SUPPLEMENTARY MATERIALS

Computational Modeling of Neuropsychological Test Performance to Disentangle Impaired Cognitive Processes in Cancer Patients

Joost A. Agelink van Rentergem, Ph.D.[1,2], Ivar E. Vermeulen, Ph.D.[3], Philippe R. Lee Meeuw Kjoe, M.Sc.[2], Sanne B. Schagen, Ph.D.[1,2]

[1]Department of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, The Netherlands, [2]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands, [3]Department of Communication Science, VU University Amsterdam, Amsterdam, The Netherlands

Supplementary Methods

Outlier removal

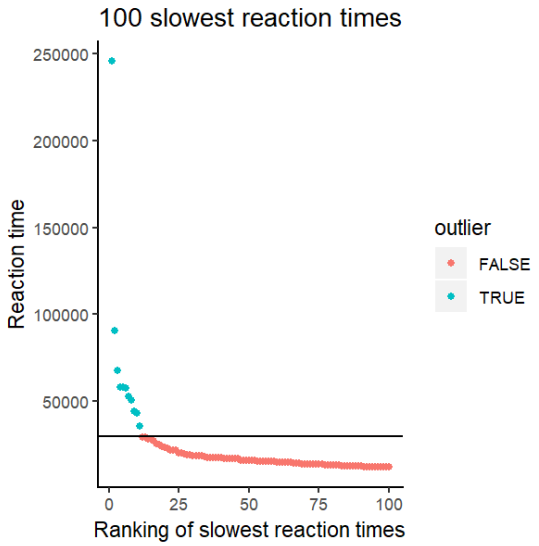
It is preferable to not remove outliers, as we cannot be sure that outlying values are not a part of normal variation. However, the procedure as followed in the data collection was susceptible to outliers, as testing was unsupervised and took place at the participants' home. Therefore, a phone ringing or someone distracting the participant could easily lead to extremely slow reaction times that are not part of normal variation, because they come from a distinct contaminating process rather than slow processing.

To define outliers, we considered using a predefined outlier criterion, like a criterion based on the number of standard deviations that a reaction time is removed from the mean, the number of absolute deviations that a reaction time is removed from the median, or the number of Inter-Quartile Ranges that a reaction time is removed from the Quartiles. These are of varying

use, as some of these make an assumption of normality (which is demonstrably violated in the case of reaction times), or may be susceptible to the effects of outliers themselves. Furthermore, we were hesitant to use any criterion that would remove 5%, 2.5% or 1% of values, as with 20 thousand reaction times, this would be a substantive amount of data that would be removed without any argumentation.

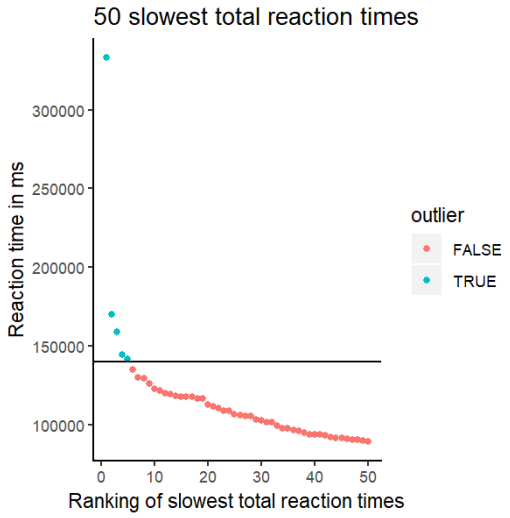
To detect outliers, we looked at the distribution of reaction times, formulating a criterion based on the distribution of observed values, because we had no a priori reason to formulate any cut-off, and wanted to remove as few data points as possible. As demonstrated below, none of the conclusions of our analyses changed with and without participants with outlying values according to these criteria.

10 participants took more than 30 seconds to move from one circle to the next, and were removed. This is illustrated in Supplementary Figure 1.



Supplementary Figure 1. 100 Ordered Reaction Times per Trial, from Slowest to Fastest.

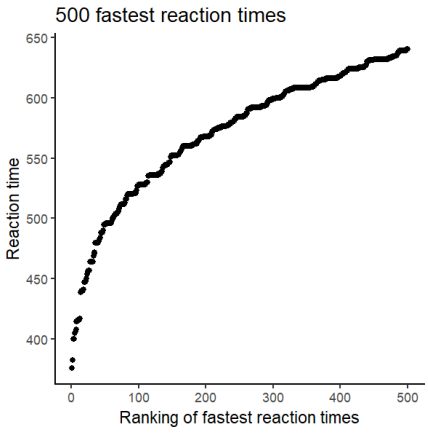
One additional participant was excluded, because the total time it took this participant to complete parts A and B exceeded 140 seconds, twice. This is illustrated in the following Supplementary Figure 2.



