Online behavioral studies are safe for now: unusual RTs do not imply bots (Reply to Van der Stigchel et al., 2026)

Andrey Chetverikov
University of Bergen, Norway
andrey.chetverikov@uib.no

Van der Stigchel, Strauch, de Zwaart, and Van Maanen (2026; further referred to as SSZM) recently raised concerns about bots in online behavioral data collection. While concerns are warranted for open-ended responses, they are premature for typical cognitive studies.

SSZM analyzed Posner cueing task data using three response time (RT) metrics: the correlation between RTs and their normal quantiles, the mean and variability relationship across conditions, and autocorrelation across trials. Strikingly, their Figure 1 shows a suspected bot with no autocorrelation. Reanalysis of their data (https://osf.io/cdu98), selecting three "suspected bots" based on the lowest autocorrelation, revealed one with near-chance performance (Figure 1A). If it is a bot, it is a very bad one, but likely it is a poorly motivated, randomly responding human. Another had the slowest RT, suggesting a deliberate strategy. None of the three stood out based on the other two metrics. Furthermore, there is no bimodality, indicative of mixed data sources, in any of the metrics (Figure 1B-D). A lack of autocorrelation is common in their data.

To test whether the suggested bot-detection metrics are meaningful, I reanalyzed data from an offline study with the Posner cueing task (Hedge, Powell, and Sumner 2018). Despite better performance, likely due to a simpler task and higher motivation (Figure 1E), there was even less autocorrelation and similar variation in RT distributions' shapes. The only clearly different metric was the RT mean and variability correlation, illustrating that a positive correlation cannot be assumed. By SSZM's criteria, one should suspect bots among the undergraduates in Cardiff.

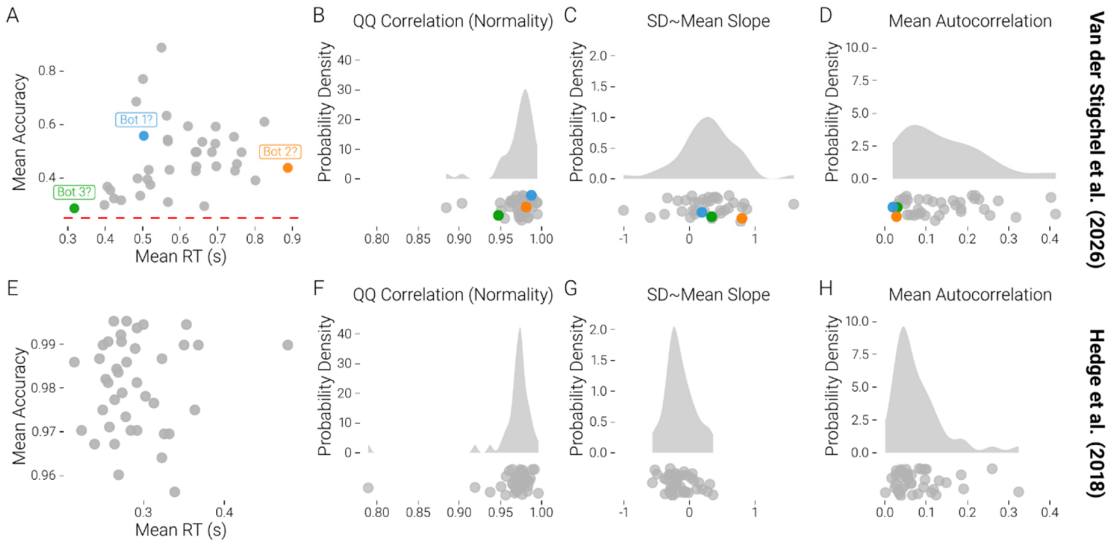

Figure 1: Comparison of online data from Van der Stigchel et al. (2026; top row), with three participants having the lowest autocorrelation highlighted as "suspected bots", and offline data from Hedge et al. (2018; bottom row). **A**, **E**: Mean RT and accuracy for each observer. The dashed red line in A indicates random performance level. **B-D**, **F-H**: Distributions of the three bot-detection metrics (top) and individual values for each observer (bottom). QQ correlation is a correlation between RTs and their normal quantiles. SD~Mean Slope is the regression slope between mean RT and its variability across conditions. Mean Autocorrelation is the mean autocorrelation across lags of 1 (immediately preceding) to 40 trials. All metrics are computed following SSZM's code.

There are also general reasons to be wary of bot detection metrics. Participants rarely conform perfectly to psychological laws, as these laws describe average behavior. Furthermore, when many metrics are used, every participant will be an outlier on one or more of them, as illustrated by Daniels' (1952) report on

anthropometric measurements of US Air Force personnel, with none of 4000+ pilots qualifying as “average”. Like in open-text responses, where AI reveals itself by too fast typing, AI in behavioral data can only be revealed by clearly inhuman speed or accuracy. Indirect measures cannot be applied, similarly to how grammar quality cannot indicate AI in textual responses.

But should we worry about bots even if there is no evidence for them in SSZM data? Creating an AI agent to mimic human behavior in an online study seems theoretically possible, but economically unfeasible given typical participant rewards. After half an hour using the latest code-focused LLM to develop a bot for my recent visual working memory experiment (Chetverikov and Hansmann-Roth 2026), I gave up. While still possible, the required time likely exceeds the experiment’s duration, and the trial-and-error development process would easily be detected. Alternatively, one could imagine a “general” bot, able to complete any study on the first pass, but until such bot is demonstrated, online studies seem safe.

SSZM call for “routine screening of all online RT datasets”. Scrutiny is always beneficial, but we should not stigmatize online data from cognitive behavioral studies, especially when evidence for bots is lacking.

References

Chetverikov, Andrey, and Sabrina Hansmann-Roth. 2026. “Noise in Competing Representations Determines the Direction of Memory Biases.” : 2025.12.17.694673. doi:10.64898/2025.12.17.694673.

Daniels, Gilbert S. 1952. *The Average Man?* Wright-Patterson Air Force Base.

Hedge, Craig, Georgina Powell, and Petroc Sumner. 2018. “The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences.” *Behavior Research Methods* 50(3): 1166–86. doi:10.3758/s13428-017-0935-1.

Van der Stigchel, S., C. Strauch, B. de Zwart, and L. Van Maanen. 2026. “Will Online Behavioral Research Follow the Fate of Online Survey Research?” *Proceedings of the National Academy of Sciences* 123(8): e2535585123. doi:10.1073/pnas.2535585123.