Supplementary Materials

Sampling Methods

For each model we used three times as many chains as model parameters. In a first step, sampling was carried out separately for individual participants in order to get good start points for hierarchical sampling. During an initial burn-in period there was a probability of .05 that a crossover step was replaced with a migration step. After burn in only crossover steps were used and sampling was continued until the proportional scale reduction factor ($\hat{R}$) was less than 1.1 for all parameters, and also the multivariate version was less than 1.1 (Brooks & Gelman, 1998). Hierarchical estimation assumed independent normal population distributions for each model parameter. Population-mean start points were calculated from the mean of the individual-subject posterior medians and population standard deviation from their standard deviations, with each chain getting a slightly different random perturbation of these values. Hierarchical sampling used probability .05 migration steps at both levels of the hierarchy during burn in and only crossover steps thereafter with thinning set at 5 (i.e., only every 5th sample was kept), with sampling continuing until $\hat{R}$ for all parameters at all levels, and the multivariate $\hat{R}$ values, were all less than 1.1. The final set of chains were also inspected visually to confirm convergence and consisted of 300 iterations (i.e., 5400 samples in total for the LBA and 5100 for the choice Wald and 3900 for the detection Wald).

Starting Point Variability for the Wald Models

We compared models for each task using the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Table S1 defines the parameterization and number of estimated parameters for each model. Smaller values indicate a better model in terms of providing the best tradeoff between simplicity and goodness of fit. DIC is on a logarithmic scale, so a difference of 10 can be considered strong evidence in favor of the model with the smaller DIC value. DIC values can be converted to model weights (*w*, Wagenmakers & Farrell 2004), which under the appropriate prior and assuming the true model is in the set of models under consideration, approximate the probability that a model in a given set of models is the data-generating model.

For both simple and choice DRTs we compared three models that differed on whether start point noise was assumed to either be absent (i.e., $A = 0$), the same for all conditions and accumulators, or the same for all conditions but different between accumulators. As shown in Table S1, for the simple DRT data, the model with a common estimate of start-point variability was selected ($w = .985$), followed by the model where start-point variability differed with load ($w = .015$), with the model with no start-point noise clearly rejected ($w = 0$). For the choice data, in contrast, the model with no start point variability was selected ($w = 0.996$) followed by the model with a single ($w = 0.004$) and two ($w = 0$) estimates. Start-point variability was not needed to model choice, likely because in that case detection of the stimulus enabled participants to avoid sampling evidence before the stimulus appeared, which is thought to cause start-point variability (Laming, 1968). By definition, participants cannot avoid such premature sampling in a simple DRT, and so a larger level of start-point noise is likely in the simple compared to the choice DRT, consistent with our model selection results.

Table S1

Summary of models and model selection for the simple and choice DRTs.

Task Type	N	Parameterization	DIC
Simple	12	$B \sim L, R; v \sim S, L; t_0 \sim L; p_f \sim L$	-1501
	13	$A \sim 1; B \sim L, R; v \sim S, L; t_0 \sim L; p_f \sim L$	**-1553**
	14	$A \sim L; B \sim L, R; v \sim S, L; t_0 \sim L; p_f \sim L$	-1544
Choice	16	$B \sim L, R; v \sim S, L, M; t_0 \sim L; p_f \sim L$	**10488**
	17	$A \sim 1; B \sim L, R; v \sim S, L, M; t_0 \sim L; p_f \sim L$	10500
	18	$A \sim S, R; B \sim L, R; v \sim S, L, M; t_0 \sim L; p_f \sim L$	10512

Note: The model parameters are: *A* (start-point), *B* (threshold), *v* (rate), t_0 (non-decision time) and p_f (omission probability). The parameters can vary with experimental factors of *S* (stimulus, *bright* vs. *dim*) and *L* (load: *absent* vs. *present*), and for the choice DRT, factors describing the accumulators, either *R* (response accumulator: *bright* vs. *dim*), or in terms of the relationship between the accumulator and stimulus, *M* (*matching* vs. *mismatching*). A "1" indicates that a single value of the parameter was estimated that was assumed to be the same for all conditions and accumulators. N indicates number of estimated parameters. Bold DIC values indicate the preferred models.

