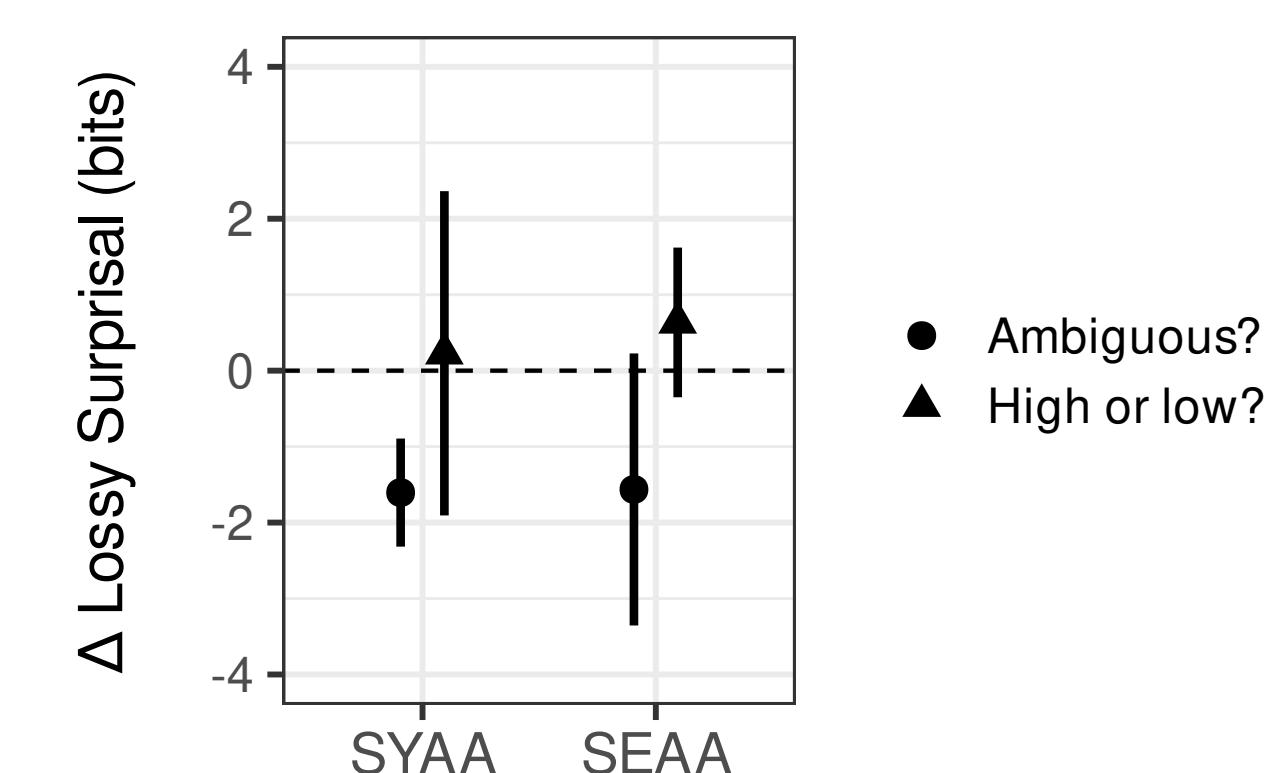
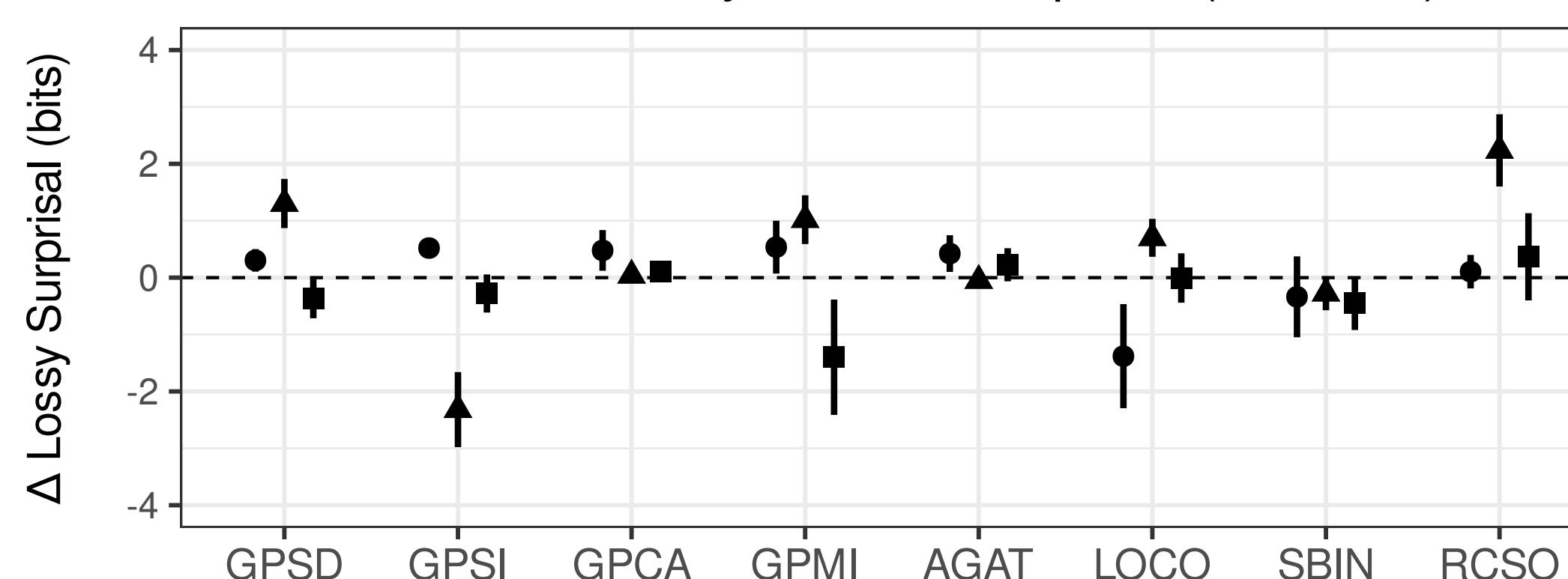
Predictions From Psycholinguistic Theory (Qualitative)



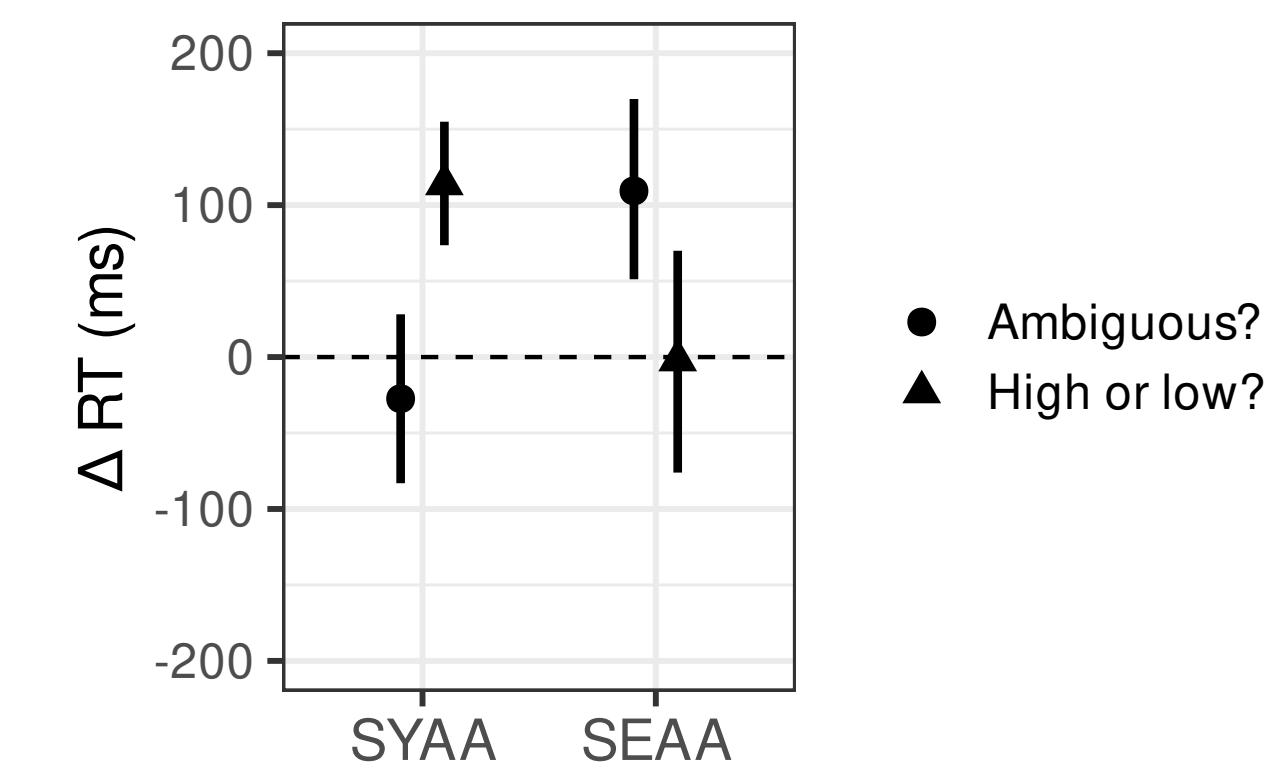
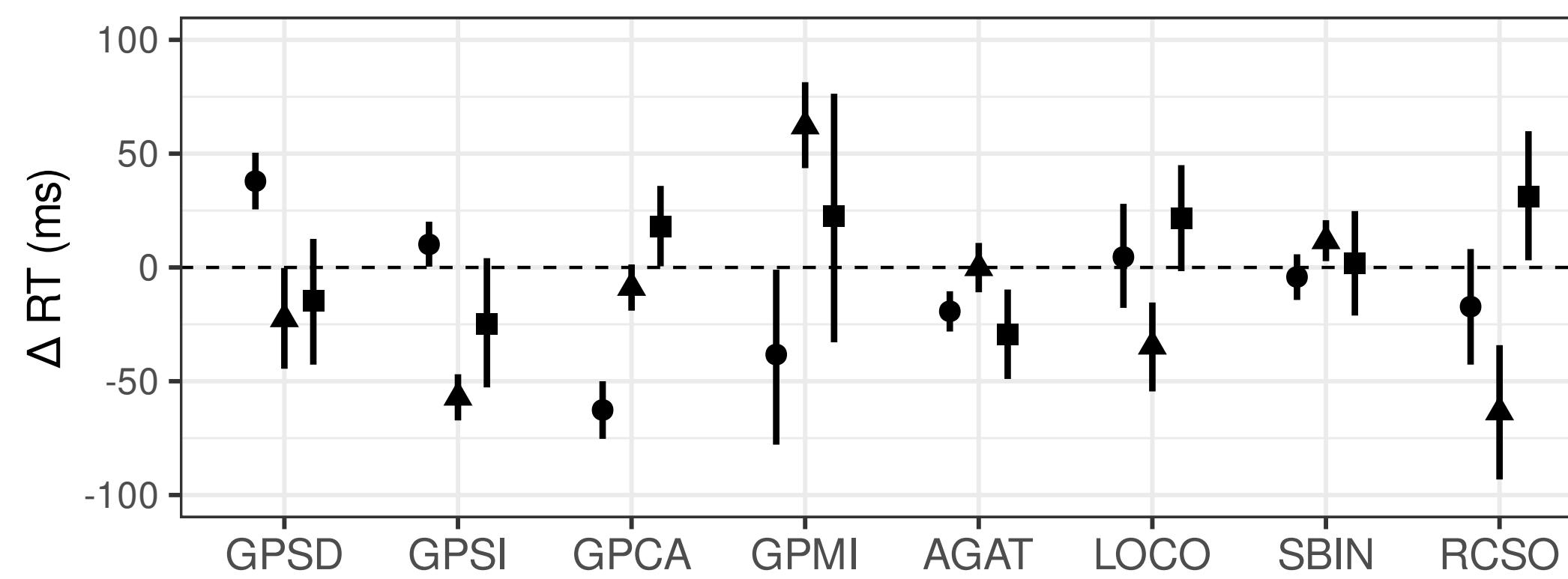
Predictions From Surprisal (95% CIs)



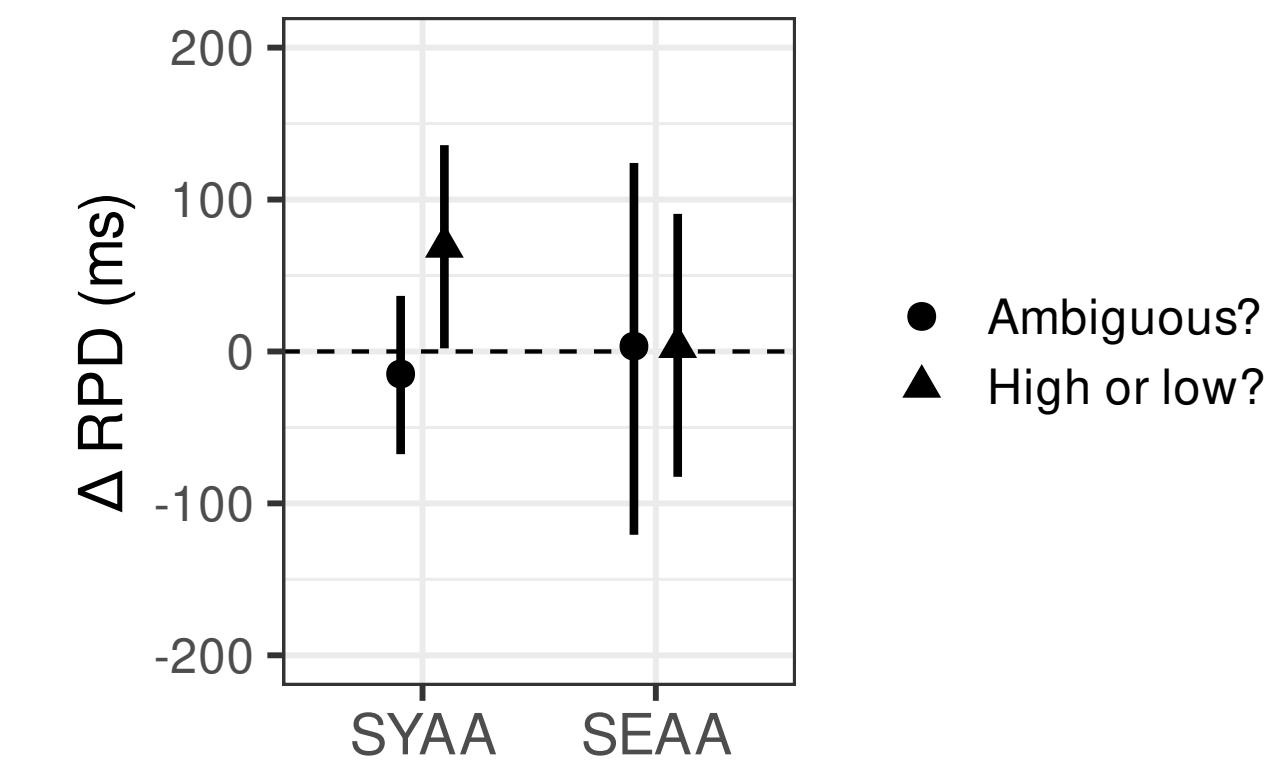
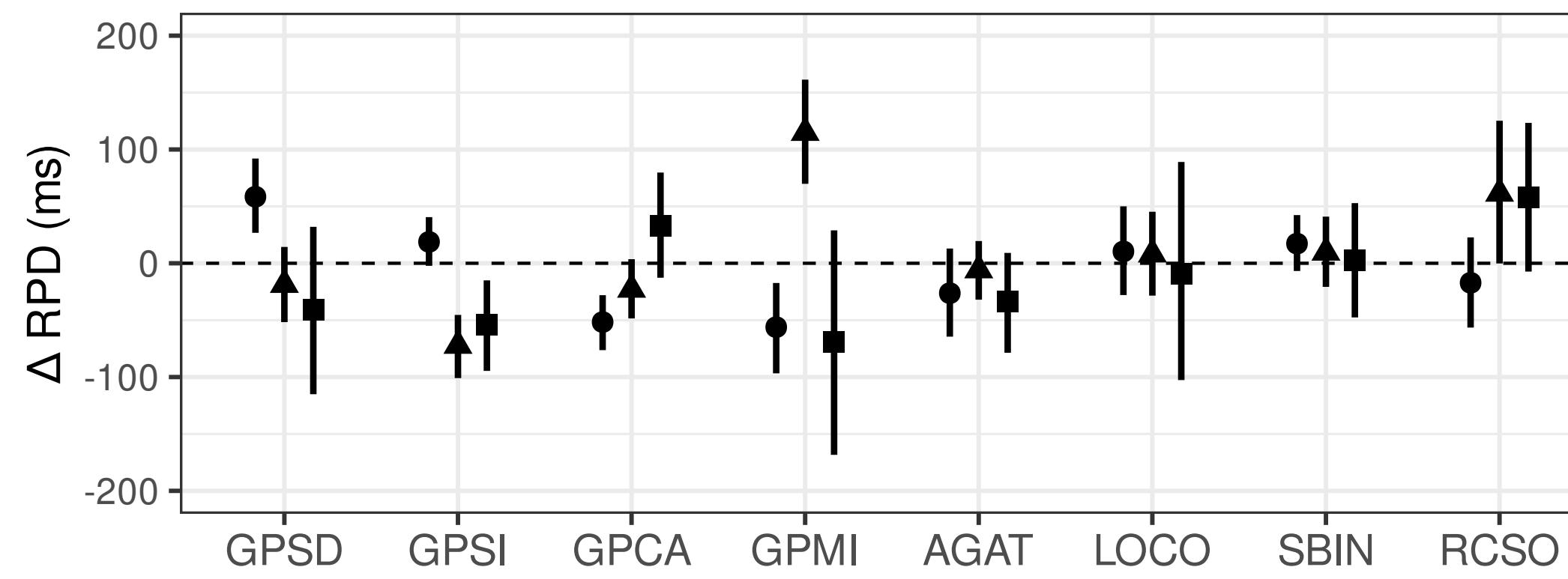
Predictions From Lossy-Context Surprisal (95% CIs)



Self-Paced Reading: Reading Times (RT), Critical Region (95% Crls)



