Statistical Power in Response Signal Paradigm Experiments

Pavel Logačev (pavel.logacev@boun.edu.tr)
Boğaziçi University, Department of Linguistics
34342 Beşiktaş, İstanbul

M. İlteriş Bozkurt (ilteris.bozkurt@metu.edu.tr)
Middle East Technical University, Cognitive Science Department
06800 Çankaya, Ankara

Abstract

The speed-accuracy tradeoff (SAT) method has produced several prominent findings in sentence processing. While a substantial number of SAT studies have yielded statistical null-results regarding the degree to which certain factors influence the speed of sentence processing operations, the statistical power of the SAT paradigm is not known. As a result, it is not entirely clear how to interpret these findings. We addressed this problem by means of a simulation study in which we simulated SAT experiments for a range of known effect sizes in order to determine the statistical power in typical SAT experiments. We found that while SAT experiments appear to have quite satisfactory power to detect differences in asymptotic accuracy, that is not the case for speed-related parameters, especially for the multiple-response variant of the technique. We conclude that the failure to find an effect in speed-related parameters in SAT experiments may be less meaningful than previously thought.

Keywords: response signal paradigm; speed-accuracy tradeoff; statistical power; simulation

Introduction

Experimental paradigms for measuring the speed-accuracy tradeoffs in cognitive tasks have been essential in many areas of cognitive science (e.g., Heitz, 2014) because they address a potentially problematic aspect of typical reaction time (RT) tasks: Most cognitive tasks can be performed more accurately at the cost of lower speed, or faster at the expense of accuracy (e.g., Pachella, 1974). As a result, the average RT and accuracy obtained in such tasks reflect not only the processing speed and accuracy on that task, but also the participant's *response criterion*, i.e., the mechanism by which they determine that they have processed a stimulus to a sufficient degree to respond. *Speed-accuracy tradeoff functions (SATFs)*, which describe the increase in response accuracy over time, are unaffected by the participants' response criteria and thus offer a way to obtain uncontaminated estimates of the relative speed of mental processes separately from their ultimate probabilities of success.

Differences between the SATFs can yield important insights into the timing of cognitive processes and have been successfully used in areas as diverse as attention, vision, memory, and psycholinguistics. For example, McElree (2000) showed that in the resolution of filler-gap dependencies in sentences like (1), increased distance between the verb *'admired'* and the head noun of its subject (*'book'*) decreased the *probability of successful dependency resolution*, but had no effect on the *processing speed*. That is, while participants did quite well at correctly judging grammatical sentences like (1a) and their ungrammatical counterparts like (1a') after a sufficient amount of time, they did not perform as well with sentences like (1b) and (1b'). However, the speed of relative increase in accuracy was the same in both condition pairs. Because discriminating between grammatical and ungrammatical sentences in (1) arguably requires the retrieval of *book* from working memory, this finding suggests that in sentence comprehension, distance affects the probability of successful retrieval of dependents from memory, but not the speed of the retrieval process.

(1a/a') This was the *book* that the editor *admired/*amused*.

(1b/b') This was the *book* that the editor who the receptionist married *admired/*amused*.

This finding, as well as a number of other results in the SAT literature (e.g., Foraker & McElree, 2007; Martin & McElree, 2008; McElree, Foraker, & Dyer, 2003; Van Dyke & McElree, 2011, among many others) rest on the absence of a significant difference in speed-related parameters of the speed-accuracy tradeoff function. Because SAT data analysis typically involves numerical estimation of several parameters describing the SATF in each condition, followed by a variant of a *stepwise model selection procedure* (e.g., Thompson, 1978), it is unfortunately not clear how much statistical power such experiments have in order to find differences in processing speed (however, see Pankratz, Yadav, Smith, & Vasishth, 2021, who conducted a power analysis using data from Franck & Wagers, 2020, a published SAT study, and showed a lack of statistical power). In order to understand how to interpret such statistical null-results, we conducted a simulation study to determine the amount of statistical power in typical SAT experiments.

Importantly, although our simulations are meant to model SAT experiments typical for the area of sentence processing, our results will likely be relevant beyond this specific setting. We further make all code available online,[1] which allows anyone to re-run the power simulations presented here under different assumptions with a modest amount of effort.

