Statistical properties of the speed-accuracy trade-off (SAT) paradigm in sentence processing

Elizabeth Pankratz (pankratz1@uni-potsdam.de)
Department of Linguistics, Universität Potsdam
14476 Potsdam, Germany

Himanshu Yadav (hyadav@uni-potsdam.de)
Department of Linguistics, Universität Potsdam
14476 Potsdam, Germany

Garrett Smith (gasmith@uni-potsdam.de)
Department of Linguistics, Universität Potsdam
14476 Potsdam, Germany

Shravan Vasishth (vasishth@uni-potsdam.de)
Department of Linguistics, Universität Potsdam
14476 Potsdam, Germany

Abstract

Studies of the speed-accuracy trade-off (SAT) have been influential in arguing for the direct-access model of retrieval in sentence processing. The direct-access model assumes that long-distance dependencies rely on a content-addressable search for the correct representation in memory. Here, we address two important weaknesses in the statistical methods standardly used for analysing SAT data. First, these methods are based on non-hierarchical modelling. We show how a hierarchical model can be fit to SAT data, and we test parameter recovery in this more conservative model. The parameters most relevant to the direct-access account cannot be accurately estimated, and we attribute this to the standard SAT model being overparameterised for the limited data available to fit it. Second, the power properties of SAT studies are unknown. We conduct a power analysis and show that inferences from null results to the null hypothesis, though commonplace in the SAT literature, may be unwarranted.

Keywords: speed-accuracy trade-off; sentence processing; Bayesian modelling; power analysis

Introduction

The speed-accuracy trade-off (SAT) captures the idea that, the faster one is forced to make a decision, the less accurate one's response will probably be (Heitz, 2014). In the sentence processing literature, SAT studies have often been used to argue for a particular model of retrieval processes, namely the so-called direct-access model, proposed by McElree (2000). The direct-access model claims that representations of previously-encountered words are directly retrieved from working memory based on semantic and syntactic cues, rather than by, e.g., a serial search through all representations that concludes when the best match is found. A serial search would mean that more complex sentences take longer to be processed than simpler ones, while the direct-access model predicts no such difference. SAT studies have frequently been used to study this question (e.g., Foraker & McElree, 2011; Franck & Wagers, 2020; McElree, 2000).

In this paper, we will discuss two critical issues with the statistical methods standardly used for analysing SAT data. First, the usual method is based on non-hierarchical modelling, which fails to take all sources of variation into account. Second and more importantly, the statistical power of SAT designs has not yet been investigated (though see Logačev & Bozkurt, 2021). The latter issue is potentially a very serious one, since the arguments for the direct-access model depend on finding null results, but when statistical power is low, null results do not provide any information about whether the effect is actually absent (Greenland et al., 2016; Hoenig & Heisey, 2001). If the statistical power of SAT studies is found to be low for the numbers of participants that are realistic to test, then this paradigm should potentially not be used to provide evidence for the direct-access theory.

To address the first issue, we use data from a published SAT study (Franck & Wagers, 2020, Experiment 2) to fit a hierarchical Bayesian model using the probabilistic programming language Stan (Carpenter et al., 2017). The hierarchical model not only characterizes the sources of variance in the data more accurately than the standard analysis, but it also allows us to quantify the uncertainty associated with each parameter to be estimated, improving interpretability of the results (Liu & Smith, 2009). And by working with data that contains known effect sizes, we can assess the model's ability to recover the specific parameters that are critical for the direct-access account.

To address the second issue, we conduct a power analysis of the SAT design, again using the data from Franck and Wagers (2020). The power analysis will show a lack of statistical power at the sample sizes that are feasible to test in a design as demanding as the SAT for the effect sizes observed in that study.

Before turning to these issues, we first provide a backdrop for our work by exploring how the SAT has conventionally been modelled.

