# A Multilingual Cross-Method Dataset for Evaluating Measurement Reliability of Individual Differences in Naturalistic Reading

**Shan Gao**[*1], **Cui Ding**[1], **Lena A. Jäger**[1]

[1]Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

*contact: shan.gao@uzh.ch

**Background:** Incremental reading paradigms are widely used in psycholinguistic research to measure language-processing difficulty during reading. Previous comparisons of online reading paradigms, including eye-tracking, self-paced reading (SPR), and the maze task, have shown considerable similarities in the general pattern of effects, while also revealing differences in the strength, timing, and localization (Witzel *et al.*, 2012). More recent work has further refined existing tools, such as the A-maze task (Boyce *et al.*, 2020), bidirectional SPR (BSPR; Paape and Vasishth, 2022), and introduced new paradigms, such as Mouse Tracking for Reading (MoTR; Wilcox *et al.*, 2024), often with explicit comparisons to established methods. At the same time, these paradigms are increasingly being used to investigate individual differences in sentence processing. Although the strengths and limitations of these paradigms for detecting population-level effects are increasingly well understood, most comparisons have focused on English. It remains unclear whether methodological differences across paradigms systematically affect their reliability for measuring individual differences across languages. To address this gap, we present a multilingual cross-method dataset in which participants read naturalistic texts using SPR, maze, and MoTR, with two sessions per method. We use this dataset to estimate the within-method cross-session measurement reliability and the cross-method reliability of individual differences in naturalistic reading.
**Method:** Native speakers of English, Chinese, Croatian, and Farsi complete the experiment in two sessions. In each session, participants read naturalistic materials using three reading paradigms: the maze task, SPR, and MoTR. The order of paradigms is randomly assigned. The stimuli include two types of materials. First, participants read out-of-context naturalistic sentences sampled from the same books or source types as the MultiplEYE materials, but from passages that do not overlap with the original MultiplEYE texts (Jakobi *et al.*, 2025). Second, participants read longer texts composed of multiple passages, using the original MultiplEYE text stimuli. For each paradigm, participants first complete the out-of-context sentence-reading task and then read one MultiplEYE text in that paradigm. Each text is followed by the original, carefully designed MultiplEYE comprehension questions. To assess reliability, we use the two-task Bayesian hierarchical model adapted to naturalistic reading by Haller *et al.* (2026). We estimate both within-method cross-session reliability and cross-method reliability, focusing on individual differences in sensitivity to three key psycholinguistic predictors of reading: word length, lexical frequency, and predictability.
**Results & Discussion:** Data collection is scheduled to begin shortly. The resulting dataset will provide a multilingual resource for evaluating the suitability of web-deployable reading paradigms for individual-differences research across languages and writing systems. The planned analyses will assess how reliably SPR, maze, and MoTR capture participant-level sensitivity to word length, lexical frequency, and predictability. By estimating within-method cross-session reliability, the study will assess whether each paradigm captures stable individual-level variation over time. By estimating cross-method reliability, it will assess whether different paradigms capture convergent participant-level differences in naturalistic reading.

## References

Boyce V., *et al.* (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *JML*.

Haller P., *et al.* (2026). Replicate me if you can: Assessing measurement reliability of individual differences in reading. *Cogn. Sci.*.

Jakobi D. N., *et al.* (2025). MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proc. ETRA*.

Paape D., Vasishth S. (2022). When nothing goes right, go left: A large-scale evaluation of bidirectional self-paced reading.

Wilcox E. G., *et al.* (2024). Mouse tracking for reading (MoTR): A new naturalistic incremental processing measurement tool. *JML*.

Witzel N., *et al.* (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *J. Psycholinguist. Res.*.