Supplemental Materials for: "A cognitive model-based approach to testing mechanistic explanations for neuropsychological decrements during tobacco abstinence"

Alexander Weigard[1], Cynthia Huang-Pollock[1], Andrew Heathcote[2], Larry Hawk Jr.[3] and Nicolas J. Schlienz[4]

[1]The Pennsylvania State University, [2]University of Tasmania, [3]The State University of New York at Buffalo, [4]Johns Hopkins Medicine

Correspondence concerning the article and materials presented here should be addressed to Alexander Weigard, Department of Psychology, Moore Building, Penn State University, University Park, PA, 16802. Email: asw5097@psu.edu

Estimation Procedures and Priors

Posterior distributions were estimated using the differential evolution method for Markov Chain Monto Carlo sampling to address the problem of parameter covariation (Turner, Sederberg, Brown, & Steyvers, 2013). Start points for the sampling of individual- and group-level parameter values were determined using means of posteriors from earlier fits at the single-subject level. Following previous work with the hierarchical LBA (Turner et al., 2013), relatively broad and uninformative priors were posited for both tasks. Group σ parameters were exponential distributions with a scale parameter of 1. Group μ priors were positive normal distributions, truncated at .1 for *t0* and 0 for all other parameters, with the following values:

$$A_\mu \sim TN(\mu=1,\sigma=0.5)$$
$$b_\mu \sim TN(\mu=1,\sigma=0.5)$$
$$v_\mu \sim TN(\mu=2,\sigma=1)$$
$$sv_\mu \sim TN(\mu=1,\sigma=1)$$
$$t0_\mu \sim TN(\mu=1,\sigma=0.5)$$

The number of chains used in sampling for each parameter was set at the number of parameters in the model times 3 (e.g., 54 chains for a model with 18 parameters). Samples were thinned by retaining only every 10^{th} iteration to address autocorrelation and burned in for 300-1100 iterations (depending on time to convergence) before running 400 for analysis. Visual inspection and the Gelman-Rubin statistic (Gelman & Rubin, 1992) both indicated that chains were stable and had converged prior to this 400-iteration period for all models (Gelman-Rubin <1.1 for all parameters). This procedure produced ≥10,800 posterior samples per parameter for analysis.

Plots of Absolute Model Fit

Although the model selection procedure described in the text addresses which of several

plausible models have the best *absolute fit*, it does not allow determination of whether or not the models explain key features of the behavioral data well (e.g., latency of RT quantiles for correct and error responses, differences in accuracy between conditions). To do so, model fit was displayed with cumulative distribution function plots that compared the cumulative probability of a correct or error response over time for both empirical data and posterior predictive data generated from the model. Plots for the two best-fitting models selected for analysis (Models A and B) are displayed in Supplemental Figure 1.

The plots indicate that the two best-fitting models both generally describe the behavioral data well, with the greatest misfits occurring for error trials in the congruent condition. This pattern is expected, as the likelihood function used to fit the model places little weight on these relatively uncommon trials compared to other trial types. Crucially, although misfit also occurred for errors in the incongruent condition, inspection of the RT quantiles for correct vs. error responses indicates that the models correctly predict faster error responses than correct responses for incongruent trials. This suggests that our description of the flanker manipulation with both *b* an *v* differences successfully accounted for the hallmark fast error effect described by White et al. (2011), and that the LBA models with the best relative fit also display adequate absolute fit to the behavioral data.

Calculation of *P* for Hypothesis Testing

To test main effects of a factor (e.g., Smoke Condition), individuals' posteriors at one level of the factor (e.g., abstinent) were averaged between levels of other factors (e.g., congruent an incongruent stimuli), and subtracted from the average of the individual's posteriors for the second level of the factor of interest (e.g., smoking). For interactions (e.g., Congruency x Smoke Condition), the effect of one factor was calculated at each level of the other factor to create

difference distributions, and these difference distributions were then subtracted from one another. In both cases, individual-level contrast distributions were averaged, and *P* was defined as the proportion of averaged samples above 0 (in the case that the majority of samples were above 0) or below 0 (in the case that the majority of samples were below 0). For main effects, *P* was calculated by sampling without replacement from the μ posterior distributions and counting the proportion of samples for which one condition was greater than the other.