In further analyses we focus on the models with the lowest DIC values in Table S1. In the supplementary materials we report the results of a parameter-recovery study (Heathcote, Brown & Wagenmakers, 2015) for the simple DRT model, which found that in contrast to the other parameters, the start-point variability parameter produced relatively uncertain estimates. However, this uncertainty did not compromise the estimation of the other parameters that were our main focus of interest in understanding the effects of the load manipulation. Both selected models provided an accurate description of the data; details of goodness-of-fit are also provided in supplementary materials.

LBA Model Specification

In the LBA model, each accumulator had starting points uniformly distributed in the interval 0-*A* that was drawn independently for each accumulator on each trial. The same value of start-point noise (*A*) parameter was assumed for all conditions. Evidence accrues linearly and deterministically at a rate drawn, independently for each accumulator and trial, from a normal distribution truncated below by zero with mean *v* and rate standard deviation s_v. In order to identify the model (Donkin, Brown & Heathcote, 2009), the s_v parameter was fixed at 1 for the accumulator that mismatched the stimulus (i.e., the bright accumulator when the stimulus was dim and the dim accumulator when the stimulus was bright), but s_v for the other (matching) accumulator was estimated. The parameter was allowed to vary with the stimulus (i.e., bright or dim). The *v* parameter was estimated for matching and mismatching accumulators for each stimulus (bright and dim). The threshold *(b*) was allowed to vary between accumulators to accommodate potential response biases (e.g., a lower threshold for the bright accumulator would cause a bias to respond “bright”), and parameterized concerning the gap (*B*) between the top of the start point noise and threshold (i.e., $B = b - A$), with *B* allowed to vary with load. Non-decision time (t_0) was assumed the same for both accumulators but allowed to vary with load.

Priors

Priors were chosen to be vague and have little influence on estimation. For the LBA in individual participant estimation, priors were normal distributions that were truncated below at zero for *B*, *A*, and s_v parameters, and truncated at 0.1s for the t_0 parameter (assuming that stimulus-contingent responses made in less than 0.1s are implausible). The t_0 prior was truncated above by 1s, and no posterior samples ever approached this limit. No other truncations were assumed, so the *v* prior was unbounded, and the same was true for the p_f parameter as it was estimated on a probit scale (i.e., as a z score, by taking an inverse cumulative normal transform, which maps p=0 to –∞ and p=1 to ∞). For *B* parameters the prior mean was 1 and for *A* 0.5. The *v* parameters for matching accumulators were given a prior mean of 1 and for mismatching accumulators a prior mean of 0 was assumed. The s_v parameters for the matching accumulator had a prior mean of 0.5, the t_0 parameters a prior mean of 0.3s, and the p_f parameter a prior mean of -1.5 (corresponding to a 6.7% omission rate). All priors had a standard deviation of 2. The mean parameters of the population distributions were assumed to have priors of the same form as for individual estimation. The standard deviations hyper parameters were assumed to have exponential distributions with a scale parameter of one. Plots superimposing prior and posterior distributions revealed strong updating (i.e., posteriors dominated priors), making it clear that the prior assumptions had little influence on posterior estimates.

The same prior means and standard deviations were used for the Wald model for the parameters that share the same names. The Wald prior differed in that the rate parameter priors were positive (i.e., truncated below by zero). Both Wald and LBA models assumed that thresholds (*B*) mean rates (*v*) and non-decision time (t_0) could vary with load. Hence, the LBA model had 18 parameters: 8 *v*s (load x stimulus x match), 4 *B*s (load x match), 2 t_0s (load), 2 p_f s (match), one *A* and one s_v. The choice Wald model had 17 parameters, as it lacked an s_v parameter. The detection Wald model had 13 parameters as it lacked the match factor, and so had only 4 *v*s.

Posterior Inference

To take account of correlations between parameter estimates in within-subject comparisons we first calculated parameter differences for each posterior sample for each participant, and then averaged the differences for each posterior sample over participants. The *p* value corresponds to the proportion of average posterior sample differences greater or less than zero, depending the direction that was consistent with the stated hypothesis. 95% credible intervals were calculated from the values below and above which 2.5% of posterior samples (or differences) occurred.

Model Fit

Choice-Data Fit. Figure S1 and S2 displays the fit of the LBA and Wald models to the choice data in terms of defective cumulative distribution functions (lines) and 10th, 30th, 50th, 70th and 90th percentiles (points from left to right) averaged over participants. Each panel corresponding to a cell of the design, and contains a pair of cumulative functions, one for each response, which are "defective" because they asymptote at the probability of that response. Note that the rate of omission equals one minus the sum of the asymptotic heights of each function, as that sum corresponds to the probability that a response was made. The thick black line and open points correspond to the data and the thin grey lines solid black points to the model prediction averaged over posterior samples. The grey points correspond to percentile predictions for 500 randomly selected sets of posterior parameter samples, so their spread gives an idea of the uncertainty in the model's predictions.

