**Title:** Scaling up psycholinguistics with Pushkin

Progress in psycholinguistics is often limited by sampling scale: it is cost-prohibitive to test a comprehensive stimulus set across the appropriate range of demographics and with adequate statistical power (cf. Pashler & Harris, 2012; Yarkoni, 2022). Studying individual differences or demographic effects only multiplies these challenges. Massive online experiments (MOEs), in which tens of thousands of volunteers take part, would allow for much faster scientific progress in theory (Hartshorne et al., 2019). In practice, deploying them requires substantial time and Web development expertise. While there are good tools for smaller, paid-subject studies (PCIbex, Gorilla, jsPsych), MOEs present additional design challenges (e.g. gamifying to attract volunteers) and technical challenges (handling heavy traffic, securely saving data, privacy, etc.).

We present updates to Pushkin, a free-and-open-source integrated toolkit for building and managing MOEs. Pushkin uses familiar jsPsych syntax for stimulus presentation and response recording; and adds support for maintaining a stand-alone website such as gameswithwords.org or themusiclab.org, including setting up and managing web servers, adding and removing experiments, utilizing gamification elements, and other necessities of MOEs. Over the past year, we have expanded Pushkin's functionality and improved its reliability. The most important functionality change is adding support for persistent logins (thus supporting longitudinal studies) while protecting subject anonymity and ensuring GDPR compliance.To our knowledge, there have been no longitudinal MOEs to date, despite the obvious advantages of not requiring subjects to physically return to a laboratory. The resulting MOEs can produce datasets with multitudes more data points both between and *within* subjects than in-person lab experiments, enabling finer-grained group comparisons and individual-level developmental trajectories.

Reliability has been improved through infrastructure modernization: cloud services with built-in error handling and retry logic, serverless compute that automatically scales with traffic demand, and an automated testing suite to validate performance before deployment. As an example, we recently implemented load testing. MOEs depend heavily on viral traffic, but large numbers of subjects participating simultaneously can overwhelm Web servers. We developed an automated testing framework to predict website capacity before deployment. Using this framework, we validated infrastructure scalability by testing two site configurations: a baseline configuration and an "upgraded" configuration with doubled database size, doubled experiment server memory, and database connection pooling. Both sites ran an identical serial reaction time experiment (7 blocks × 60 trials; Nissen & Bullemer, 1987). We measured success rates—the proportion of users successfully completing the experiment with data saved (Figure 1)—and P95 response times—the latency threshold below which 95% of server requests complete (Figure 2). We tested these metrics across loads from 1 to 30 concurrent participants.

Results revealed precise capacity limits: baseline configuration sustained 100% success (P95=176-238ms) until 10 concurrent users (success=29%); the "upgraded" configuration improved capacity (P95=66-176ms) but failed beyond 15 users (success=48% at 20 users). Both configurations handled 300 simultaneous logins with 100% success (median=31ms; P95=529ms), confirming robust support for persistent user logins and identifying experiment concurrency as the primary scaling bottleneck. In real-world terms, the "upgraded" configuration currently supports ~2.7K experiments/day at ~$87/month, and our load testing demonstrates an empirically-validated path for iterative infrastructure optimization toward production targets.

We illustrate both new and existing functionality of Pushkin with examples of prior MOEs (e.g., Aguasvivas et al., 2020; Hartshorne & Snedeker, 2013) and potential uses for future work.

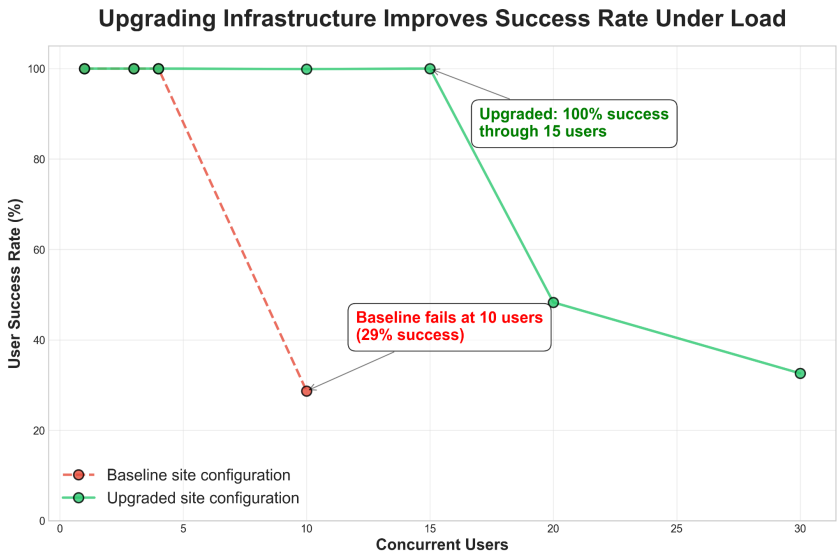


*Figure 1:* Comparison of experiment completion success rate at increasing concurrent user loads between baseline and upgraded configurations of Pushkin site infrastructure

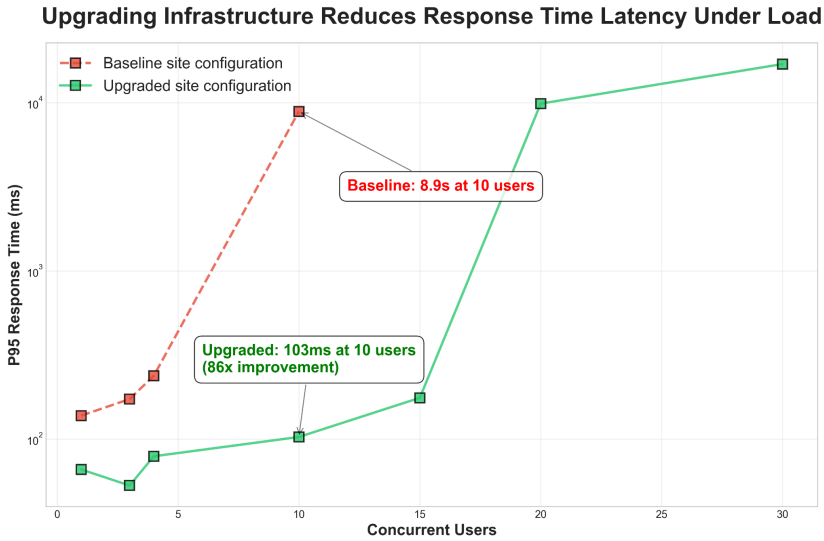


*Figure 2:* Comparison of latency threshold below which 95% of server requests complete (P95) at increasing concurrent user loads between baseline and upgraded Pushkin site configurations.

**Pashler, H., & Harris, C. R.** (2012). Is the replicability crisis overblown? Three arguments examined. Perspectives on Psychological Science, 7(6), 531-536. | **Yarkoni T.** (2020) The generalizability crisis. Behav Brain Sci. | **Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J.** (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. BRM, 51(4), 1782-1803. | **Nissen, M. J., & Bullemer, P.** (1987). Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19(1), 1–32. | **Aguasvivas, J., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A.** (2020). How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy. BRM, 52(5), 1867-1882. | **Hartshorne, J. K., & Snedeker, J.** (2013). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. Language and Cognitive Processes, 28(10), 1474–1508.