1

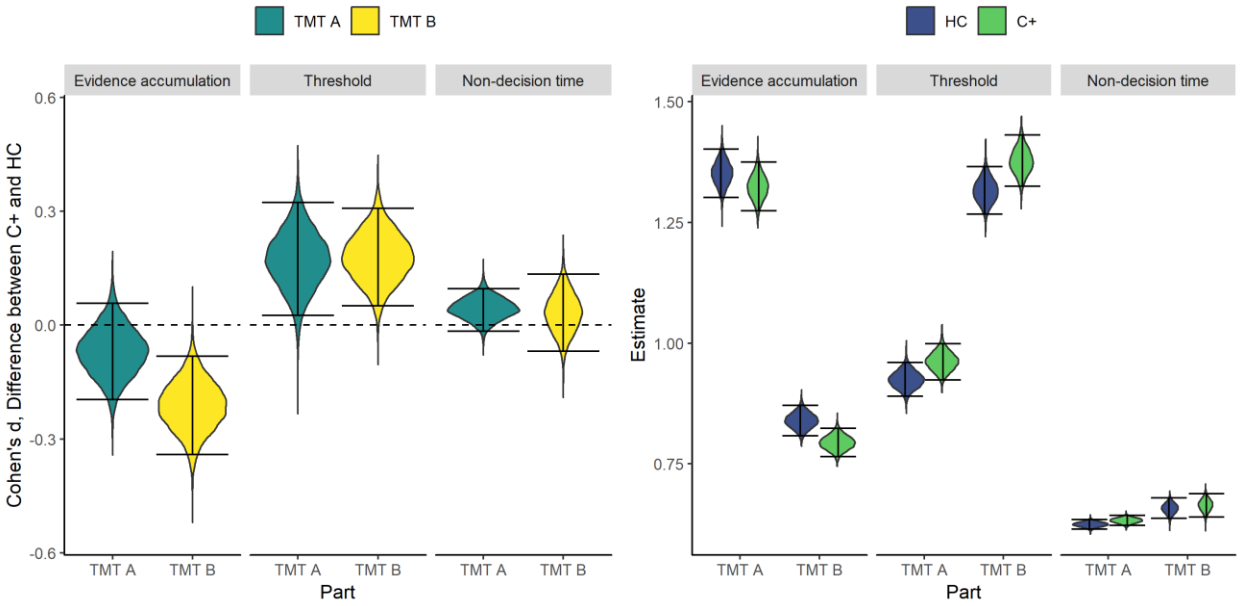
Supplementary Figure 2. 50 Ordered Total Reaction Times per Part, from Slowest to Fastest.

There were no exceptionally fast responses. All responses were above 0.350 seconds. This is illustrated in Supplementary Figure 3.



Supplementary Figure 3. 500 Ordered Reaction Times per Trial, from Fastest to Slowest.

To check whether our outlier removal procedure had an impact on the conclusions we drew, we reran the main analysis without removing any outliers, and found that all results were qualitatively the same. For this analysis, participants were matched on age between the two groups, resulting in two groups of 201 participants. The results are provided in Supplementary Figure 4.



Supplementary Figure 4. Distributions of Effect Sizes and Parameter Estimates, on Three Parameters and Parts A and B of the TMT. Intervals denote 95% Highest Posterior Density Intervals. *Notes: TMT = Trail Making Test, C+ = non-CNS cancer patients, HC = controls.*