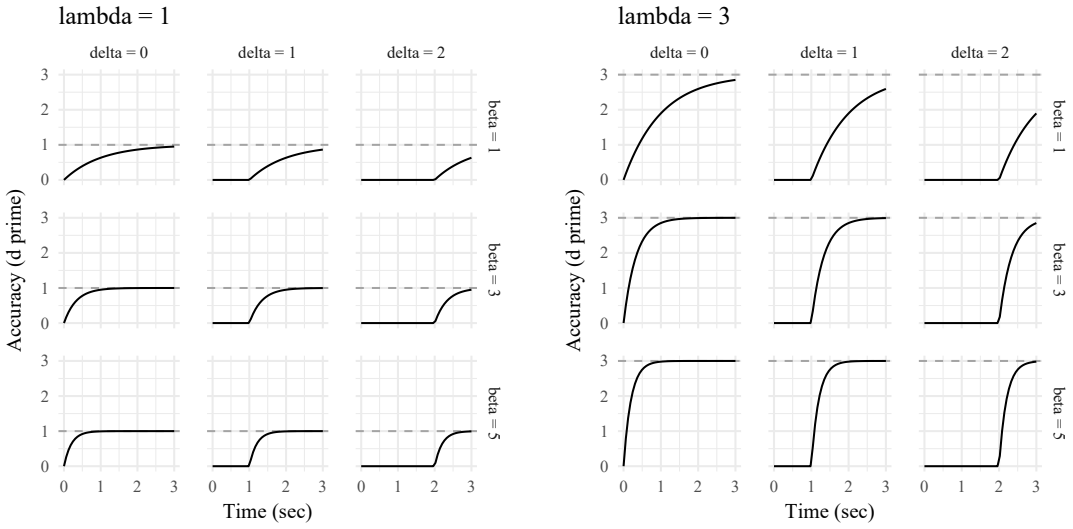

Figure 1: Parameter combinations for the SAT function

Modelling the SAT

Response accuracy can be thought of as a function of time. Imagine a task in which somebody reads a sentence and must provide an acceptability judgment after finishing it. If that person is prompted to give their judgment immediately after finishing the sentence, their decision of acceptable vs. unacceptable will be near chance, since they have not yet had enough time to process the sentence. However, if they are given more time before being prompted for a judgment, their decisions will, on average, become more accurate until eventually, the accuracy plateaus at some level.

This pattern is captured by the shifted exponential function given in Equation 1 (Wickelgren, 1977, p. 70). Accuracy is measured in d' units, a measure from signal processing.

$$d'(t) = \begin{cases} \lambda(1 - e^{-\beta(t-\delta)}) & \text{for } t > \delta \\ 0 & \text{for } t \leq \delta \end{cases} \quad (1)$$

The three parameters in this function—λ, β, and δ, henceforth the SAT parameters—admit the following interpretations: λ is the limiting value of $d'(t)$ as t grows without bound; β is the rate at which the function rises toward this asymptote; and δ is the function's point of departure from $d' = 0$, the accuracy level that represents chance. Figure 1 illustrates this function for a range of parameter combinations.

Researchers use SAT tasks like the hypothetical one described above to study how long it takes to process sentences that range in complexity. Take, for instance, the following examples from the foundational text by McElree (2000).

(1) This was the book that the editor admired __.

(2) This was the book that the editor who the receptionist married admired __.

(3) This was the book that the editor who the receptionist who quit married admired __.

These sentences contain a long-distance dependency between *book* and the gap following *admired*. While processing these sentences, the reader keeps the representation of *book* in their working memory until the gap is encountered, at which point the dependency can be completed (McElree, 2000, p. 113). The question of interest for the direct-access model is whether the time it takes to retrieve *book* differs between these three types of sentences, each of which contains progressively more material between the filler and the gap.

If people retrieve representations from working memory using a serial search mechanism rather than a direct-access one, then it should take longer to retrieve the correct filler *book* in conditions with more intervening material. In other words, (3) should take longer to process than (2), which should take longer than (1). In contrast, if the direct-access theory were true, then the processing speed in all three conditions should be the same, since we would access the representation of the filler directly. McElree (2000) introduces the SAT model as a way to answer this question.

Reframed in sentence processing terms, the three SAT parameters from Equation 1 receive more meaningful interpretations. λ is "a measure of the ultimate probability of correct retrieval, since additional time does not improve performance" (Franck & Wagers, 2020, p. 4). This may vary between sentences with different levels of complexity, since some sentences (like (3)) are simply harder to understand than others (like (1)).