The Response Signal Paradigm

A method commonly used to estimate SATFs is the *response-signal paradigm* (McElree, 1993; Reed, 1973). In this

[1] https://git.io/JsfPR

paradigm, participants see stimuli of different types and are asked to respond to them after varying amounts of time relative to the onset of the last phrase of the sentence. Typically, an auditory cue presented after a variable amount of time is used as cue to respond immediately, even if participants have not yet finished making a decision. In the so-called *single-response* variant (*SR-SAT*), participants are prompted to respond once per trial. In the *multiple-response* variant (*MR-SAT*), participants respond several times per trial, at different lags.

As in the McElree (2000) experiment, the experimental design needs to ensure that discrimination between different types of stimuli (such as *'acceptable'* and *'unacceptable'*) requires participants to deploy the cognitive process being studied. As a result, the difference between SATFs (at least partially) reflects the timing of the cognitive process involved – the number of trials on which the relevant process has terminated will increase with the passage of time, and so will accuracy for both types of stimuli. In keeping with *Signal Detection Theory (SDT)* terminology (Wickens, 2001), we will refer to the two types of stimuli as *signal* and *noise* without loss of generality.

In order to obtain estimates of sensitivity to stimulus type which is unaffected by response bias towards either response, the accuracy at each lag in each experimental condition pair is computed as the SDT sensitivity measure $d' = \Phi(\textit{hits}) - \Phi(\textit{false alarms})$, where Φ is the Gaussian distribution function, and *hits* and *false alarms* are the proportions of 'signal' responses in signal and noise conditions, respectively. The resulting d' values at each lag allow us to estimate the underlying SATF in each experimental condition, which is typically well-approximated by the negatively accelerated shifted exponential function in equation 1 (Dosher, 1979) which is illustrated in Figure 1. In it, λ (*asymptotic accuracy*) determines the highest attainable level of accuracy given unlimited processing time, while δ (*intercept*) and β (*rate*) jointly determine the processing speed: δ determines at which point the accuracy rises above chance, while β determines how quickly the SATF increases. The reciprocal of the rate β^{-1} can be interpreted as the time required for the function to reach approximately 63% of the asymptote, *once accuracy has departed from* 0. A joint measure of dynamics, $(\delta + \beta^{-1})$, can be interpreted as the time requires to reache 63% of the asymptote, *starting at 0* (e.g., Foraker, Cunnings, & Martin, 2018; Liu & Smith, 2009).

$$d'_t = \lambda \cdot \left(1 - e^{-\beta(t-\delta)}\right), \; for\, t > \delta;\; else\; d'_t = 0 \qquad (1)$$

Analysis of data from response signal paradigm experiments follows a variant of a stepwise model selection procedure sometimes referred to as a *hierarchical model testing scheme* (e.g., Foraker et al., 2018): The aim is to select the most parsimonious of a range of models of varying complexity according to several criteria. For two-condition experiments, there are 8 candidate models. The simplest model ($1\delta - 1\beta - 1\lambda$) posits a single intercept, rate, and asymptote for both experimental conditions. The most complex model ($2\delta - 2\beta - 2\lambda$) posits separate intercepts, rates, and asymptotes for both experimental conditions. Further models of intermediate complexity that are also considered in the process are $1\delta - 1\beta - 2\lambda$, $1\delta - 2\beta - 1\lambda$, $1\delta - 2\beta - 2\lambda$, $2\delta - 1\beta - 1\lambda$, $2\delta - 1\beta - 2\lambda$, and $2\delta - 2\beta - 1\lambda$. Each model corresponds to a particular pattern of differences between experimental conditions: Differences in asymptotes (λ) can be interpreted as differences in the success probability of the target process, while differences in rate and intercept (δ and β), jointly considered *the dynamics*, can be interpreted as differences in processing speed.

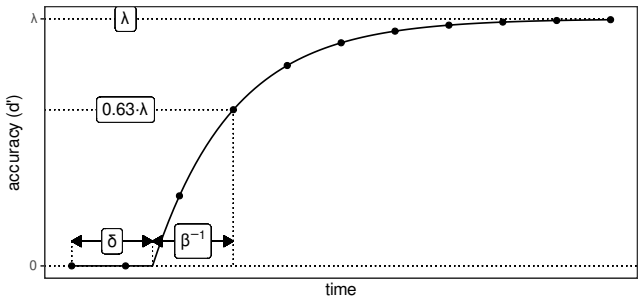

Figure 1: A speed-accuracy tradeoff following equation 1. After an initial period of chance performance, accuracy begins to increase at the intercept (δ). The growth rate (β) determines how quickly the function approaches the asymptotic performance (λ). The reciprocal of the rate β^{-1} can be interpeted as the time required for the function to reach approximately 63% of the asymptote.