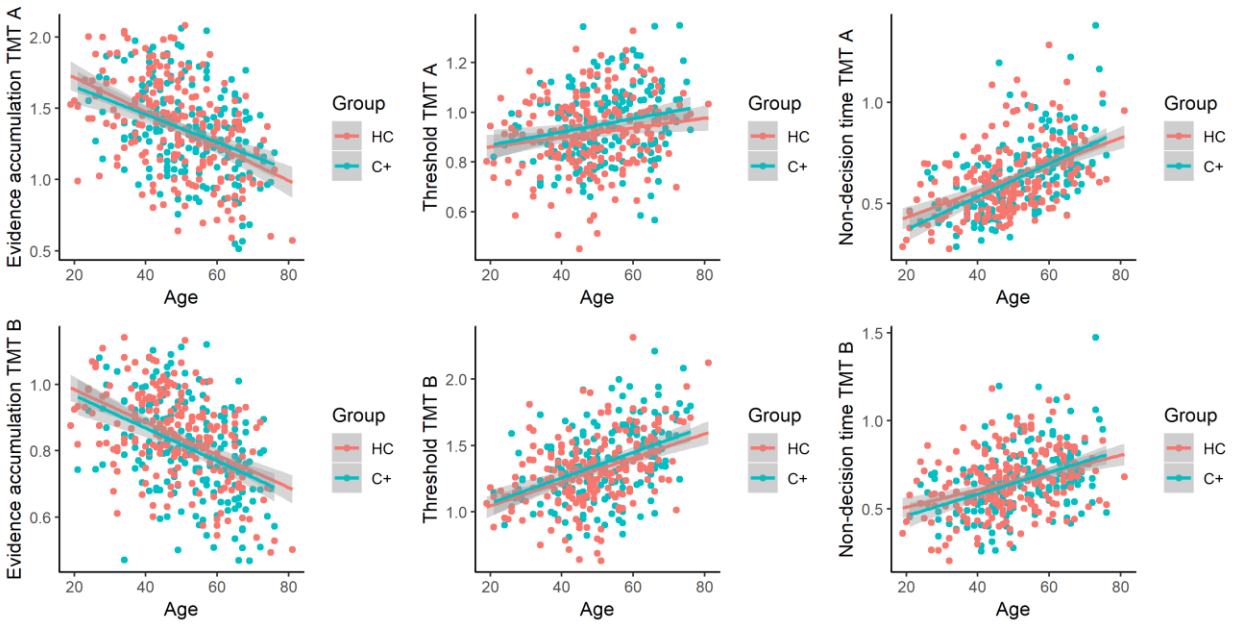
Investigation of potential age confound

The effects that were found between patient and control groups are consistent with age-related differences in the literature, for the same cognitive model (but a different task). Therefore, age may have a confounding effect on comparisons between groups with and without cancer, which is why we matched the groups on age. To investigate whether this age matching was necessary, we investigated whether there is an age difference between groups, using classical statistics.

After outlier removal (see above), the full sample consisted of 192 patients (112 women, mean (sd) age: 52.4 (11.9)) and 215 controls without a history of cancer that were recruited via participants (136 women, mean (sd) age: 48.8 (12.7))

In the full sample, the age difference is significant, but small, $t(405)=-2.90$, $p = 0.004$, $d = -0.29$, with the patient group being older. To remove the confound, we matched the two groups on age, removing 23 participants from the control group for the main analysis. After this correction, the age difference was no longer significant, $t(382)=-1.10$, $p = 0.271$, $d = -0.11$.

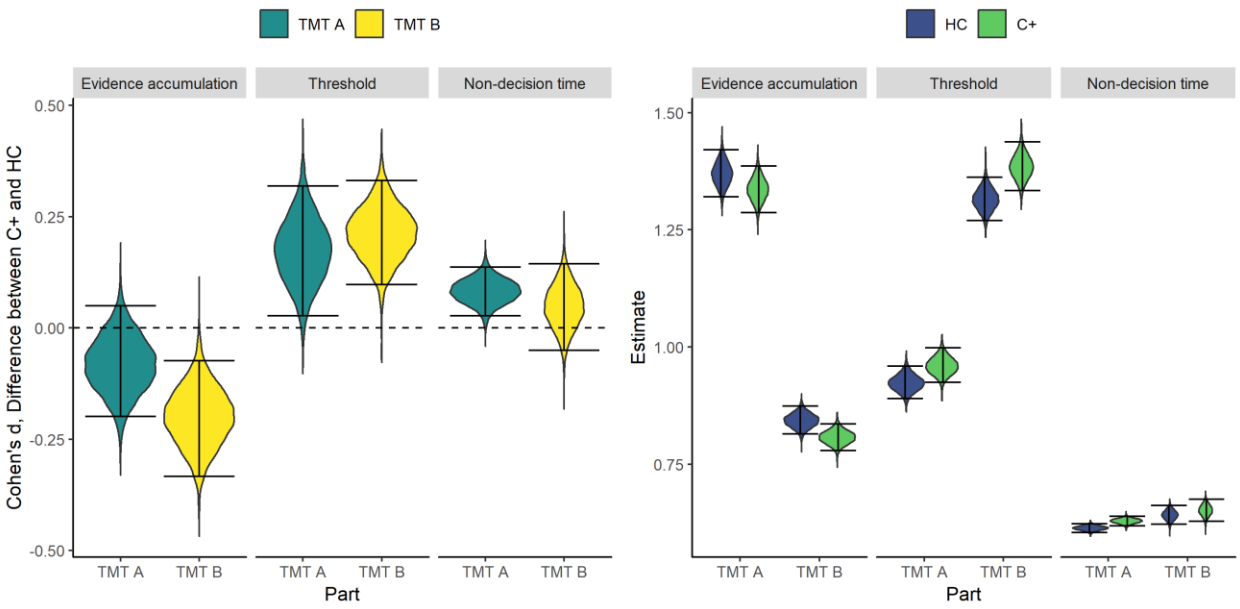
We also refitted the model to the full sample without matching, and investigated whether there was an effect of age on the different parameter estimates. For this, we used the medians of the MCMC chains as the parameter estimates per patient. For each of the six parameters, there was a significant effect of age (see Supplementary Figure 5). Therefore, age was indeed an important confound. There were no significant interactions between age and patient status, for any of the parameters.