Crucially, both β and δ concern the speed of processing. β is "the rate of information accrual" (Foraker & McElree, 2011, p. 768), or in other words, how quickly the sentence is understood starting from the time δ, the point "when information first becomes available" (Franck & Wagers, 2020, p. 4). Sometimes beta and delta are also combined into a single combined measure of processing dynamics. This type of

measure will be elaborated on below. So, to tease apart serial search vs. direct access, we can ask: do the parameters β and/or δ (or some combination) vary between conditions?

SAT experiments are designed to try to answer this question. Subjects are presented with sentences like the ones in (1)–(3) word by word (or phrase by phrase), and then they are prompted with an audio signal to give acceptability judgments at different time points following completion of the sentence. As one might imagine, this is a very demanding task, requiring extensive training and multiple lengthy experimental sessions (see, e.g., Franck & Wagers, 2020, p. 16). As such, the number of subjects is generally kept quite small.

The data gathered from an SAT experiment is processed as follows. At every time point t, each subject's response accuracy $d'(t)$ for each condition is computed as the Z-score of their true positive rate at t minus the Z-score of their false positive rate at t (Foraker & McElree, 2011, p. 768; Franck & Wagers, 2020, p. 16). This has the effect of aggregating all of the individual items into a single accuracy score for each condition at each time t. The result is that each subject has a vector of $d'(t)$ values for each condition, and the SAT function can be fit to these values.

An early method of analysing data using SAT curves was used by McElree (2000). This method uses model selection for hypothesis testing. It fits a number of combinations of curves to each subject's data, ranging from a "null model" (a single SAT curve with only one λ, one β, and one δ, abbreviated in the literature as 1λ-1β-1δ) to a "fully saturated" model (three different SAT curves—one per condition—yielding a total of three different λs, three βs, and three δs, abbreviated as 3λ-3β-3δ). Then, a goodness-of-fit measure (generally an adjusted R^2; Judd & McClelland, 1989) and a qualitative "evaluation of the consistency of the parameter estimates across subjects" (p. 117) are used to determine which model describes the data best. Liu and Smith (2009) illustrate how information criteria (AIC and BIC) can also be used for model selection. If the best model based on these selection criteria contains only one β and one δ, then the result is interpreted in support of the direct-access model.

A more recent approach used by, e.g., Franck and Wagers (2020) uses parameter estimation rather than model selection to test hypotheses. This method fits a fully-saturated model to each subject's data, or in other words, it fits one SAT curve per subject per condition. This results in a vector of estimates for each SAT parameter. Then, for each of these parameters a linear model is fit, predicting parameter estimates by condition. If there is no significant difference in the values of β and/or δ between conditions, then the data is again interpreted in support of direct access. (We conduct the power analysis below using this more recent method, since no qualitative judgment is involved in its interpretation.)

Generally, the finding from SAT analyses like these is that there is no evidence for a difference in β and δ between conditions (see Foraker & McElree, 2011 for a review), thus supporting the direct-access model.

In both the model selection and parameter estimation approaches mentioned above, the data from each subject is modelled separately. However, assuming that the tested subjects belong to some population, that population will contain some average effect that is modulated by the individual differences in each subject. We argue that the aim should be to model the general effect in the population, and as already observed by Liu and Smith (2009), hierarchical models offer a way to do this. Part of the current paper's contribution is a novel illustration of how a hierarchical Bayesian model can be fit to SAT data.

A hierarchical SAT model

Before detailing our model, we briefly discuss the data that it was fit on. This data comes from a published SAT study, Experiment 2 from Franck and Wagers (2020), which studies agreement attraction, a phenomenon by which a verb's agreement is incorrectly "attracted" by some element that is not its subject. For instance, in *The label on the bottles are rusty*, the verb agrees with the plural attractor *bottles* rather than the singular subject *label* (example from Franck & Wagers, 2020, p. 1).