The parameters for each such model are typically estimated separately for each participant by means of numerical optimization minimizing the *root mean squared error (RMSD)* of the model fit.

For inference, Foraker et al. (2018) recommend *forward model selection* starting with the asymptote parameter λ. As illustrated in Figure 2A, this process works by successively comparing nested models, beginning with the simplest model $1\delta - 1\beta - 1\lambda$ and $1\delta - 1\beta - 2\lambda$ in order to determine whether to assume one or two asymptotes. The choice made at this point affects which models will be considered later. If there is sufficient evidence for the two-asymptote model, this means that there is a difference in asymptotes between the two experimental conditions, and only 2λ models are considered at later stages; otherwise only 1λ models are considered. Similarly, the choice regarding the number of intercept parameters affects which models are considered at the third stage of the forward model selection procedure, when the number of rates is determined in the example in Figure 2A.

An alternative is *backward model selection*, illustrated in Figure 2B. This model selection process starts with the most complex model $2\delta - 2\beta - 2\lambda$, and involves comparing it to increasingly less complex nested models. If no evidence is found for the more complex model, the simpler model is adopted. During the stepwise selection procedure, models are compared based on (i) *adjusted* R^2, which takes into account the fit and

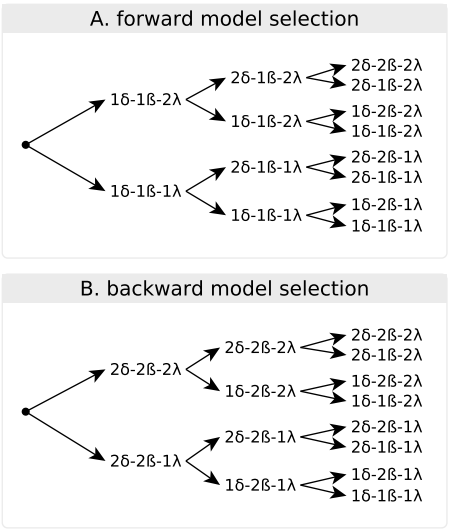

Figure 2: Illustration of the forward and backward model selection processes, in which the number of asymptotes is determined first, then the number of intercepts, and then the number of rates.

penalizes the number of parameters, and (ii) the consistency of the direction of the difference of the parameter estimates in the more complex models, as assessed by a statistical test.

A third alternative we will consider is non-stepwise inference based on the estimates of a $2\delta-2\beta-2\lambda$ model. That is, we will assume that the two experimental conditions differ in a parameter if we observe a significant difference between the parameter estimates, in no particular order.

Simulation Study

In order to determine the statistical power of the model selection methods outlined above with different sample sizes, we repeatedly simulated data for a number of different effect sizes for asymptotic accuracy and dynamics, performed model selection, and calculated the proportion of cases in which a difference with the correct sign was identified.

Method

We repeatedly simulated SR-SAT and MR-SAT experiments with two experimental condition pairs (two signal conditions, and two noise conditions). Figure 3 provides an overview of the process. We used parameter values which are relatively typical for sentence processing SAT experiments and assumed that the average population SATF followed equation 1 with $\delta = 0.8\,sec$, $\beta^{-1} = 0.8\,sec$, $\lambda = 2.25$. We further assumed that the individual SATF parameters for each subject s (λ_S, β_S^{-1}, δ_S) were log-normally distributed around the population parameters ($SD_\delta = 0.8$, $SD_{\beta^{-1}} = 0.8$, $SD_\lambda = 1.25$). We simulated data for a range of differences between conditions in the three SATF population parameters.

For each combination of parameters, we simulated a participant pool of $2,000$ participants with responses at 17 lags starting from $0\,sec$ to $5.6\,sec$, increasing in steps of $0.35\,sec$. We simulated different numbers of responses per: 20, 50, and 80 in each of the four experimental conditions for each of the 17 lags. To simulate experiment replications, we then drew $1,000$ bootstrap samples of 10, 20, 30, 40, or 50 participants and carried out model comparison for each bootstrap sample.