As shown in Figure S1, the fit of the LBA model in terms of response probability is excellent, with a close match between asymptotic model and data response probabilities. The observed response time distributions generally fall well within the region of uncertainty for the model response time distributions, with a few minor exceptions, mainly for slower and less common responses. Figure S2 shows that the Wald model with no start point variability provides a similarly good fit. For the simple detection data, Figure S3 shows that the average fit of the selected Wald model was excellent.

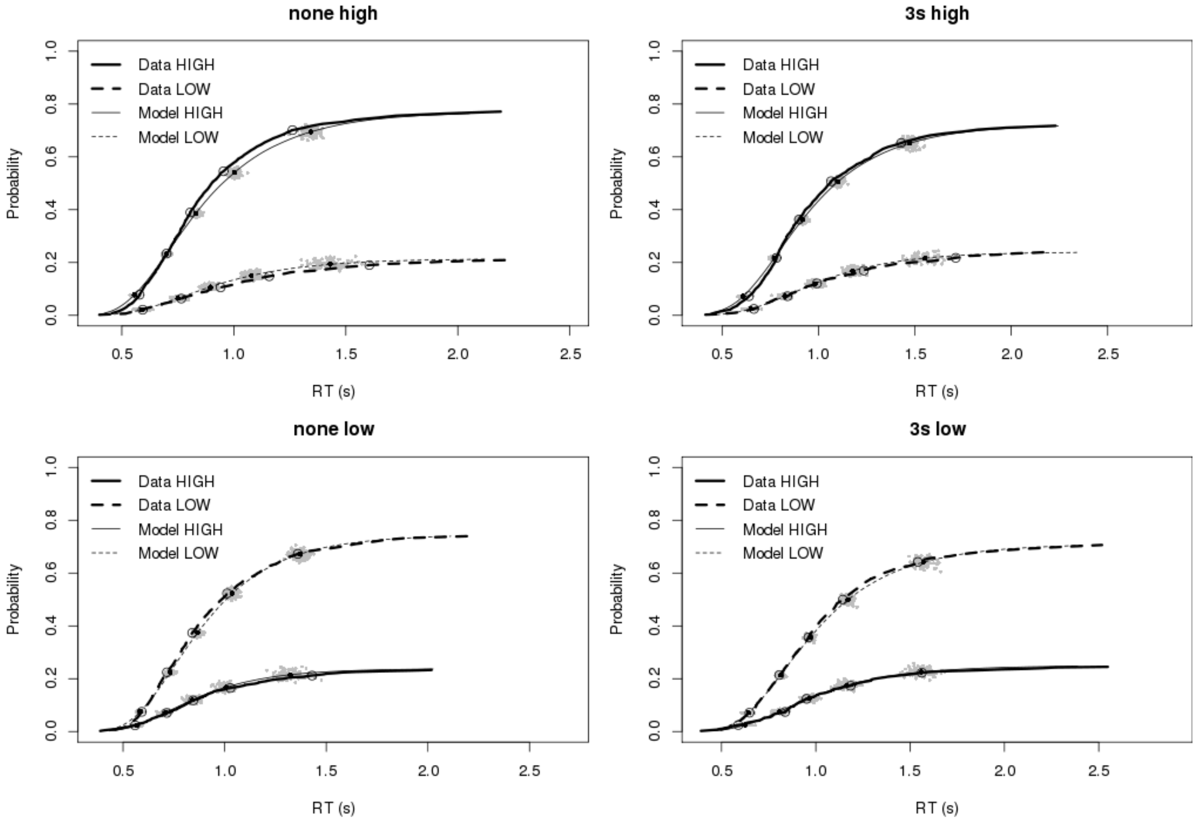

Figure S1. Cumulative distribution functions for data (thick lines) and fits (thin grey lines) of the LBA model to the choice data. Each panel contains results for both HIGH (bright) and LOW (Dim) responses. Symbols mark the 10th, 30th, 50th, 70th and 90th percentiles (solid for average fits, open for data). Grey points are 500 percentile estimates from fits for random draws from posterior parameter samples; the grey line and black solid points are the average of these 500 fits.

