Spillover as a Test of Cognitive Models of Reading: E-Z Reader vs. SWIFT

Cui Ding (cui.ding@uzh.ch)[1], Lena A Jäger[1],Erik D. Reichle[2]

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

[2]School of Psychological Sciences, Macquarie University

Cognitive models of reading explain how lexical processing, attention, and oculomotor control shape eye movement behavior. Two leading models embody different theories of reading: E-Z Reader [1] assumes serial attention allocation and staged lexical processing that controls saccade programming, whereas SWIFT [2] assumes parallel lexical processing within a distributed attentional gradient and an autonomous saccade timer. These differences should affect both aggregate reading measures—fixation durations, skipping, and regressions—and the way linguistic effects of the stimulus unfold over time and across words.

Prior work. Existing comparisons of E-Z Reader and SWIFT are mostly qualitative or focus on aggregate surface-level reading measures for an "average" reader. Recent studies that use both models treat them mainly as baselines for evaluating deep neural generative models of eye movements. It remains unclear whether E-Z Reader and SWIFT capture how linguistic predictors shape reading: local effects, spillover, temporal effect distribution, and individual-level predictor effects.

Current study. We test E-Z Reader and SWIFT using spillover and individual differences in linguistic effects as diagnostics of reading mechanisms. We ask whether the models simulate population-level effects of word length, lexical frequency, and surprisal; how these effects spill over to the subsequent word; and how these effects vary across readers.

Method. For each reader, we fit E-Z Reader and SWIFT separately and evaluate their simulations on held-out test data from three sentence-reading corpora: English CELER [1], Dutch RaCCooNS [2], and German OCULO [3]. We compare model performance across surface reading measures, local and spillover effects, temporal effect distribution, and individual-level predictor effects.

Results. Model performance differed by the level of analysis. SWIFT better simulated participant-level reading measures, whereas E-Z Reader better captured word-level measures. E-Z Reader was generally better calibrated for population-level effects of word length, frequency, and surprisal (Fig. 1). SWIFT overestimated local effects, particularly word length and frequency, and showed larger temporal distribution mismatches (Tab. 1). For individual differences, both captured general reading speed better than the reader-specific effects of linguistic predictors. Neither model recovered individual spillover effects (Fig. 2).

Summary. Analyzing spillover reveals how models distribute processing difficulty over time, a critical step for teasing apart theories of serial versus parallel processing. Weak individual-level recovery suggests that current models might require systematic, theoretically motivated mechanisms for individual differences, beyond reader-specific parameter fitting.

References

[1] Engbert et al. (2005). SWIFT model of saccade generation. Psychological Review, 112.

[2] Veldre et al. (2023). Visual constraints on lexical processing. JEP: General, 152, 693.

[3] Berzak et al. (2022). CELER L1/L2 English reading corpus. Open Mind, 6, 41–50.

[4] Frank & Aumeistere (2024). Eye-tracking/EEG sentence corpus. LRE, 58, 641–657.

[5] Chandra et al. (2020). Mirrored/inverted text reading. Scientific Reports, 10, 4210.

Figures and Tables

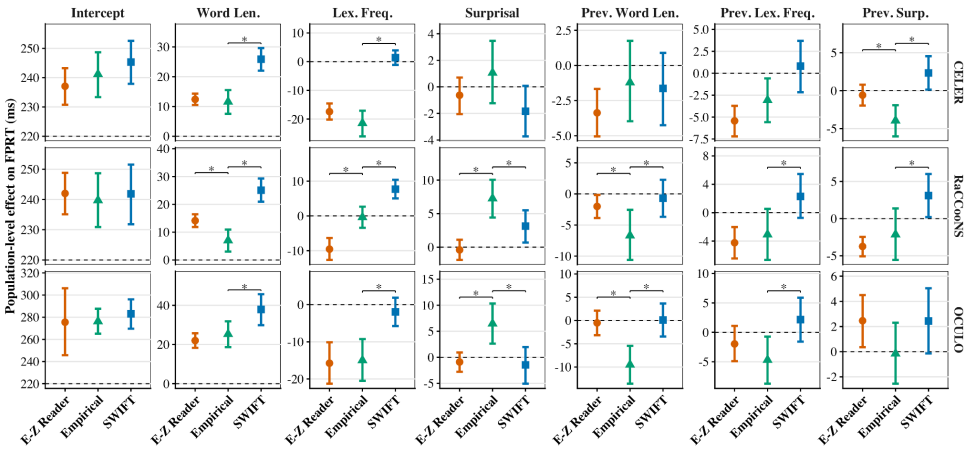

Fig. 1: Population-level effects on gaze durations in empirical data and model simulations across datasets. Points show posterior means; error bars show 95% CrIs; asterisks mark where the model vs. empirical difference CrI excludes zero.

Tab. 1: Temporal distribution of local and spillover effects. Values are posterior means with 95% credible intervals; bold indicates that the directions of the models' estimates match those in the empirical data.