In simulating responses, we assumed that $P(R_{t,q} = 0)$, i.e. the probability of a 'noise' response in condition q at time t is given by equation 2. We assumed that responses were equi-biased towards 'signal' and 'noise' responses, and that the response criterion c at time t was always $c_t = d_t/2$.[2]

$$P(R_{t,q} = 0) = \Phi(c_t - \psi_{t,q}), \text{ where}$$
$$\psi_{t,q} = \begin{cases} d_t & \text{in signal trials} \\ 0 & \text{in noise trials} \end{cases} \tag{2}$$

Because in MR-SAT experiments, participants respond several times per trial, it stands to reason that the responses within the trial are correlated as the amount of evidence in favor of a particular response at lag k would at least partially depend on the amount of evidence available in its favor at lag $k-1$. We modeled this serial dependence based on assumptions akin to a simple random walk model in which evidence adds up in accordance with the increase in d' since the last lag: We assumed that ψ'_k, the amount of evidence in favor of a 'signal' response on a particular trial at lag k was the sum of ψ_k, the average amount of evidence in its favor at this lag, and normally distributed serially correlated noise with mean 0 and $\sigma = 1$, as in equation 3. We further assumed that a 'signal' response was given when $\psi'_k > c_k$, and a 'noise' response otherwise, thus accounting for serial correlation between responses.

$$\psi'_k = \psi_k + \sum_{i=1}^{k} \varepsilon_i/\sqrt{k}$$
$$\text{where } \varepsilon_i \sim N(0,1) \tag{3}$$

Analysis

We used *R* (R Core Team, 2018) and the *tidyverse* packages (Wickham, 2017) for simulation and data pre-processing, and the *mrsat* R package (Van Dyke, Wagers, Cho, & Matsuki, 2015) to fit eight models of varying degrees of complexity to each simulated participant's data.

We used five different model selection procedures on each of the simulated experiments: As a baseline analysis method, we tested for significant differences between estimates of $\hat{\delta}$, $\hat{\beta}$, $\hat{\lambda}$ in the two conditions based on $2\delta-2\beta-2\lambda$ model estimates. Moreover, we carried out backward and forward model selection using two sets of criteria: The first method was model selection based on the results of a t-test on the parameter difference estimates ($\hat{\Delta\delta}$, $\hat{\Delta\beta}$, $\hat{\Delta\lambda}$) of the more complex model.

[2]This assumption was made in order to obtain best-case power estimates, as a bias towards either response would increase the variance of the sampling distribution of d'. (cf. Liu & Smith, 2009, for an approximation of the variance of the d' sampling distribution.)