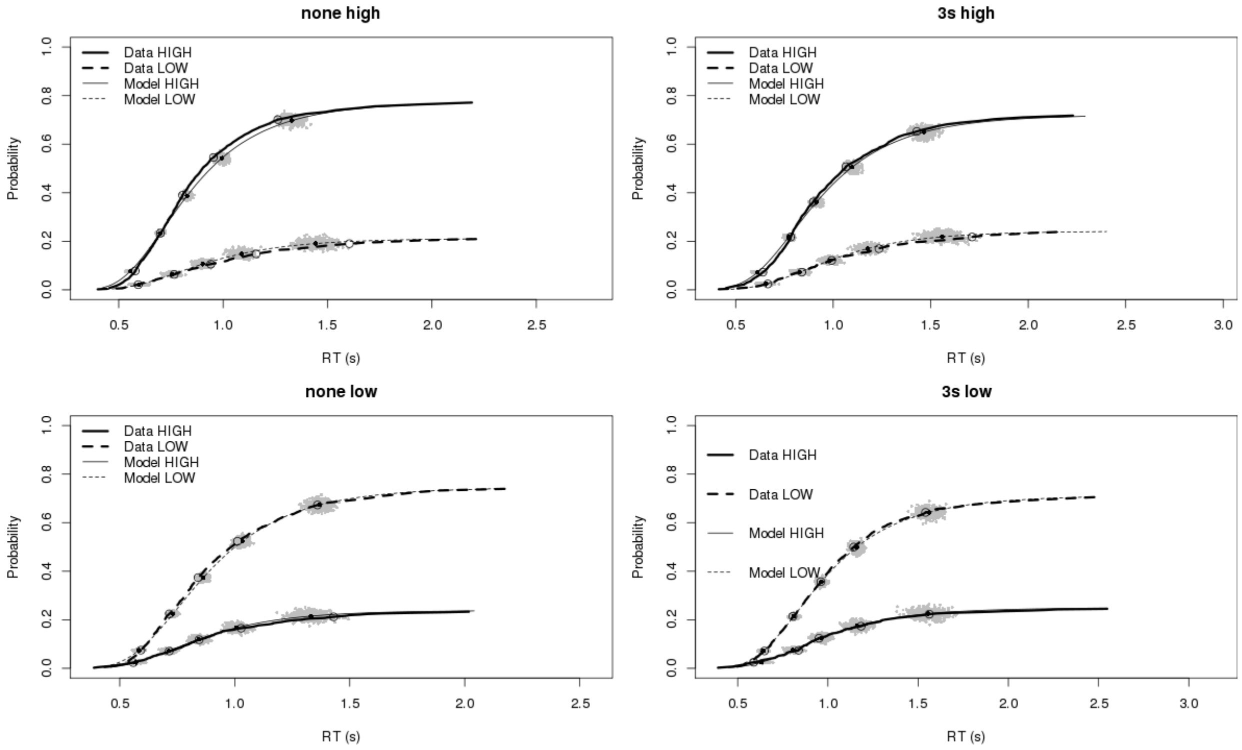

Figure S2. Cumulative distribution functions for data (thick lines) and fits (thin grey lines) of the Wald model with no start-point variability to the choice data. Each panel contains results for both HIGH (Bright) and LOW (Dim) responses. Symbols mark the 10th, 30th, 50th, 70th and 90th percentiles (solid for average fits, open for data). Grey points are 500 percentile estimates from fits for random draws from posterior parameter samples; the grey line and black solid points are the average of these 500 fits.