In Franck and Wagers' Experiment 2, subjects read sentences in chunks of a few words at a time, each displayed for pre-determined time intervals. Subjects were then prompted to decide whether a given probe word appeared in the sentence they had just read. The prompts occurred at 18 different time intervals after the sentence was finished, ranging from 250 to 6000 milliseconds. This was a multiple-response SAT experiment, so subjects responded multiple times per trial.

The materials in this experiment follow a complex design that crosses three variables: probe type (with two levels, target and distractor), the probe word's syntactic position in the sentence (subject, high in the object NP, low in the object NP), and structure (whether the sentence contains a fronted direct object or a nested PP modifier). For simplicity's sake, we focus here only on the structure variable. Example sentences for the object level and the modifier level are shown in (4) and (5) respectively. (The materials are in French, and they contain jabberwocky nouns so that any observed effects would be due to the sentences' syntactic structure, rather than their semantics; Franck & Wagers, 2020, p. 7.)

(4) Quel dafran du brapou dis-tu que les bostrons défendent?
Which dafran of.the brapou do.you.say that the bostrons defend

(5) Les dafrans du brapou du bostron dorment.
The dafrans of.the brapou of.the bostron sleep

The hierarchical model we construct predicts whether the three SAT parameters vary as a function of the structure variable. We use R (R Core Team, 2020) and the package RStan (Stan Development Team, 2020) as an interface to Stan (Carpenter et al., 2017).

Since the sampling distribution of d' can be approximated by a normal distribution when the sample size is large (Liu

& Smith, 2009, p. 191), our model predicts d' values based on a Gaussian likelihood. The mean of the likelihood is an adapted version of the exponential approach to the limit function given in Equation 1 (see the Appendix for details).

Each of the three SAT parameters is computed from a linear function that predicts their value based on the sentence's structure. Importantly, these linear models all contain by-subject adjustments to slope and intercept, allowing individual subjects to differ in their processing speed and ultimate response accuracy. By including by-subject adjustments to the estimates of the slope and intercept parameters, we allow the model to focus on the general effects within the population, incorporating but generalising over variation in individual subjects (Pinheiro & Bates, 2000).

At this point, the reader may be wondering: why include adjustments by subject but not by item, since in principle, item-level variability should play a similar role? It should, but as mentioned above, we cannot include adjustments for each item because SAT data is pre-aggregated by item in order to obtain the values for d'.

The likelihood and its components in the hierarchical model are listed in Equation 2 (the priors, which are not shown, are weakly regularising). In the model definition, i indexes the i^{th} data point; the α and γ parameters are the intercept and slope, respectively, of the linear functions predicting the three SAT parameters; the u parameters represent the adjustments to the intercepts and slopes for each subject $subj[i]$; and $structure_i$ is the ± 0.5 contrast coding of the structure variable.

$$
\begin{aligned}
\lambda_i &= (\alpha_\lambda + u_{\alpha_{\lambda_{subj[i]}}}) + (\gamma_\lambda + u_{\gamma_{\lambda_{subj[i]}}}) \cdot structure_i \\
\beta_i &= (\alpha_\beta + u_{\alpha_{\beta_{subj[i]}}}) + (\gamma_\beta + u_{\gamma_{\beta_{subj[i]}}}) \cdot structure_i \\
\delta_i &= (\alpha_\delta + u_{\alpha_{\delta_{subj[i]}}}) + (\gamma_\delta + u_{\gamma_{\delta_{subj[i]}}}) \cdot structure_i \\
\mu_i &= \lambda_i (1 - e^{-\beta_i (t_i - \delta_i)}) \\
d'_i &\sim \text{Normal}(\mu_i, \sigma) \text{ for } d'_i > 0
\end{aligned}
\tag{2}
$$

In addition to easy incorporation of the data's hierarchical structure, modelling within the Bayesian framework also allows us to quantify uncertainty about the parameters of interest. We want to know about the effect of condition on the SAT parameters, so the parameters of interest are the slope parameters of the three linear functions: γ_λ, γ_β, and γ_δ (and in particular the latter two, since these reflect the condition's effect on processing speed). Ideally, the posterior distributions of these parameters should tell us whether there is a difference between conditions for any of the three SAT parameters.