Analysis of Behavioral Summary Statistics

Behavioral summary statistics (mean RT, the SD of RT, accuracy) were compared with both traditional null-hypothesis significance tests (*p*-values) and Bayes Factors (BFs) using the JASP software (all tests used standard priors: JASP Team, 2016)[1]. Summary statistics were entered into Repeated-Measures ANOVAs that included Smoke Condition (smoking, abstinent) and other relevant within-subjects factors for each task to obtain *p*-values and BFs for all main and interaction effects. BFs quantify the extent to which the data support the research hypothesis – in this case, that there are differences between conditions – (above 1) relative to the null hypothesis (below 1); a BF of 3, for example, indicates that the data are 3 times more likely (a 75% chance) to have occurred under the research hypothesis than under the null hypothesis assuming each hypothesis is equally likely a priori. Based on commonly-used guidelines (Kass & Raftery, 1995), BFs between 1 and 3 are thought to provide only "anecdotal" or ambiguous evidence for the research hypothesis, BFs of 3 to 20 provide positive evidence, and BFs >20:1 provide strong evidence.

There was a main effect of Congruency on accuracy, $F(1,24)=81.25, \eta^2 =.77, p<.001$,

[1] All Bayesian tests in JASP used standard priors, including the Cauchy prior for effect size (width = .707) for Bayesian t-tests, prior effect size scale = .5 for fixed factors in ANOVAs, and prior effect size scale = 1 for random factors in ANOVAs.

BF>10.000, in which congruent trials were more accurate. There was no main effect of Smoke Condition on accuracy, $F(1,24)=1.47, \eta^2 =.06, p=.237$, BF=.32 or Smoke Condition x Congruency interaction, $F(1,24)=3.71, \eta^2 =.13, p=.066$, BF=.52.

For mean RT, participants had longer RT on incongruent trials, $F(1,24)=237.81, \eta^2 =.91, p<.001$, BF>10,000, and longer RT during abstinence, $F(1,24)=12.14, \eta^2 =.34, p=.002$, BF=2613.49. Similarly, RTs were more variable on incongruent trials, $F(1,24)=81.09, \eta^2 =.77, p<.001$, BF>10,000, and during abstinence, $F(1,24)=12.36, \eta^2 =.34, p=.002$, BF=377.59. There were no significant interactions in RT metrics.

Plausible Values Analysis Methods

To estimate posterior distributions of Pearson's *r* for the relationship between changes in individual-level parameters and covariates, the posterior distribution for the *sample*'s correlation coefficient is first estimated by assessing the correlation between the covariate and each individual-level posterior sample. Following this, the approach outlined by Ly, Marsman and Wagenmakers (2015) is used to estimate posterior distributions for the *population*'s correlation coefficient. For calculation of the population posterior, a uniform prior was posited for the population's *r* spanning the values from -1 to 1. After these distributions were estimated, their density was plotted, and *P* was calculated to quantify the probability that *r* for the relationship was above 0.

To calculate changes in *svc* and *b* (the main model parameter changes of interest), each individual's posterior for *svc* and *b* (the latter being averaged between all accumulators) in the smoking condition was subtracted from their posterior for the same parameter in the abstinent condition. Thus, a positive relationship between a covariate and these change scores would indicate that an increase in the covariate predicts a larger increase in *svc* or *b* for a given

individual during withdrawal. Due to concerns about possible overfitting in Model A, results calculated using the posteriors from Model B are reported here. However, these results did not change substantially when posteriors from Model A were used instead.

Mean Drift Rate Effects in Model A

In Model A, *vc* was faster, $P>.99$, and *ve* was slower, $P>.99$, in the abstinent session. There was also evidence for Congruency x Smoke Condition interactions in both *vc*, $P=.97$, and *ve*, $P>.99$. For *vc*, the interaction appeared to be driven by a stronger effect of Smoke Condition in the incongruent condition, $P>.99$ (average $\Delta vc = .27$), than in the congruent condition, $P>.99$ (average $\Delta vc = .20$). For *ve*, the interaction was driven by the fact that there was strong evidence for a decrease in *ve* in the Abstinent session for incongruent stimuli, $P>.99$ (average $\Delta ve = .17$), but no substantial evidence for such an effect in the congruent condition $P=.74$ (average $\Delta ve = .01$).