Temporal Distribution		Word Len.		Lex. Freq.		Surprisal	
		Total Magnitude (ms)	Local Prop.	Total Magnitude (ms)	Local Prop.	Total Magnitude (ms)	Local Prop.
CELER	Empirical	13.14 [8.71, 17.62]	0.89 [0.72, 1.00]	24.49 [19.56, 29.59]	0.88 [0.78, 0.97]	5.23 [2.66, 8.20]	0.23 [0.01, 0.51]
	E-Z Reader	**15.77 [13.19, 18.28]**	**0.79 [0.70, 0.88]**	**22.84 [19.48, 26.18]**	**0.76 [0.69, 0.83]**	1.52 [0.31, 3.28]	0.51 [0.03, 0.96]
	SWIFT	**27.66 [23.37, 32.26]**	**0.94 [0.86, 1.00]**	2.96 [0.59, 6.17]	0.53 [0.05, 0.97]	4.21 [1.55, 7.11]	0.45 [0.06, 0.87]
RaCCooNS	Empirical	13.72 [7.89, 19.26]	0.51 [0.29, 0.75]	4.39 [0.93, 8.50]	0.30 [0.01, 0.82]	9.61 [5.82, 14.03]	0.77 [0.53, 0.98]
	E-Z Reader	**16.16 [13.37, 19.07]**	**0.88 [0.78, 0.98]**	**13.80 [9.99, 17.54]**	**0.69 [0.56, 0.83]**	4.44 [3.00, 6.01]	0.15 [0.01, 0.37]
	SWIFT	**26.42 [22.02, 31.22]**	**0.95 [0.87, 1.00]**	10.08 [6.44, 14.14]	0.77 [0.56, 0.99]	6.26 [2.87, 9.69]	0.51 [0.16, 0.90]
OCULO	Empirical	34.82 [26.83, 42.81]	0.73 [0.62, 0.83]	19.68 [13.05, 26.29]	0.76 [0.59, 0.95]	7.49 [3.37, 11.65]	0.86 [0.62, 0.99]
	E-Z Reader	**23.16 [19.09, 27.31]**	**0.95 [0.87, 1.00]**	**17.87 [11.93, 23.79]**	**0.88 [0.74, 0.99]**	3.53 [1.17, 6.11]	0.30 [0.02, 0.74]
	SWIFT	39.30 [31.02, 47.43]	0.96 [0.90, 1.00]	4.71 [1.01, 9.37]	0.49 [0.04, 0.95]	4.34 [1.17, 8.30]	0.41 [0.03, 0.88]

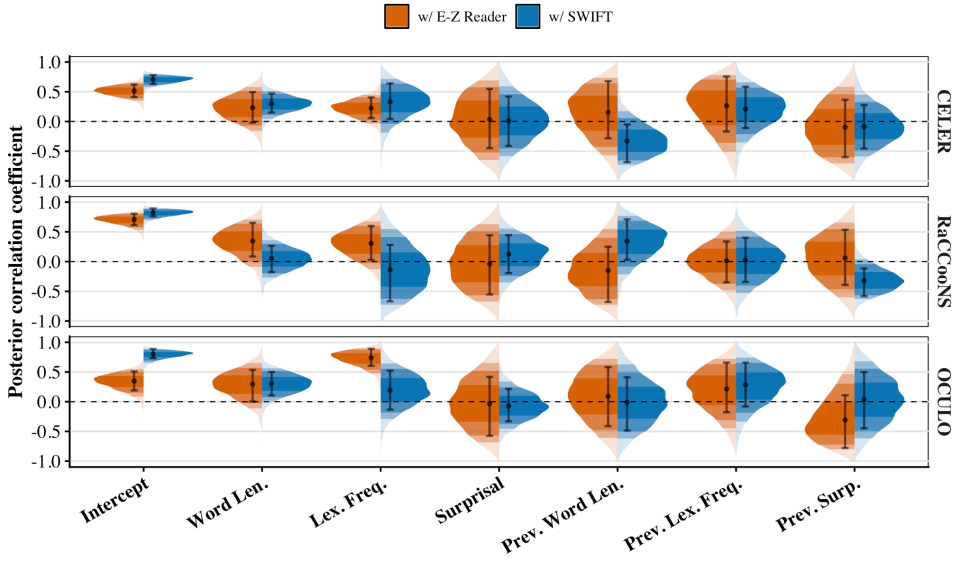

Fig. 2: Participant-level correlations btw. empirical & model-simulated predictor effects across datasets. The left and right halves show the correlation distributions for E-Z Reader and SWIFT, respectively; points/error bars show medians and 75% HPDIs.

Optional Supplemental Information

Dataset	Language	Stimulus predictors: mean (SD)			#Participants	#Word-level observations	
		Word Len.	Lex. Freq.	Surprisal		Total	FPRT
CELER	English	4.83 (2.61)	5.66 (1.43)	6.24 (4.33)	69	99,663	60,457
RaCCooNS	Dutch	4.31 (2.24)	6.01 (1.19)	5.72 (4.09)	37	84,613	50,148
OCULO	German	5.17 (2.77)	5.43 (2.37)	5.24 (3.60)	36	53,745	35,047

Tab. 2: Overview of the datasets. Predictor statistics of the stimulus are summarized as mean (SD). Lexical frequency was lemma frequency. Observations report the total number of word-level observations and the subset with valid first-pass reading time (FPRT) used in the main analyses.