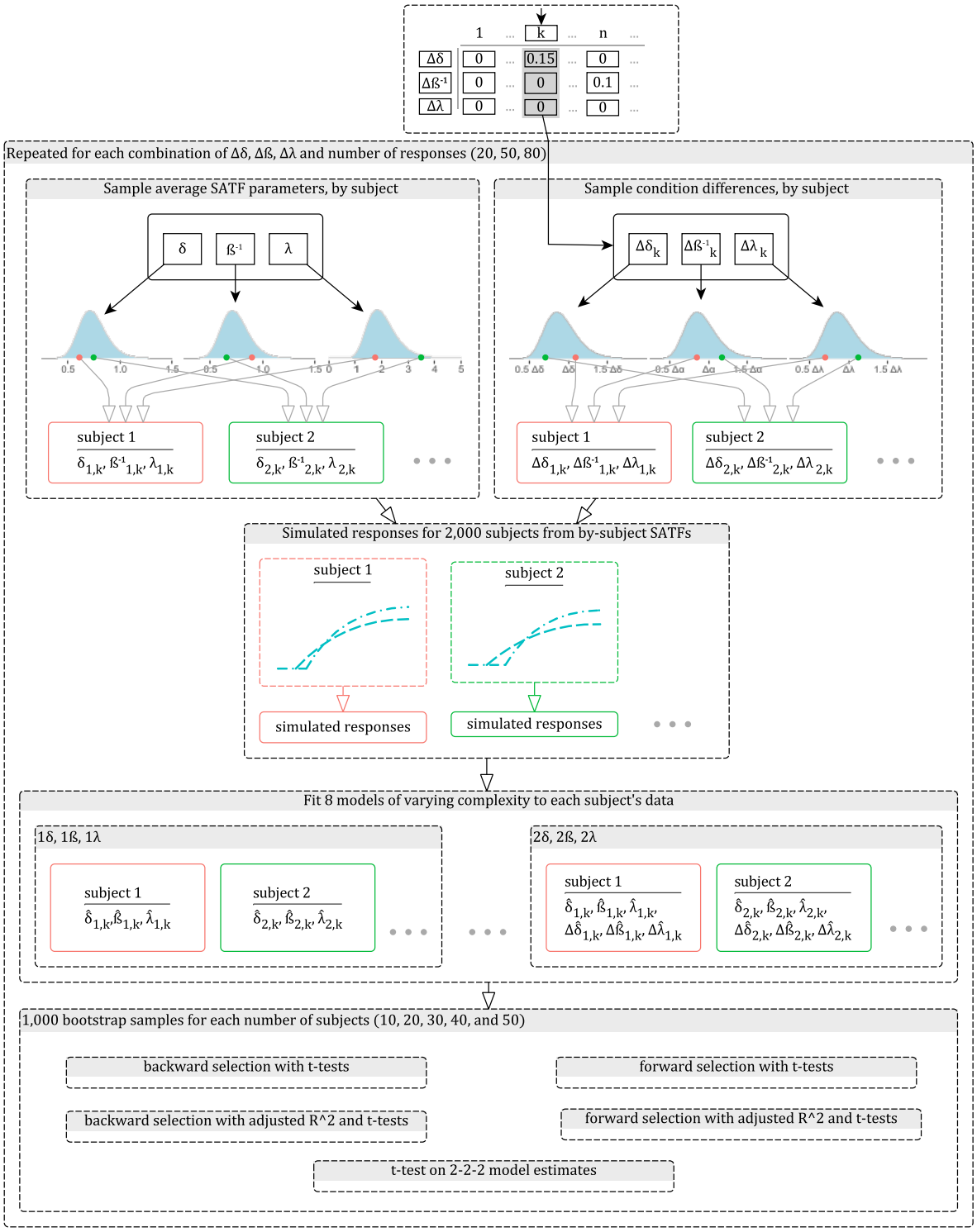

Figure 3: Illustration of the simulation process: For each parameter combination for $\Delta\delta$, $\Delta\beta^{-1}$, $\Delta\lambda$, we simulated out 2,000 participants. For each participant, we independently sampled the by-participant parameters for δ, β^{-1}, λ, $\Delta\delta$, $\Delta\beta^{-1}$, and $\Delta\lambda$ from their respective distributions. We then sampled 'signal'/'noise' responses for each of the 17 lags in each of the 4 conditions, computed d' at each lag, for each condition pair for each simulated participant, fitted several models of varying complexity to each participants data. Next, we drew 1,000 bootstrap samples of 10, 20, 30, 40, or 50 participants each (with replacement) and carried out model comparison for each bootstrap samples. Simulations for SR-SAT and MR-SAT were carried out independently.