However, in a test of parameter recovery, the model was able to accurately recover γ_λ on data simulated based on the effect sizes in Franck and Wagers' experimental data, but it consistently failed to recover both γ_β and γ_δ. The fact that it could not recover those two parameters, while succeeding for γ_λ, is not entirely surprising; there are principled reasons why this might be the case.

One explanation could be that the model receives more information about the value of λ than about the values of β and δ. Only a few data points near the beginning of the curve—those closest to the point of departure from $d' = 0$, the true value of δ—will tell us much about where δ is located. There will also only be a handful of points in between $d' = 0$ and $d' = \lambda$ that are informative about β, the rate of the function's approach to the asymptote. In contrast, many data points will be located near the asymptote λ, making the model's job of figuring out λ much easier. And if β and δ cannot be identified with any certainty, then the estimates of how they differ between conditions, γ_β and γ_δ, will naturally also suffer.

Another explanation is based on the facts that both β and δ concern the speed of processing and that they tend to trade off: high βs often co-occur with small δs, and vice versa (Franck & Wagers, 2020). Similar to fitting a linear model with two collinear predictors, it may be that neither of these parameters can be accurately estimated because they both explain similar aspects of the data. The SAT model could well be over-parameterised, making these two parameters hard to estimate simultaneously.

Either way, the problems we observed in estimating β and δ are probably not exclusive to the hierarchical model presented here. The traditional method for fitting SAT curves, "an iterative hill-climbing algorithm (Reed, 1976) similar to STEPIT (Chandler, 1969)" (McElree, 1996), may suffer from the same difficulty in identifying β and δ. In light of this finding, parameter recovery in the traditional SAT curve-fitting method should be checked.

One way the literature attempts to circumvent issues with β and δ is to estimate them individually as described in the previous section, but to combine them for analysis into a single measure that reflects the overall dynamics of processing (Liu & Smith, 2009; Franck & Wagers, 2020). This dynamics measure is generally computed as shown in Equation 3.

$$
\delta + \frac{1}{\beta} \tag{3}
$$

This measure is motivated theoretically by the trade-off between β and δ, and it should allow the limited data available for each of those two parameters to be combined into a single measure. However, as we will see in the next section, this dynamics parameter tends to obscure true differences between conditions.

Power analysis of the standard SAT method

A second pressing concern for the SAT methodology is that statistical power may be low. As mentioned above, the inferential step often made in the literature involves interpreting a null result—i.e., no significant difference between the β and δ estimates for various conditions—as evidence for a hypothesis of no effect—i.e., that β and δ actually do not differ in reality. This step is problematic. A null result cannot be interpreted as evidence that the null hypothesis is true if statistical power is low (Greenland et al., 2016; Hoenig & Heisey,

2001). To determine whether the null inferences that are commonplace in the SAT literature are warranted, we investigate the power properties of a recent SAT experiment.

The power of a model is given by the proportion of times that that model produces a statistically significant result under repeated sampling when the null hypothesis does not hold (Vasishth, Mertzen, Jäger, & Gelman, 2018). In other words, it is the proportion of times that the null hypothesis is correctly rejected. When power is computed using simulation, it can be estimated by computing the proportion of times that the model produces a statistically significant result under repeated sampling when the value of the parameter of interest has a particular value different from the null hypothesis (Miller & Miller, 2004).

We can explore the power properties of an existing experiment design for which data already exist by assuming that the estimates from that experiment reflect the true values of the parameters in question (Gelman & Carlin, 2014). Then, we use these estimates to repeatedly generate simulated datasets with an arbitrary number of participants. The simulated data is analysed using the same tests as in the original experiment. The proportion of times that the relevant predictors are statistically significant is the estimated power of the method for a future study. For reproducible code illustrating this approach, see Vasishth et al. (2018, Appendix B).

Here, we base our simulated data on the estimates of λ, β, and δ from Franck and Wagers' (2020) Experiment 2. Consequently, the power analysis presented here only holds for their analysis method of their data, assuming that the experiment will be conducted again (this is not a computation of post-hoc power).