Eye Tracking: Regression Path Durations (RPD), Critical Region (95% Crls)



References

[1] Bayerische Staatsbibliothek. <https://huggingface.co/dbmdz/german-gpt2>. 2020. [2] J. Devlin et al. In: *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (long and short papers)*. 2019, pp. 4171–4186. [3] R. Futrell et al. In: *Language Resources and Evaluation* 55 (2021), pp. 63–77. [4] J. T. Hale. In: *Proceedings of the Second Meeting of the NAACL*. Pittsburgh, PA, 2001. [5] J. Häussler. PhD thesis. University of Konstanz, 2009. [6] J. Hennert et al. Unpublished manuscript. 2025. [7] F. Hsiao and E. Gibson. In: *Cognition* 90.1 (2003), pp. 3–27. [8] K.-J. Huang et al. In: *Journal of Memory and Language* 137 (2024), p. 104510. [9] A. van Kampen. PhD thesis. Free University of Berlin, 2001. [10] L. Konieczny, B. Hemforth, and C. Scheepers. In: *German Sentence Processing*. Ed. by B. Hemforth and L. Konieczny. Springer, 2000, pp. 247–278. [11] R. Levy. In: *Cognition* 106.3 (2008), pp. 1126–1177. [12] P. Logačev. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 49.9 (2023), p. 1471. [13] M. Meng and M. Bader. In: *Language and Speech* 43.1 (2000), pp. 43–74. [14] M. Meng and M. Bader. In: *Language and Cognitive Processes* 15.6 (2000), pp. 615–666. [15] D. Paape and S. Vasishth. In: *Language and Speech* 59.3 (2016), pp. 387–403. [16] A. Radford et al. In: *OpenAI Blog* 1.8 (2019), p. 9. [17] P. Schoknecht, H. Yadav, and S. Vasishth. In: *Journal of Memory and Language* 141 (2025), p. 104599. [18] M. J. Traxler, M. J. Pickering, and C. Clifton Jr. In: *Journal of Memory and Language* 39.4 (1998), pp. 558–592. [19] A. Vehtari, A. Gelman, and J. Gabry. In: *Statistics and Computing* 27 (2017), pp. 1413–1432.