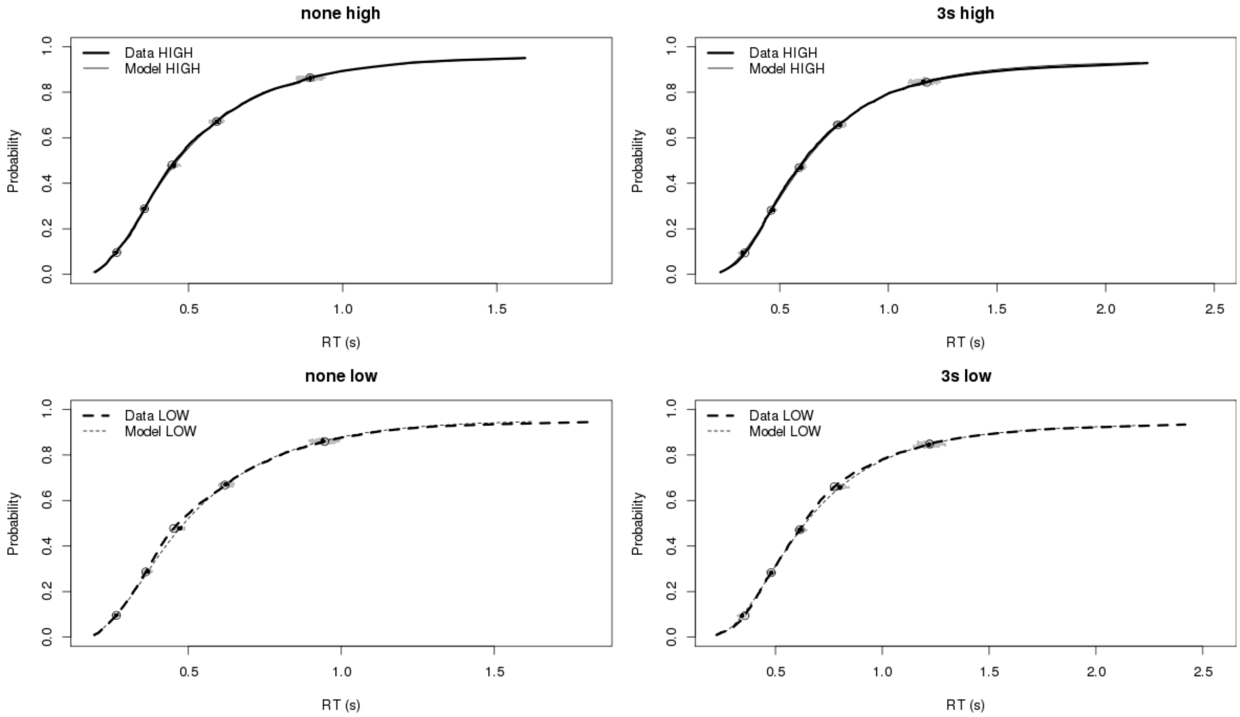

Figure S3. Cumulative distribution functions for data (thick lines) and fits (thin grey lines) of the Wald model with start-point variability to the simple detection data. Each panel contains results for either HIGH (Bright) and LOW (Dim) responses. Symbols mark the 10th, 30th, 50th, 70th and 90th percentiles (solid for average fits, open for data). Grey points are 500 percentile estimates from fits for random draws from posterior parameter samples; the grey line and black solid points are the average of these 500 fits.

LBA Choice Results

The response omission parameter was higher in the load-present condition by 2.1%, [1.5, 2.8] (3.4%, [2.9, 4.0] vs. 1.3%, [1.0, 1.7], *p* < .001), non-decision time was 0.014s [.002, .028] faster (.130s, [.122, .141] vs. 0.144s, [.132, .156], *p* = .042), and the response threshold was 0.15, [.08, .22] higher (averaged over bright and dim accumulators, 1.36, [1.30, 1.46] vs. 1.21, [1.15, 1.26], *p* < .001). The average rate for the matching accumulator, which largely determines the speed of correct responses, was clearly lower [.12, .26] in the load-present condition (2.3, [2.24, 2.37] vs. 2.5, [2.43, 2.56], *p* < .001), whereas the mismatching rate was only marginally lower [-.05, .15] in the load-present condition (1.24, [1.16, 1.33] vs. 1.29, [1.22, 1.37], *p* = .17). The difference between match and mismatch rates, which largely determines accuracy, was smaller [.05, .24] for the load-present condition (1.06, [.99, 1.13] vs. 1.2, [1.14, 1.27], *p* < .001).

Overall, these results indicate that more frequent choice errors and slower responding in the load-present condition was due to a reduced accumulation rate. The tendency for increased choice errors was partially counteracted by a higher threshold, which also exaggerated the slowing in the load-present condition, consistent with participants attempting to trade speed for accuracy to counteract the load effect. The slowing under load was also counteracted by a decrease in non-decision time, but this effect was small; without it slowing would have been 0.127 s rather than 0.113 s in median RT, a decrease of only 11%.