Taken together, this pattern suggests that participants were able to more efficiently process relevant information (increased *vc*) and more efficiently parse relevant information from irrelevant information (decreased *ve*) when they were abstinent, although these effects appear to be limited to incongruent stimuli. These results were surprising, as it was not expected that participants' efficiency of processing would be *improved* during tobacco withdrawal. The unexpected nature of these results, combined with concerns about overfitting in Model A (detailed in the body of the manuscript), further suggests that the effects of Smoke Condition and Congruency x Smoke Condition interactions were spurious.

Simulation Studies on Selective Effects in *v* and *sv*

The model selection procedure and analysis of parameter differences reported in the main body of the manuscript provided evidence that the *sv* parameter, but not the *v* parameter,

explained abstinence effects on cognition in the experimental paradigm. However, effects in mean rate parameters from sequential sampling models are notoriously difficult to separate from effects in rate variance parameters (e.g., see: Donkin, Brown, & Heathcote, 2009; Ratcliff & Van Dongen, 2011; Voss, Nagler, & Lerche, 2013). To determine whether it was plausible that the current experimental paradigm and the LBA modeling method used were able to reliably distinguish abstinence effects in *v* from those in *sv*, several additional "model recovery" analyses were conducted using simulated data generated by the LBA model.

First, two separate data sets were simulated using functions in DMC, both of which matched key features of the empirical data set used in the current study: 25 subjects with 800 trials each, split evenly across Smoke (smoking/abstinent) and Congruency (congruent/incongruent) conditions. The first, *sv-effects*, data set was simulated using group μ and σ parameter estimates (the medians of the posterior samples for all parameters) drawn from the top model in the model selection analysis for which *sv* varied by abstinence, but *v* did not (Model B). The second, *v-effects*, data set was simulated using group μ and σ parameter estimates from the top model in the model selection analysis for which *v* varied by abstinence, but s*v* did not (Model C).

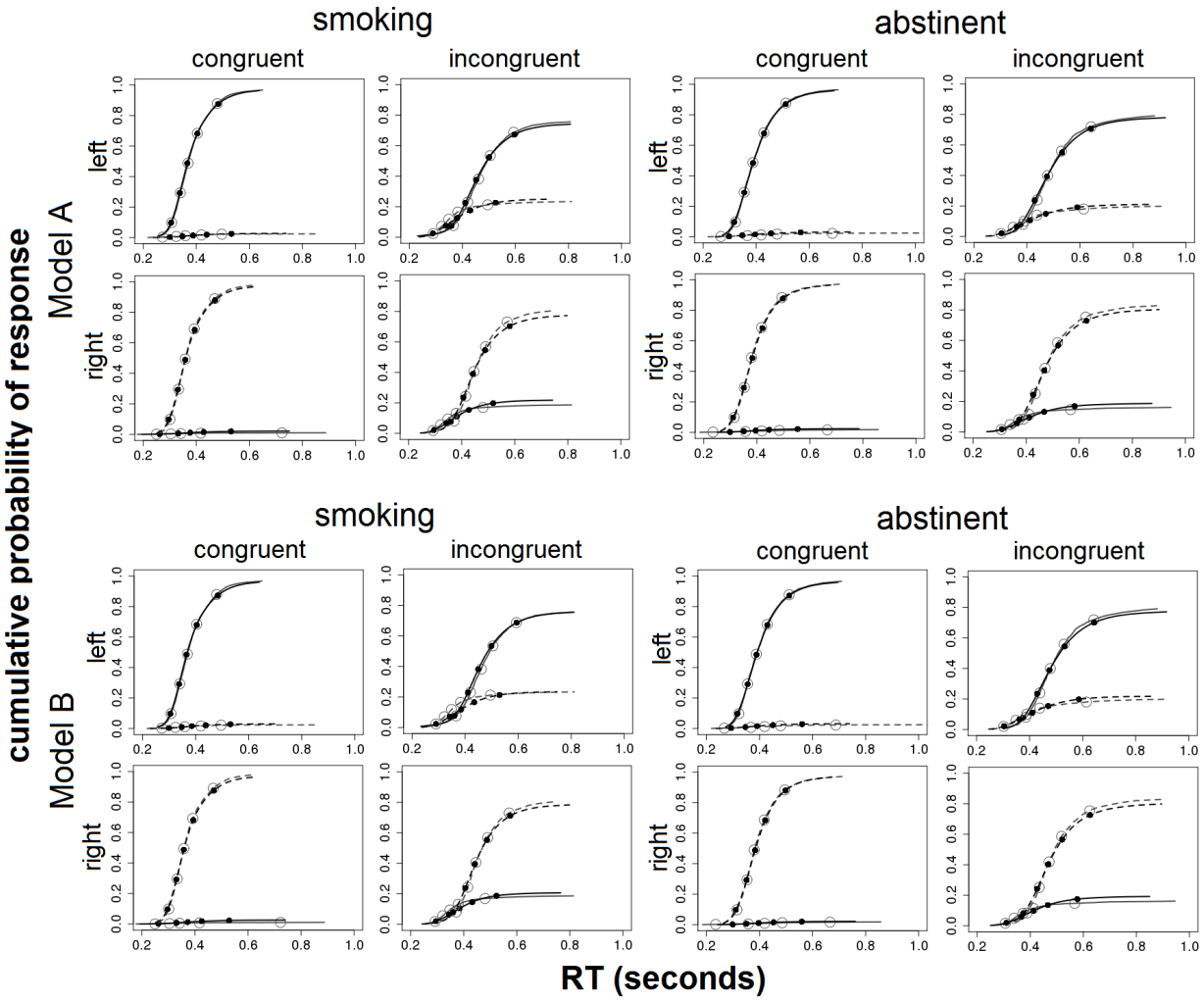
Following simulation, both the *sv-effects* and *v-effects* data sets were fit to the two respective LBA models used to generate the data: one in which *sv* and *b,* but not *v,* varied by abstinence (*sv* model), and a second in which *v* and *b*, but not *sv*, varied by abstinence (*v* model). Estimation procedures and priors used were identical to those used in analysis of the empirical data. If the LBA modeling method used is capable of separating *v* effects from *sv* effects of abstinence in this experimental paradigm, it would be expected that the fit indices used in the model selection analysis of the empirical data, DIC and WAIC, would suggest the correct model

