Maze Made Easy: Better and easier measurement of incremental processing difficulty

Veronica Boyce (MIT), Richard Futrell (UCI), Roger Levy (MIT)
vboyce@mit.edu

The Maze task (Freedman & Forester, 1985) has shown promise as a way to measure incremental processing difficulty with high sensitivity and accuracy. In particular, it has been claimed to avoid the spillover effects endemic to self-paced reading (SPR) (Witzel et al., 2012). Here we demonstrate that the Maze task can be run reliably over the web, and that it is a substantially more sensitive instrument for measuring incremental processing difficulty than SPR on Mechanical Turk. Furthermore, we demonstrate and validate a method for automatically generating materials which dramatically reduces the effort involved in preparing a Maze task experiment while yielding the same sensitivity and accuracy. The resulting "Auto-Maze" task provides all the advantages of Maze while being as easy to prepare and run as SPR.

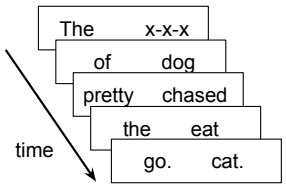
In the Maze task, participants read a sentence word by word (see Figure 1). For each word position, they see two words, one of which is the next word in the sentence and one of which is a distractor. In **G**(rammatical)**-maze**, the distractor is a real word, but not a grammatical continuation. In **L**(exical)**-maze**, the distractor is a nonce word. Participants press a key to indicate which word continues the sentence, and then see the next pair of words. The time between key presses is the dependent measure. If a participant makes a mistake, the sentence terminates, and they continue to the next sentence.

Our **Experiment 1** is a replication of the G-maze, L-maze, and SPR portions of Witzel et al. (2012) using crowdsourced participants. Witzel et al. (2012) compared the performance of G-maze and L-maze with SPR and eye-tracking on three types of temporary structural ambiguities: relative clause attachment height, adverb clause attachment height, and sentence versus noun phrase conjunctions (S v NP) (see Figure 2 for examples). They found that G-maze was the most sensitive for the Relative and Adverb conditions, localizing the slowdown strongly on the critical word, but only eye-tracking was sensitive to the S v NP ambiguity. We replicated the G-maze, L-maze, and SPR results successfully over Amazon Mechanical Turk with 50 participants for each task, using the same distractor words as Witzel et al. (2012). Each participant saw 8 practice items, and then 24 sentences of each type (half in each condition) mixed in with 24 filler items. Comprehension questions were used for the SPR task, but not for either Maze task. Figure 3 shows that G-maze had the largest and most immediate, localized effects, although none showed significant effects on the S v NP condition.

Next we introduce **A**(uto)**-maze**: a version of G-maze in which distractor words are generated automatically, thus massively reducing experimenter effort and preparation time. We generate distractors by harnessing recent advances in NLP. Distractor words are chosen to be matched in frequency and length with target words, but with much lower probability under state-of-the-art LSTM language models (Jozefowicz et al., 2016; Gulordava et al., 2018). In **Experiment 2**, we use A-maze to successfully replicate the G-maze results. As shown in Figure 3, the A-maze results are even stronger than the results using the original hand-crafted distractors: using distractors from one of the LSTM language models, we even find a significant and localized effect for the S v NP ambiguity, previously detectable only by eye-tracking. The strength of A-maze effects makes high-powered studies possible with fewer participants (see Figure 4 for results of a power simulation).

Our results unlock the potential of the Maze task by removing three hurdles to its adoption: (1) we show that it can be run reliably in a crowdsourced format; (2) we show how to automatically generate distractors; and (3) we show that the results are relatively invariant to the choice of distractors. We make our A-maze generation code, as well as the Ibex code for the web-based Maze task, freely available online at github.com/vboyce/Maze.

Figure 1: Sample Maze



Participants see two words at a time and have to select the correct word. They then see the next pair of words.

Figure 2: Sample Stimuli (disambiguating words are highlighted):

Relative Clause– Low attachment:
The son of the lady who politely introduced herself was popular at the party.
Relative Clause – High attachment:
The son of the lady who politely introduced himself was popular at the party.
Adverb clause – Low attachment:
James will fix the car he drove yesterday, but he will need some help.
Adverb Clause – High attachment:
James will fix the car he drove tomorrow, but he will need some help.
Sentence v Noun Phrase conjuction (S v NP) – With comma:
The swimmer disappointed her coach, and her mother tried to console her.
Sentence v Noun Phrase conjuction (S v NP) – No comma:
The swimmer disappointed her coach and her mother tried to console her.

Figure 3: Results

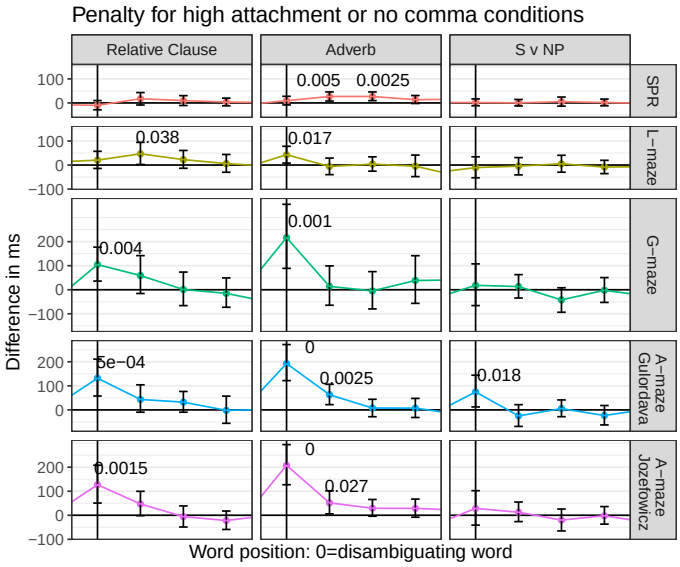
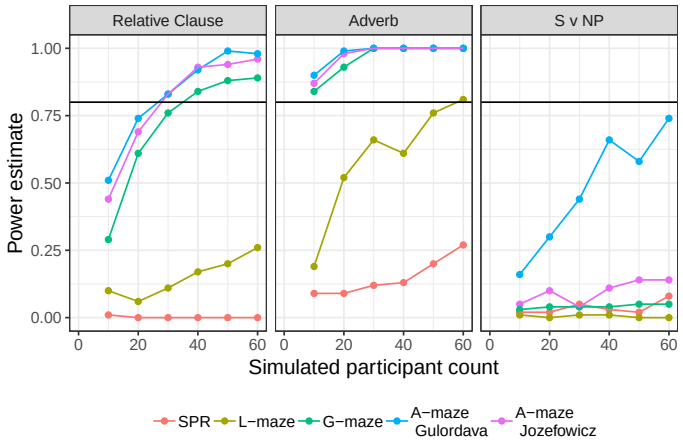


Figure 3 shows the mean difference in reading time between the dispreferred conditions (high attachment or no comma) and the preferred conditions. For the Relative and Adverb conditions, G-maze and both A-mazes show a significant difference at the disambiguation word (word 0).

Figure 4 shows the estimated power for different participant numbers. As SPR has spillover effects, power for SPR was calculated on the summed 0-3 word region. G-maze and A-maze have much higher power than SPR or L-maze in Relative and Adverb conditions.

Figure 4: Power Simulation



References: Witzel, N., Witzel, J., & Forster, K. (2012). *J Psycholinguist Res* 41(2): 105-128.• Freedman, S. and Forster, K. (1985). *Cognition* 19. • Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). *Proceedings of NAACL-HLT 2018.* 1195-1205. • Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410.