Supplementary Figure 5. Scatterplots for Medians on Six Different Parameters with Linear Regression Lines with Confidence Intervals, Plotted for Two Groups Before Matching. *Notes: TMT = Trail Making Test, C+ = non-CNS cancer patients, HC = controls.*

We also reran the main analyses with the non-matched group of 407 participants, for which the results are provided in Supplementary Figure 6. All results are qualitatively the same, except for a new group effect on non-decision time for TMT A, with the Highest Posterior Density Interval now not overlapping zero. Therefore, there seems to be an effect between groups on non-decision time, where C+ is slower than HC. Because this effect is most likely caused by the small difference in age between groups before matching, we chose not to report it in the main article. The rest of the found effects are unaffected, and age did not induce or mask any other effects that were previously undiscovered.

We have also considered confounding by other demographic variables. However, the samples are well-matched in education (numbers provided below), and sex (112 women /192 vs. 123 women/192) , suggesting that the results will not be confounded by these factors.



Supplementary Figure 6. Distributions of Effect Sizes and Parameter Estimates Before Matching, on Three Parameters and Parts A and B of the TMT Before Matching. Intervals denote 95% Highest Posterior Density Intervals. *Notes: TMT = Trail Making Test, C+ = non-CNS cancer patients, HC = controls.*

Inclusion / exclusion criteria

Inclusion criteria:

1. adults
2. sufficient proficiency of the Dutch language,
3. basic computer skills (i.e., being able to operate the mouse and send emails independently),
4. access to a computer with an Internet connection.

An added inclusion criterion for the cancer group was

1. prior treatment with chemotherapy, radiotherapy, hormonal therapy, or immunotherapy (current hormonal therapy allowed).

Exclusion criteria for the control group were

1. history of cancer,
2. self-reported neurological or psychiatric conditions that could influence cognitive functioning (e.g., schizophrenia, psychosis, clinical depression, substance dependence, or brain pathology).

Exclusion criteria for the cancer group were

1. tumor or metastases in the central nervous system
2. distant metastases
3. disease progression
4. psychiatric/neurologic symptoms hampering test completion.

Tumor types

Patients with different non-CNS tumor types were included. Frequent types were breast cancer (42%), prostate cancer (14%), testicular cancer (7%), lung cancer (5%), colorectal cancer (4%), and bladder cancer (3%).

Treatments

Treatments in descending order of frequency were chemotherapy (79%), radiotherapy (78%), surgery (71%), hormonal therapy (46%), and immunotherapy (10%).