for each data set; the *sv-effects* data would be better fit by the *sv* model, while the *v-effects* data would be better fit by the *v* model.

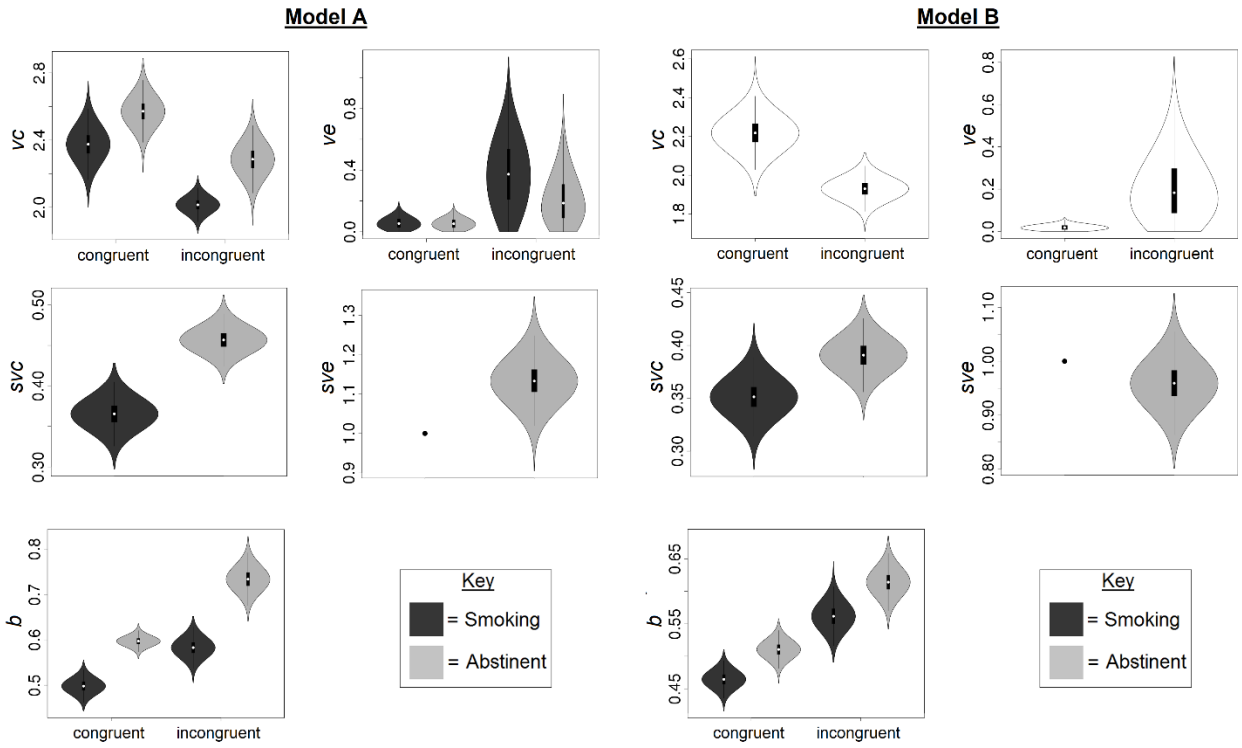
Model fit indices are displayed in Supplemental Table 2a. As expected, the correct model for each respective data set is indicated by both DIC and WAIC. The "paired estimate" method used to compare WAIC scores indicated the *sv* model displayed credibly better fit to the *sv-effects* data than the *v* model (ΔWAIC = 126.05, SE = 23.59), and that the *v* model displayed credibly better fit to the *v-effects* data than the *sv* model (ΔWAIC = 457.04, SE = 43.92). Finally, the simulation and model estimation procedures were repeated a second time to ensure that findings of the model recovery study were robust. Results of this repeated analysis were no different than those of the first analysis with respect to the models that were selected by DIC and WAIC (Supplemental Figure 2b). Therefore, the results of this model recovery study provide evidence that, at least in data generated by the LBA model, it is possible to discriminate between abstinence effects in *sv* and those in *v* using data from this experimental paradigm.