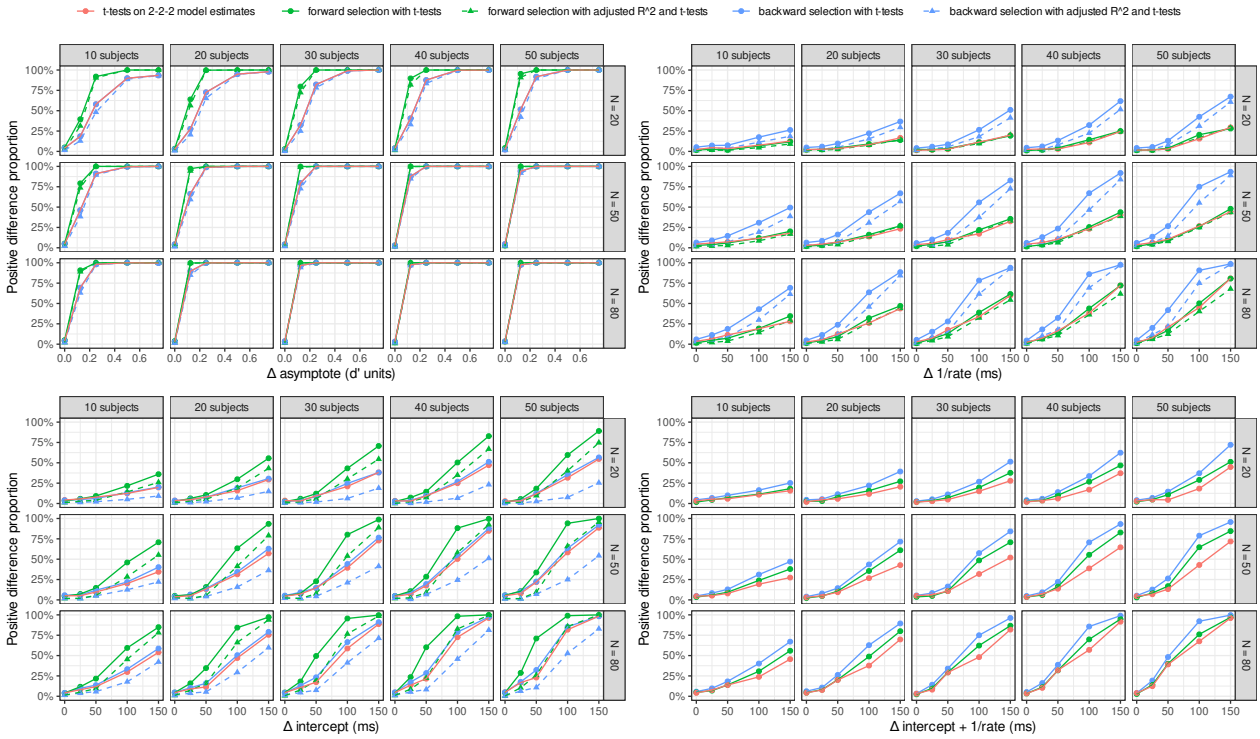

Figure 4: Simulation results for *single-response SAT*. Each panel shows the simulation results for one of the three SATF parameters with a separate panel for the pooled dynamics ($\delta + \beta^{-1}$). Each cell shows the power curve estimates of a particular combination of participants (columns) and number of resposes per condition (rows). In each cell, the x-axis represents the magnitude of the population difference, while the y-axis corresponds to the proportion of simulated experiments in which a positive value for the respective difference was (correctly) detected.