LBA Single Accumulator Model

We fit a 13 parameter LBA model with start-point noise (the same for all conditions), thresholds varying with load and accumulator, mean rates with stimulus, load and match, rate standard deviation varying with match, omission rate varying with load and the same non-decision time for all conditions. Priors were the same as for the choice case for analogous parameters. This model had a DIC = -1130.4, substantially greater than the Wald model with start-point variability (DIC = -1552.7), although as shown in Figure S4 it fits was generally good.

Inferences from parameter comparisons matched those for the Wald. They were essentially identical for p_f, and consistent for non-decision time in that it was higher for load-present than load-absent ($p < .001$) but the difference was miniscule 0.1015s [0.101 – 0.102] vs. 0.1045 [0.1031 – 0.1061], due to samples being piled up on the lower bound of the prior. This was not a problem for the remaining parameters: thresholds for load-present (0.852, [0.827-0.879]) were higher than for load-absent (0.707, [0.683-0.731]), $p < .001$, whereas mean rates were higher for load-absent (3.43, [3.36-3.51]) than load-present (2.85, [2.77-2.92]), $p < .001$.

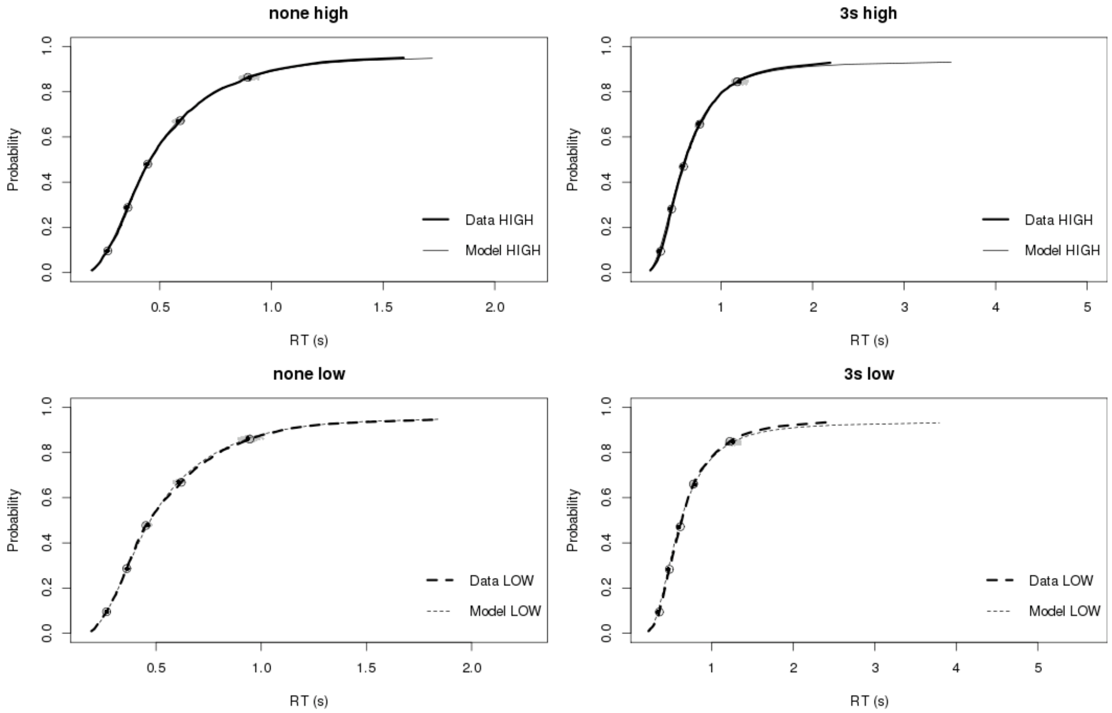

Figure S4. Cumulative distribution functions for data (thick lines) and fits (thin grey lines) of the LBA model with start-point variability to the simple detection data. Each panel contains results for either HIGH (Bright) and LOW (Dim) responses. Symbols mark the 10th, 30th, 50th, 70th and 90th percentiles (solid for average fits, open for data). Grey points are 500 percentile estimates from fits for random draws from posterior parameter samples; the grey line and black solid points are the average of these 500 fits.