As mentioned in the introduction, Franck and Wagers fit one SAT curve per subject per condition. The two levels of the probe type variable are combined to estimate d', yielding d' values for each participant in six conditions: the cross of the two-level structure variable and the three-level probe variable. For 25 subjects, this means 150 curves, yielding 150 estimates of each parameter.

To generate simulated data in each condition, we defined normal approximations of Franck & Wagers' observations of β and δ by taking the mean and standard deviation of these estimates in each condition, shown in Table 1.

The estimates of β and δ were negatively correlated ($r = -.32$), so we sampled simulated values for those two parameters from a multivariate normal distribution defined by the means and standard deviations from Table 1 and a correlation coefficient of $-.32$. Simulated data for λ was sampled from a univariate normal distribution. Then, to produce the combined dynamics parameter, we computed $\delta + \frac{1}{\beta}$ from the sampled values of β and δ in each iteration.

Although the β and δ distributions are similar to one another, they are not identical. This observation lines up with the speculation voiced by Nicenboim and Vasishth (2018, p. 5) that the reason why studies do not find evidence for differences in β and δ may simply be that "the differences were too small to be detected", not that the differences do not exist. At this point, the question may be raised of whether the differences are also too small to be theoretically meaningful. This is for direct-access researchers to decide; we will abstain from pre-judging what is theoretically meaningful and simply work with the sizes of effects that have actually been observed. Thus, we ask: how many subjects are required in order to reliably identify the differences in Franck and Wagers' data?

Condition		λ		β		δ	
Str.	Pr.	M	SD	M	SD	M	SD
Mod.	High	3.09	0.33	3.39	2.20	1.00	0.26
Mod.	Low	3.14	0.37	3.33	2.19	1.01	0.27
Mod.	Subj.	3.22	0.20	3.65	2.66	0.96	0.26
Obj.	High	2.93	0.46	3.76	2.99	0.98	0.25
Obj.	Low	2.69	0.68	3.54	3.22	0.94	0.25
Obj.	Subj.	3.22	0.27	3.77	2.76	0.94	0.23

Table 1: Means and standard deviations of the SAT parameter estimates in the six conditions of Franck and Wagers' Experiment 2 (the cross of structure and probe)

We generated data as described above for a range of subject counts from 5 to 2000, and then analysed the simulated data with the linear models used in Franck and Wagers' original analysis. These models predict each parameter's value as a function of the structure and probe conditions as well as their interaction, with varying intercepts by subject.

In order to compute the power under repeated sampling, we simulated data for each subject count 1000 times. Figure 2 presents the results. The shading of each cell indicates the proportion of significant outcomes under repeated sampling for each coefficient in the linear model. This proportion indicates the power of the model: the darker the shade, the higher the power.

The coefficients shown on the vertical axis can be understood as follows. The structure condition was sum-coded; it compares the object level to the modifier level. The probe condition was Helmert-coded. The subject level was compared to the average of the high- and low-probe levels, and then the high- and low-probe levels were compared to one another. The interaction of both Helmert comparisons with the structure condition was also computed. For a detailed discussion of contrast coding, see Schad, Vasishth, Hohenstein, and Kliegl (2020).

In their study, Franck and Wagers analysed λ and the combined dynamics parameter with data from 25 participants. For λ, they found significant effects of probe (subj./non-subj.) and structure, and significant interactions between structure and probe (subj./non-subj.) and between structure and probe (high/low). For dynamics, they also found a significant effect of probe (subj./non-subj.).