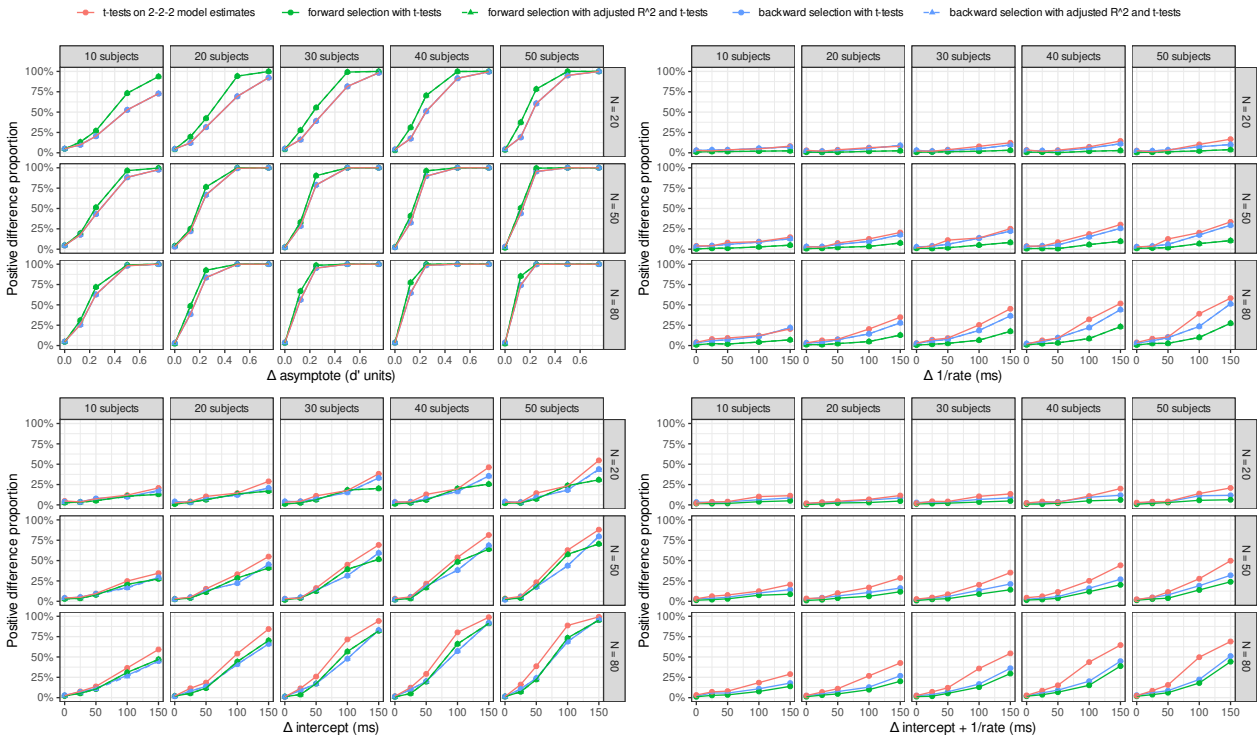

Figure 5: Simulation results for *multiple-response SAT*. Layout as in figure 4.

For example, in forward model selection as illustrated in fig. 2A, we selected model $1\delta - 1\beta - 2\lambda$ if the estimates of $\Delta\lambda$ significantly differed from 0, or in other words: if there was a significant difference in the asymptote estimates of the two experimental conditions. Otherwise, we selected the simpler model $1\delta - 1\beta - 1\lambda$. Because Foraker et al. (2018) suggest the use of the *adjusted* R^2 metric to supplement hypothesis tests, we also tested a more conservative method, in which the more complex model was only selected if its average *adjusted* R^2 across participants was higher than that of the simpler model, in addition to the relevant parameter difference being significantly different from 0.

For the analysis of pooled dynamics $(\delta + \beta^{-1})$, we considered only $2\delta - 2\beta - 2\lambda$, $1\delta - 1\beta - 2\lambda$, $2\delta - 2\beta - 1\lambda$ and $1\delta - 1\beta - 1\lambda$ models, and based inference on the results of t-tests on the asymptote and the joint dynamics measure.

Results and Discussion

Figures 4 and 5 show the results of our simulations. In each simulation, we calculated the proportion of samples in which a positive effect was detected, because we were interested in obtaining estimates of the probability that a difference with the *correct* sign is detected. As a result, the figures show estimates of *(Type I error rate)/2* when there was no difference between conditions, and *(Power - Type M error rate)* otherwise (Gelman & Carlin, 2014). We will refer to this quantity as the detection rate.

Importantly, since the joint dynamics measure is the sum of the intercept and the reciprocal of the rate, we could vary either to manipulate it. As a result, several detection rates are available for each value of the joint dynamics measure, depending on the contribution of the two underlying parameters. We present the maximum detection rate at each point in order to obtain optimistic detection rate estimates.

SR-SAT

The results for SR-SAT show a high overall probability of detecting a difference in asymptotes: It is reasonably high ($> 80\%$) for differences of more than 0.6 d' units, even for small sample sizes, except for the combination 10 subjects, 20 responses. Detection rates for the intercept parameter was rather low ($< 50\%$) for many sample sizes. Detection rates were even lower for the rate parameter, and reached more than 80% only for relatively large effect sizes, and only when the sample size was high.

Interestingly, while the forward model selection methods appeared to perform better than their alternatives for the asymptote and intercept parameters, backward model selection showed the highest detection rates for the rate parameter. This discrepancy is likely due to the fact that the rate parameter is selected last in both methods. As a result, backwards model selection is less likely to attribute differences between conditions to the asymptote or to the intercept as long as the models also assume two rates. When it does not, any substantial differences between conditions have to be attributed to the rate. As a result, backwards selection shows relatively low detection rates for asymptote and intercept differences, and relatively high detection rates for rate differences.

Somewhat surprisingly, the detection rates for the pooled dynamics metric $(\delta + \beta^{-1})$ did not show an improvement over detection rates for the intercept alone.