Single Accumulator Parameter Recovery

The ability of a model and associated estimation method to recover its parameters can be determined by generating data from the model and then obtaining parameter estimates from that simulated data using the estimation method (Heathcote, Brown & Wagenmakers, 2015). Recovery can vary both with the region of parameter space and experimental design from which data are simulated. We simulated exactly the same design as our experiment, except that we examined a wider range of sample sizes for each condition: 50, 100, 200 and 400 trials per cell of the design. We preformed 20 sets of parameter recoveries, with each set based on the posterior median of the parameters for each participant. For each set we carried out 200 replications, sampling a fresh set of data using the participant specific parameters and the same methods and priors as we did for the real data. For each parameter we computed the average of the posterior means, the average 95% Credible Interval, and the standard error of the posterior mean. This allowed us to quantify our ability to recover the data-generating parameter values. Figure S5 showing example results for parameters derived from participant two, who showed some differences between estimates for the models with and without start-point noise. Estimation is generally excellent for 400 trials per condition, and usually very good for 200 trials, which is a little less than the average for our data, and the same was true for other simulations based on other participants' parameters (for the full set of results see https://osf.io/e8kag/). We also assessed how well the Bayesian procedure estimates uncertainty in the parameter estimates (i.e., the validity credible interval estimates) by assessing "coverage", that is, by counting the number of times the data generating values fell within each replication's 95% credible intervals. As show in Figure S5, and again as was the case generally, coverage was close to nominal for 400 trials per condition and usually quite good for 200 trials per condition.

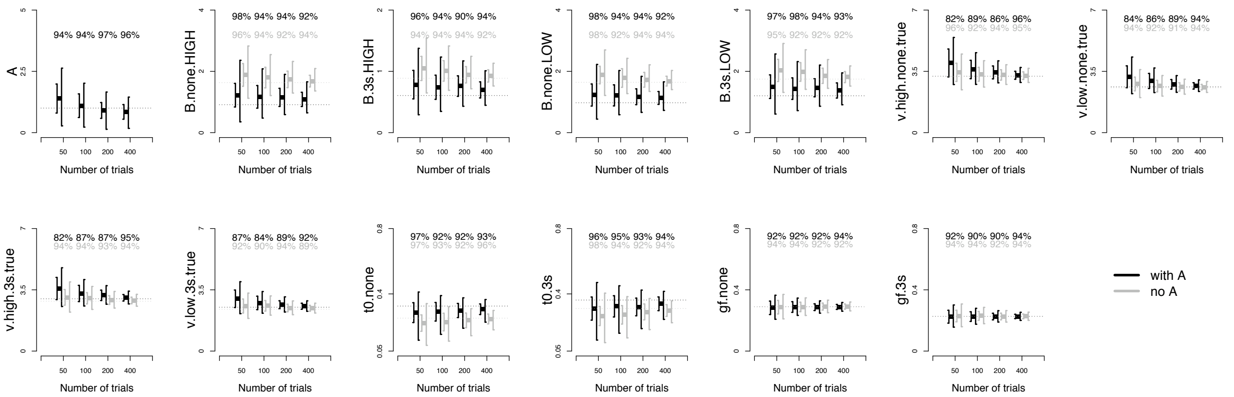

Figure S5. Recovery results for parameters derived from participant 2. Square symbols show the average of the posterior means. The double bars show the average 95% Credible Interval (right side), and the standard error of the posterior mean (left side). Results with no start-point noise are presented in a lighter shade than those with start-point noise. True values are indicated by horizontal dotted lines with corresponding shading. The numbers at the top of each panel indicate the percentage of times over the 200 replications that the data-generating values fell in the replications' 95% credible interval, again in corresponding shades.

Parameter and Pursuit Tracking Correlations for the Wald Model

Results of fixed and random effect plausible value correlation analyses are listed in Table S2 for the choice Wald model and Table S4 and S5 for the Wald model for the bright and dim DRT.

Table S2

Plausible values for the correlation between Pursuit Tracking RMSE for the Two-Alternative Choice Task and the parameters of the winning Wald model according to DIC