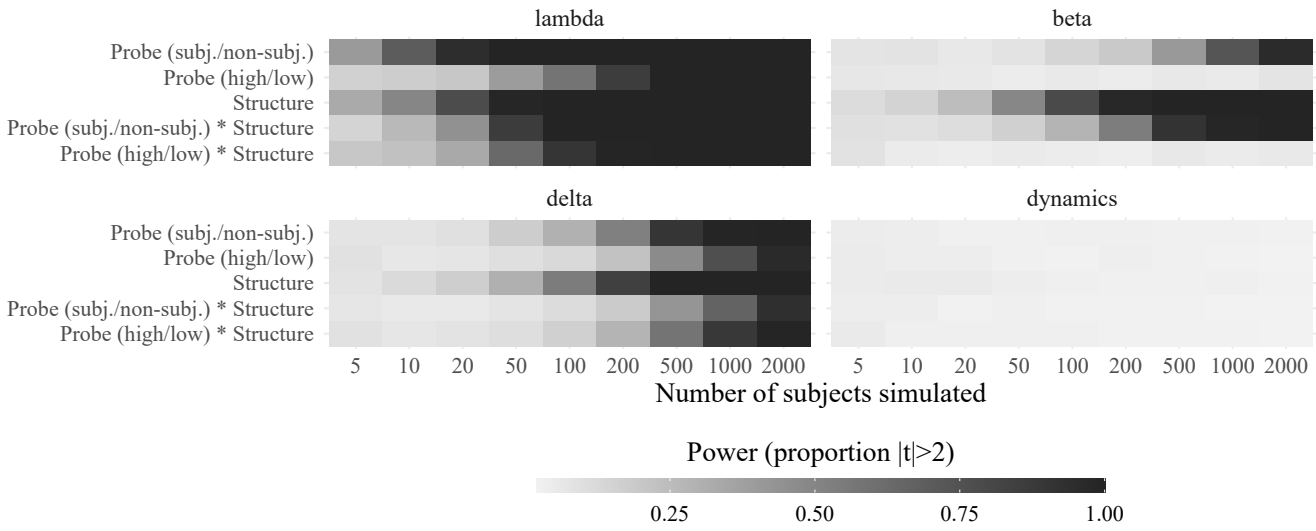

Figure 2: Proportion of significant ($|t| > 2$) results for each coefficient for a range of simulated subject counts

As Figure 2 shows, for the original parameters of interest β and δ, power only reaches an acceptable level (80% is the usual standard) when the subject count becomes very large. And, strikingly, the dynamics parameter computed from the covarying β and δ does not begin to approach the desired level within the range of subjects we simulated. Thus, subject counts that are feasible in real-life SAT studies are unlikely to identify true effects of the magnitude estimated by Franck and Wagers (2020) and will often return a null result, even when a true effect exists. In sum, this power analysis suggests that, for the effect sizes found by Franck and Wagers (2020), the standard inference to the null hypothesis based on null results is not warranted.

A reviewer noted that the effects of the dynamics predictor observed by Franck and Wagers are on the smaller end, and larger effects are possible, which should make the power situation less dire. This is true, but we would add that in underpowered studies, effects tend to look larger than they are (Gelman & Carlin, 2014), and our focus here has been on conservative effect sizes that may be more realistic than those reported in other studies.

Finally, we reiterate that this finding only holds for the Franck and Wagers (2020) study. Results could well be different in, e.g., a single-response SAT study in which subjects respond only once per trial. Because no single-response SAT datasets are presently available, we have not analysed this paradigm here, though see Logačev and Bozkurt (2021), who show using simulated data that single-response SAT has somewhat higher power than multiple-response SAT.

Conclusion and outlook

The goal of this paper was to highlight general issues with the SAT methodology on the basis of a case study using published SAT data. The paper's contributions are twofold.

First, we have argued that SAT data should be analysed using a hierarchical model, and we have shown how such a model can be implemented in Stan. This model calls into question the assumption that β and δ parameters can be accurately estimated with the amount of data commonly available. Second, we have shown that the standard method for analysing SAT data may not have enough statistical power to identify the effects that the direct-access model depends on.

We wish to emphasise that, although we use their data, our intention in this paper is not to single out and criticise the study by Franck and Wagers (2020). Quite the opposite: Franck and Wagers making their data publicly available is highly admirable. To our knowledge, theirs is the only published SAT dataset available at the time of writing, and we are grateful for their contribution, since it made the present study possible. We also wish to highlight that our critique of the SAT method only minorly affects the conclusions drawn in their paper, since their discussion primarily focuses on interpreting λ, and as we have shown, λ is the parameter that can be estimated most reliably.