Education

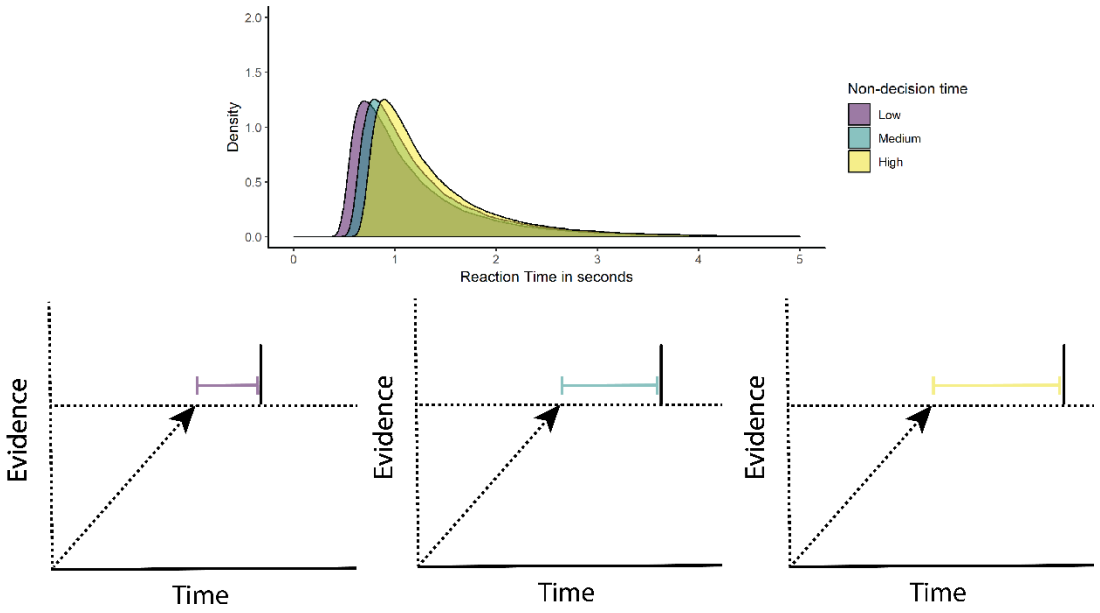
Of the patients, 21% had completed university education, 41% had completed higher vocational education, 29% had completed intermediate vocational education, 8% had completed lower vocational education, 2% had not completed a degree. Of the controls, 21% had completed university education, 53% had completed higher vocational education, 20% had completed intermediate vocational education, 6% had completed lower vocational education, and 1% had not completed a degree.

Parameters' relation to the reaction time distribution

Reaction times are disentangled into three parameters: evidence accumulation, threshold, and non-decision time. Each parameter relates to a property of the reaction time distribution. This means that from the shape and location of the reaction time distribution, we can infer estimates for each of the parameters. The reaction times are assumed to follow from a random walk, the overall direction of which is determined by the evidence accumulation parameter, until the random walk stops at the threshold. Each parameter is estimated separately for Part A and Part B of the Trail Making Test. Therefore, there are six parameters: evidence accumulation for Parts A and B, threshold for Parts A and B, and non-decision time for Parts A and B. These will be discussed (in reversed order) below.

Non-decision time

The non-decision time parameter is determined by the location of the distribution; i.e., where the distribution "starts" on the left. Non-decision time is related to the minimum amount of time it takes to respond to a trial. Non-decision time is always positive, and can be interpreted on the same time scale as the original reaction time (in this article, time in seconds). Because it is not related to a cognitive process or decision process, but is defined as what remains after we take those processes into account, it typically represents the remaining processes, like motor processes. If non-decision time is low, participants waste little time on non-cognitive tasks and respond quickly. In Supplementary Figure 7, three distributions are displayed, for equal evidence accumulation and equal threshold. The yellow distribution, with a high non-decision time, is displaced to the right. The purple distribution, with a low non-decision time, is displaced to the left. The shape of the distribution is unchanged.