MR-SAT

While the MR-SAT findings regarding the model selection method were qualitatively similar to the SR-SAT results for asymptotes and rates, the model selection method appears to matter less than for SR-SAT data. Importantly, the MR-SAT detection rates were substantially lower across the board. This is not surprising given the fact that the responses on each trial were correlated, which is expected to increase the width of the sampling distribution and to decrease the *effective sample size* (e.g., Berger, Bayarri, & Pericchi, 2014). As a result, SR-SAT and MR-SAT seem to exhibit substantial differences in terms of power in detecting a difference in rates and intercepts, with remarkably low power especially for rate parameter differences of less than 50% even for relatively large effect sizes such as $150ms$, and less than 25% in most cases.

Surprisingly, the detection rates for the pooled dynamics metric $(\delta + \beta^{-1})$ were even lower than for the intercept parameter, which was likely due to a wide sampling distribution of the rate parameter due to a lot of uncertainty. This uncertainty would increase the variance of the dynamics estimate, leading to lower detection rates.

A possible concern about our findings for MR-SAT is that the exact detection rate estimates are potentially heavily dependent on the assumptions of our ad-hoc model accounting for the serial correlation between responses on a given trial. While that is correct, our results show that statistical power for MR-SAT *may* be substantially lower than for SR-SAT, due to a small effective sample size. The fact that we don't know *how much lower* it is, presents a major challenge in the interpretation of results from the MR-SAT methodology.

Conclusions

Our findings suggest that the statistical power for dynamics-related parameters in typical MR-SAT experiments (20 participants, with 50 responses per time lag per condition) may be relatively low for typical effect sizes in psycholinguistics $(50 - 100ms)$. This is due to the fact that as a result of the serial correlation of responses on a trial, the effective sample size may be significantly lower than the nominal sample size. As a result, the failure to find a difference in dynamics between two experimental condition pairs should be interpreted with caution. We hope that the power can be improved and its dependence on the model selection mechanism may be remedied by simultaneous estimation of all three SATF parameters using hierarchical models such as used by Niklaus, Singmann, and Oberauer (2019) and Pankratz et al. (2021).

References

Berger, J., Bayarri, M., & Pericchi, L. (2014). The effective sample size. *Econometric Reviews*, *33*(1-4), 197–217.

Dosher, B. A. (1979). Empirical approaches to information processing: Speed-accuracy tradeoff functions or reaction time—a reply. *Acta Psychologica*, *43*(5), 347–359.

Foraker, S. M., Cunnings, I., & Martin, A. E. (2018, Dec). *Speed-accuracy tradeoff modeling and its interface with experimental syntax.* PsyArXiv. Retrieved from psyarxiv.com/8zpaj doi: 10.31234/osf.io/8zpaj

Foraker, S. M., & McElree, B. (2007). The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, *56*(3), 357–383.

Franck, J., & Wagers, M. (2020). Hierarchical structure and memory mechanisms in agreement attraction. *Plos one*, *15*(5), e0232163.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 150.

Liu, C. C., & Smith, P. L. (2009). Comparing time-accuracy curves: Beyond goodness-of-fit measures. *Psychonomic Bulletin & Review*, *16*(1), 190–203.

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*(3), 879–906.

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language*, *32*(4), 536–571.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.

McElree, B., Foraker, S. M., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.

Niklaus, M., Singmann, H., & Oberauer, K. (2019). Two distinct mechanisms of selection in working memory: Additive last-item and retro-cue benefits. *Cognition*, *183*, 282–302.

Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In *Human information processing: Tutorials in performance and cognition* (p. 95). Hillsdale, N.J: Erlbaum.

Pankratz, E., Yadav, H., Smith, G., & Vasishth, S. (2021, Feb). *Statistical properties of the speed-accuracy trade-off (sat) paradigm in sentence processing.* PsyArXiv. Retrieved from psyarxiv.com/puqkv doi: 10.31234/osf.io/puqkv

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*(4099), 574–576.

Thompson, M. L. (1978). Selection of variables in multiple regression: Part i. a review and evaluation. *International Statistical Review/Revue Internationale de Statistique*, 1–19.

Van Dyke, J., Wagers, M., Cho, P., & Matsuki, K. (2015). *mrsat: Multiple Response Speed-Accuracy Tradeoff.* Retrieved from https://github.com/matsukik/mrsat

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. Oxford University Press.

Wickham, H. (2017). tidyverse: Easily install and load the 'tidyverse' [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tidyverse (R package version 1.2.1)