Although the SAT method has generally been considered the de facto standard for evaluating the direct-access account, other types of data can also be used (for example, Lissón et al., 2021 use listening times, and Nicenboim & Vasishth, 2018 use reading times). Our results suggest that perhaps these alternative measures should be prioritised over the SAT paradigm when studying the direct-access model of retrieval processes.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1), 1–32.

Chandler, J. P. (1969). Subroutine STEPIT—Finds local minimum of a smooth function of several parameters. *Behavioral Science*, *14*, 81–82.

Foraker, S., & McElree, B. (2011). Comprehension of linguistic dependencies: Speed-accuracy tradeoff evidence for direct-access retrieval from memory: Comprehending dependencies. *Language and Linguistics Compass*, *5*(11), 764–783. doi: 10.1111/j.1749-818X.2011.00313.x

Franck, J., & Wagers, M. (2020). Hierarchical structure and memory mechanisms in agreement attraction. *PLOS ONE*, *15*(5), 1–33. doi: 10.1371/journal.pone.0232163

Gelman, A., & Carlin, J. (2014, November). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. doi: 10.1177/1745691614551642

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. doi: 10.1007/s10654-016-0149-3

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 1–19. doi: 10.3389/fnins.2014.00150

Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, *55*(1), 19–24. doi: 10.1198/000313001300339897

Hoffman, M. D., & Gelman, A. (2011). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv:1111.4246*.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. San Diego: Harcourt Brace Jovanovich.

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., van het Nederend, M. L., Burchert, F., … Vasishth, S. (2021). A Computational Evaluation of Two Models of Retrieval Processes in Sentence Processing in Aphasia. *Cognitive Science*, *45*(4), e12956. doi: 10.1111/cogs.12956

Liu, C. C., & Smith, P. L. (2009). Comparing time-accuracy curves: Beyond goodness-of-fit measures. *Psychonomic Bulletin & Review*, *16*(1), 190–203. doi: 10.3758/PBR.16.1.190

Logačev, P., & Bozkurt, M. İ. (2021). Statistical power in response signal paradigm experiments. *PsyArXiv Preprints*. doi: 10.31234/osf.io/st6de

McElree, B. (1996). Accessing short-term memory with semantic and phonological information: A time-course analysis. *Memory & Cognition*, *24*(2), 173–187. doi: 10.3758/BF03200879

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

Miller, I., & Miller, M. (2004). *John E. Freund's Mathematical Statistics with Applications*. Prentice Hall.

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34. doi: 10.1016/j.jml.2017.08.004

Pinheiro, J., & Bates, D. (2000). Linear mixed-effects models: Basic concepts and examples. In *Mixed-effects models in S and S-PLUS* (pp. 3–56). New York, NY: Springer New York. doi: 10.1007/0-387-22747-4

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria.

Reed, A. V. (1976). The time course of recognition in human memory. *Memory & Cognition*, *4*, 16–30.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. doi: 10.1016/j.jml.2019.104038

Stan Development Team. (2020). *RStan: The R interface to Stan*.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. doi: 10.1016/j.jml.2018.07.004

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85.

Appendix: Technical details

In Equation 1, the conditional statement depends on the comparison of t with δ, while Equation 2 computes the exponential function only when d' exceeds zero. This condition may seem circular or tautological—how can we use the value of d' to compute the value of d'?—but it can be understood by recognising that Stan is not actually computing any values for d', since d' is not a parameter in the model, but the outcome. Instead, Stan is using the observed d' values to compute the likelihood in each sampling iteration. Therefore, its values are known and may be used in a conditional statement like the one in Equation 2.

How does changing the conditional statement improve the sampling procedure? The sampling algorithm that Stan uses, NUTS (Hoffman & Gelman, 2011), works poorly in a model in which t_i is compared to δ_i. This is because, in each iteration, Stan might propose a different value for δ_i, and comparing t_i to this fluctuating threshold results in a posterior "landscape" that is too inconsistent for the sampler to traverse. However, when we compare d'_i to the constant value of zero, the sampler is able to work much more efficiently. Posterior predictive checks confirmed the equivalence of this new condition with the original one from Equation 1.