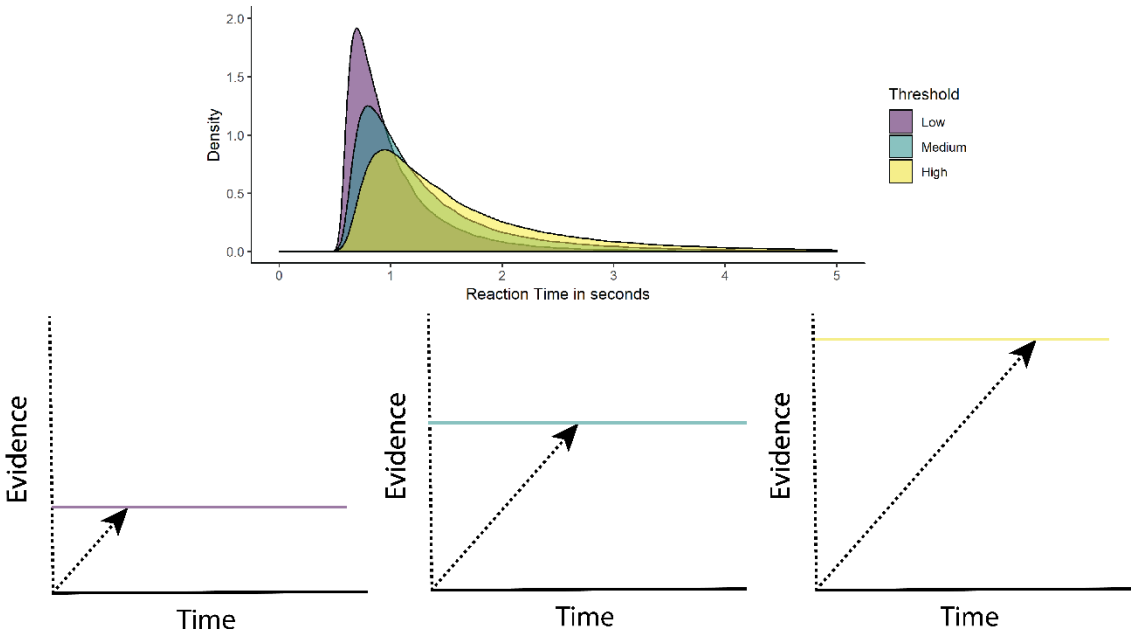


Supplementary Figure 7. Illustration of how non-decision time affects the reaction time distribution.

Threshold

The threshold parameter is determined by the variability around the modal reaction time. The threshold is related to the amount of processing that is done before a decision is taken. If the threshold is low, participants accumulate little evidence before they take a decision, and there is little randomness in the random walk, as the walk is cut off quickly. Therefore, for a low threshold, there is little variability. For a high threshold, there are many possibilities for the random walk to proceed. A random walk might slowly proceed to the high threshold on some trials, and on other trials may quickly proceed towards the high threshold. Therefore, for a high threshold, there is much more room for variability in reaction times. The threshold is positive by definition. In Supplementary Figure 8, three distributions are displayed, for equal evidence accumulation and equal non-decision time. The yellow distribution, with a high threshold, is wider, and the mode is shifted to the right. The purple distribution, with a low threshold, is

narrower, and the mode is shifted to the left. The location —where the distribution starts on the left— is unchanged.

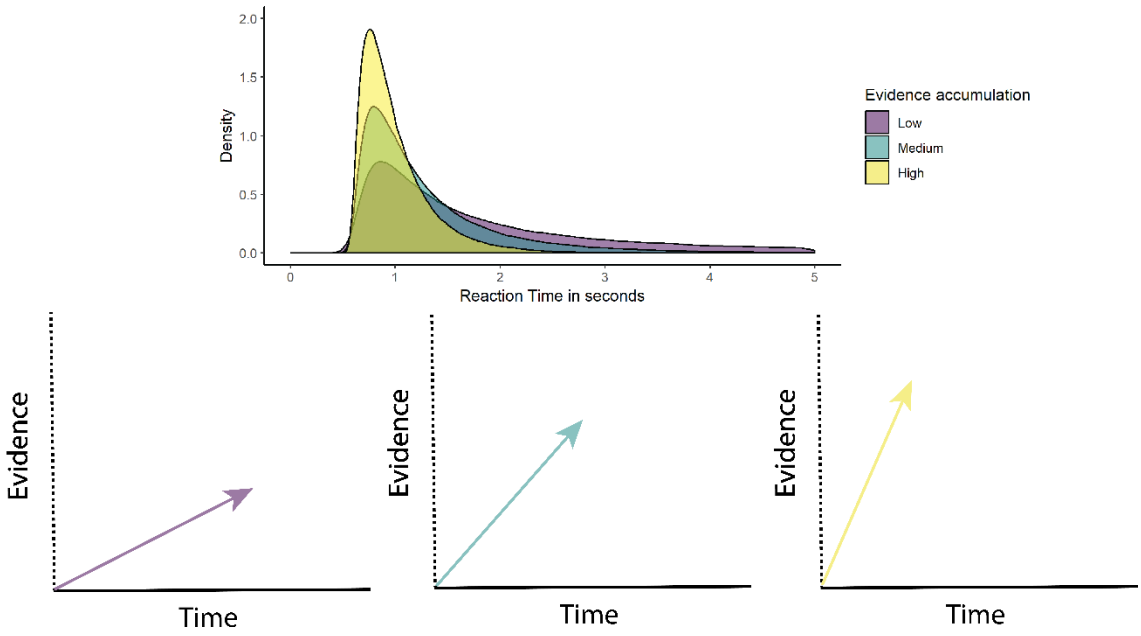


Supplementary Figure 8. Illustration of how the threshold affects the reaction time distribution.