Effect	**Cognitive Workload**	**Parameters**				
		t_0	p_f	*B*	*vMatch*	*vMatch - vMismatch*
Fixed (participant)	Absent	-.19 (.08)	**.38 (.01)**	.27 (.04)	**-.48 (<.001)**	**-.36 (<.001)**
	95%CI	[-.45, .08]	**[.07, .63]**	[-.03, .54]	**[-.59, -.35]**	**[-.54, -.17]**
	Present	**-.36 (<.001)**	**.52 (<.001)**	.30 (.07)	**-.69 (<.001)**	**-.52 (<.001)**
	95%CI	**[-.54, -.19]**	**[.37, .67]**	[-.11, .64]	**[-.77, -.61]**	**[-.68, -.34]**
Random (population)	Absent	-.16 (.27)	.32 (.11)	.22 (.20)	-.40 (.04)	-.30 (.10)
	95%CI	[-.62, .35]	[-.21, .72]	[-.03, .66]	[-.75, .05]	[-.70, .17]
	Present	-.30 (.10)	.45 (.03)	.25 (.19)	**-.61 (<.001)**	- .44 (.03)
	95%CI	[-.70, .17]	[-.01, .77]	[-.32, .71]	**[-.86, -.24]**	[-.78, .02]

Note. Values in **bold** pertain to correlations with 95% credible intervals that do not include zero. Values in parentheses are Bayesian *p* values, representing the probability of a value being greater than zero for negative correlations and less than zero for positive correlations.

Table S3

Plausible values for the correlation between steering RMSE for the ISO Detection Response Task (Bright) and the parameters of the winning Wald model according to DIC

Effect	**Cognitive Workload**	**Parameters**			
		t_0	p_f	*B*	*vMatch*
Fixed (participant)	Absent	**-.52 (<.001)**	**-.11 (.02)**	**.19 (.002)**	**-.48 (<.001)**
	95%CI	**[-.61, -.43]**	**[-.22, -.01]**	**[.06, .31]**	**[-.59, -.36]**
	Present	**-.43 (<.001)**	**.10 (.01)**	.07 (.22)	**-.45 (<.001)**
	95%CI	**[-.53, -.34]**	**[.02, .18]**	[-.11, .25]	**[-.56, -.34]**
Random (population)	Absent	**-.45 (.02)**	-.09 (.35)	.15 (.26)	-.41 (.04)
	95%CI	**[-.77, -.01]**	[-.53, .37]	[-.32, .57]	[-.75, .04]
	Present	-.36 (.06)	.08 (.37)	.06 (.42)	-.38 (.05)
	95%CI	[-.72, .09]	[-.38, .51]	[-.42, .51]	[-.73, .07]

Note. Values in **bold** pertain to correlations with 95% credible intervals that do not include zero. Values in parentheses are Bayesian *p* values, representing the probability of a value being greater than zero for negative correlations and less than zero for positive correlations.

Table S5

Plausible values for the correlation between steering RMSE for the modified Detection Response Task (Dim) and the parameters of the winning Wald model according to DIC

Effect	**Cognitive Workload**	**Parameters**			
		t_0	p_f	B	*vMatch*
Fixed (participant)	Absent	**-.49 (<.001)**	-.08 (.05)	**.20 (<.001)**	**-.49 (<.001)**
	95%CI	**[-.57, -.39]**	[-.19, .02]	**[.08, .32]**	**[-.64, -.34]**
	Present	**-.45 (<.001)**	**.10 (.01)**	-.10 (.10)	**-.54 (<.001)**
	95%CI	**[-.55, -.36]**	**[.02, .18]**	[-.24, .05]	**[-.62, -.45]**
Random (population)	Absent	-.41 (.03)	-.07 (.39)	.17 (.24)	-.42 (.04)
	95%CI	[-.75, .03]	[-.51, .39]	[-.31, .58]	[-.76, -.03]
	Present	-.38 (.05)	.08 (37)	-.08 (.38)	**-.46 (.02)**
	95%CI	[-.73, .07]	[-.38, .51]	[-.52, .39]	**[-.78, -.03]**

Note. Values in bold pertain to correlations with 95% credible intervals that do not include zero. Values in parentheses are Bayesian *p* values, representing the probability of a value being greater than zero for negative correlations and less than zero for positive correlations.

References

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

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.

Heathcote, A., Brown, S.D. & Wagenmakers, E-J. (2015). An introduction to good practices in cognitive modeling. In Forstmann, B. U., & Wagenmakers, E.-J. (Eds.). *An Introduction to Model-Based Cognitive Neuroscience*. Springer, New York.

Ly, A., Boehm, U., Heathcote, A., Turner, B.M., Forstmann, B., Marsman, M. & Matzke, D. (2017). A flexible and efficient hierarchical Bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. In A.A. Moustafa (Ed.) *Computational models of brain and behavior* (pp. 467-480). Wiley Blackwell.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347-356.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.s

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.