Supplemental Tables and Figures

Supplemental Figure 1. Joint cumulative distribution function plots of empirical (grey lines, circles) and posterior predictive (black lines, dots) data for "left" (solid lines) and "right" (dotted lines) responses. Panels are separated by flanker condition (congruent/incongruent) and correct direction of the arrow stimulus (left/right). Circles and dots denote the .10, .30, .50, .70, and .90 RT quantiles.



Supplemental Figure 2. Posterior distributions of group μ model parameter estimates from Models A and B for all parameters of interest. Distributions are displayed as violin plots, which contain a box plot of posterior samples inside a kernel density plot of the same samples. White violin plots for Model B represent estimates of parameters that are held constant across abstinence conditions.



Supplemental Table 1. Means and group standard deviations (parentheses) for behavioral task summary statistics.

Flanker	Visit	Accuracy		Mean RT		SD RT	
Congruent	**Smoke**	0.982	(0.017)	0.380	(0.053)	0.055	(0.025)
	Abstinent	0.978	(0.022)	0.404	(0.055)	0.070	(0.022)
Incongruent	**Smoke**	0.788	(0.118)	0.448	(0.065)	0.081	(0.025)
	Abstinent	0.818	(0.115)	0.483	(0.076)	0.092	(0.030)

Supplemental Table 2a. Relative model fit indices from the model recovery study using simulated data.

Data	Model	WAIC	SE - WAIC	DIC
sv-effects	***sv* model**	-41246.48	319.03	-1649.73
	***v* model**	-41120.43	320.70	-1644.27
v-effects	***sv* model**	-37765.36	316.24	-1509.43
	***v* model**	-38222.40	314.52	-1528.79

Supplemental Table 2b. Relative model fit indices from the repetition of the model recovery study using simulated data.

Data	Model	WAIC	SE - WAIC	DIC
sv-effects	***sv* model**	-43146.70	311.19	-1725.87
	***v* model**	-42987.96	312.58	-1718.90
v-effects	***sv* model**	-44139.03	318.13	-1764.63
	***v* model**	-44536.30	316.23	-1781.54

Supplemental References

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16*(6), 1129-1135.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association, 90*(430), 773-795.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2015). Analytic Posteriors for Pearson's Correlation Coefficient. *arXiv preprint arXiv:1510.01188*.

Ratcliff, R., & Van Dongen, H. P. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences, 108*(27), 11285-11290.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods, 18*(3), 368.

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. *Experimental psychology*.

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive psychology, 63*(4), 210-238.