Evidence accumulation

Evidence accumulation is determined by the thickness of the distribution's right tail, and can be interpreted as a cognitive speed component. If the evidence accumulation rate is high, participants gather evidence for a decision quickly. If the evidence accumulation rate is low, participants are more variable in the evidence that they gather, sometimes moving towards the decision threshold, sometimes moving away from the threshold. In the long run, they will reach the same conclusion, but this may take arbitrarily long. Therefore, very long reaction times do occur with a low evidence accumulation rate. In Supplementary Figure 9 three distributions are displayed, for equal threshold and equal non-decision time. The yellow distribution, with a high evidence accumulation, has little to no right tail. The purple distribution, with a low evidence

accumulation, has a thick right tail. The location —where the distribution starts on the left— and the mode —the most frequently occurring reaction time— are unchanged.



Supplementary Figure 9. Illustration of how evidence accumulation affects the reaction time distribution.

Number of parameters

Each of the six parameters is estimated for each participant. Each participant only provides 48 reaction times. Therefore, participant-level estimates will be unstable, unless we constrain them with group-level information. Hierarchical estimation provides the optimal balance between preserving the unique information available on the participant, and the constraining information available from the group. Hierarchical estimation in this way allows us to leverage the large sample size of 384 participants, to improve participant-level parameter estimates. This is especially important in participants who are potentially fatigued, like cancer patients often are, for whom prolonged test administration may be unfeasible or lead to biased results when participants start to underperform later on in the test.

To constrain participant-level parameter estimates, we estimate 27 parameters at the group level: six mean parameters, six standard deviations, and fifteen correlations. A multivariate normal distribution of parameters across participants is assumed.

The six mean parameters constrain the expectation for the participant-level estimates for each of the six parameters. The six standard deviations constrain how much group-level information is used in the estimation of the participant-level parameters. If standard deviations are infinitely small, all participants have the same parameter estimate, and group-level information dominates participant-level information. If standard deviations are infinitely large, all participants have entirely dissimilar parameter estimates, and participant-level information dominates group-level information. It is important to note that standard deviations are themselves estimated from the data, and are not predefined in the analysis.

Parameters may be correlated across participants. Participants who have a higher-than-average evidence accumulation for part A may also have a higher-than-average evidence accumulation for part B. Such dependencies, if they exist, may also constrain parameter estimates, allowing us to estimate participant-level parameters with more precision. Between the six parameters, there are fifteen correlations to be estimated.

Each of the group-level mean parameters are estimated with a weakly informative prior, of a normal distribution with mean 0 and standard deviation 2. For the group-level standard deviations, half-normal distribution are used, with mean 0 and standard deviation 2. The prior for the correlation matrix is a LKJ distribution with a shape parameter η of 2.

Stan model code

```
functions {
  // Shifted-Wald Likelihood function
  // Aided by https://mrunadon.github.io/Shifted-Wald-distribution-
for-response-time-data-using-R-and-Stan/
  real SW_log(real x, real gamma, real alpha, real theta){
    return log( alpha / (sqrt(2 * pi() * (pow((x - theta), 3)))) *
```

```
    exp(-1 * (pow((alpha - gamma * (x-theta)),2)/(2*(x-theta)))));
  }
}
data {
  // Number of responses
  int<lower=0> I;
  // Number of participants
  int<lower=0> N;
  // Response times, i.e. time to move from one circle to the next (in
seconds)
  vector<lower=0>[I] rt;
  // Dummy-coded variable for Part A (=1), and Part B (=0) (per
response)
  vector<lower=0>[I] partA;
  // Dummy-coded variable for Part B (=1), and Part A (=0) (per
response)
  vector<lower=0>[I] partB;
  // Participant ID (per response)
  int<lower=0> id[I];
}
parameters {
  // There are 6 latent variables,
  // - Evidence accumulation rate (gamma) during Part A
  // - Threshold (alpha) during Part A
  // - Non-decision time (theta) during Part A
  // - Evidence accumulation rate (gamma) during Part B
  // - Threshold (alpha) during Part B
  // - Non-decision time (theta) during Part B
  // Correlation matrix of size 6 by 6 for the six parameters at the
// group level
  corr_matrix[6] cormat_pars;
  // Mean vector of size 6
  vector[6] mean_pars;
  // Vector of standard deviations of size 6, bounded at 0
  vector<lower=0>[6] sd_pars;
```

```
  // Matrix of individual level parameters, of size N by 6
  vector[6] pars[N];
}

model {
  // Priors
  // Weakly informative normal priors on all group-level means
  // Although we assume all to be positive, these are unbounded.
  mean_pars[1] ~ normal( 0, 2);
  mean_pars[2] ~ normal( 0, 2);
  mean_pars[3] ~ normal( 0, 2);
  mean_pars[4] ~ normal( 0, 2);
  mean_pars[5] ~ normal( 0, 2);
  mean_pars[6] ~ normal( 0, 2);

  // Weakly informative half-normal priors on all group-level standard
deviations
  // Half-normal, because of <lower=0> in the parameter definition
  sd_pars[1] ~ normal( 0, 2);
  sd_pars[2] ~ normal( 0, 2);
  sd_pars[3] ~ normal( 0, 2);
  sd_pars[4] ~ normal( 0, 2);
  sd_pars[5] ~ normal( 0, 2);
  sd_pars[6] ~ normal( 0, 2);

  // Weakly informative prior on the group-level correlation matrix
  cormat_pars ~ lkj_corr(2);

  // Individual-level parameters come from a multivariate normal
distribution
  pars ~ multi_normal( mean_pars, quad_form_diag(cormat_pars,
sd_pars));

  // Likelihood
  for( i in 1:I){
```

```
    // Likelihood of reaction times given Shifted Wald model
    rt[i] ~ SW(
      // Likelihood for Part A is defined by parameters 1 as gamma, 2
as alpha, 3 as theta
      // Likelihood for Part B is defined by parameters 4 as gamma, 5
as alpha, 6 as theta
      (pars[id[i],1] * partA[i] + pars[id[i],4] * partB[i]),
      (pars[id[i],2] * partA[i] + pars[id[i],5] * partB[i]),
      (pars[id[i],3] * partA[i] + pars[id[i],6] * partB[i]));
  }
}
```

Convergence diagnostics

Finding good starting values for the hierarchical model was done in a two-step fashion. First, a non-hierarchical model was fitted, to find starting values for the mean parameters. These mean parameter estimates were set as starting values for the hierarchical model specified above, with standard deviations initialized close to zero. Therefore, the initialized model was de facto a non-hierarchical model. None of the Highest Posterior Density Intervals for the standard deviations included 0 for the hierarchical model, indicating that none of the chains were stuck at the initialized non-hierarchical setting.

5 chains, each with iter=20000; warmup=10000; thin=1;
post-warmup draws per chain=10000, total post-warmup draws=50000.

The highest R-hat for any parameter was 1.003, indicating that all R-hats are below the frequently used threshold of 1.1, indicating that the five chains have mixed well.

Divergences:
0 of 50000 iterations ended with a divergence.

```
Tree depth:
0 of 50000 iterations saturated the maximum tree depth of 20.


Energy:
E-BFMI indicated possible pathological behavior:
  Chain 1: E-BFMI = 0.173
  Chain 2: E-BFMI = 0.123
  Chain 3: E-BFMI = 0.192
  Chain 4: E-BFMI = 0.179
  Chain 5: E-BFMI = 0.137
```

According to http://mc-stan.org/misc/warnings.html (retrieved 8-20-2019), the E-BFMI warnings indicate an efficiency issue, rather than an issue with the substantial results.

Effect size intervals

To compute effect sizes and Highest Posterior Density Intervals, the effsize and HDInterval packages were used. Cohen's d is computed in the standard way, i.e., dividing the difference between the mean of the control group and the mean of the patient group, by the standard deviation. Because we were not considering one parameter estimate, but had 50,000 samples from the posterior distribution of parameter estimates, we could compute the uncertainty in the effect size. First, we calculated Cohen's d between the two groups, for each of the 50,000 samples from the Monte Carlo chains, for each of the six parameters. The following R code was used:

```
> cohensdmat <- matrix( NA, nrow = 50000, ncol = 6)
> for( par_no in 1:6){
>   cohensdmat[,par_no] <- sapply( 1:50000, function(x){cohen.d(
pars_samples_pat[ x,,par_no], pars_samples_con[ x,,par_no])$estimate})
>}
```

```
Second, we calculated the middle interval of most likely parameters.
To obtain the mean Cohen's d and 2.5th and 97.5th quantiles of the
Highest Posterior Density Interval, the following function was used:
> mean_hdi <- function( x ){ data.frame( y = mean(x), ymin =
hdi(x)["lower"], ymax = hdi(x, )["upper"])}
```

Supplementary Tables

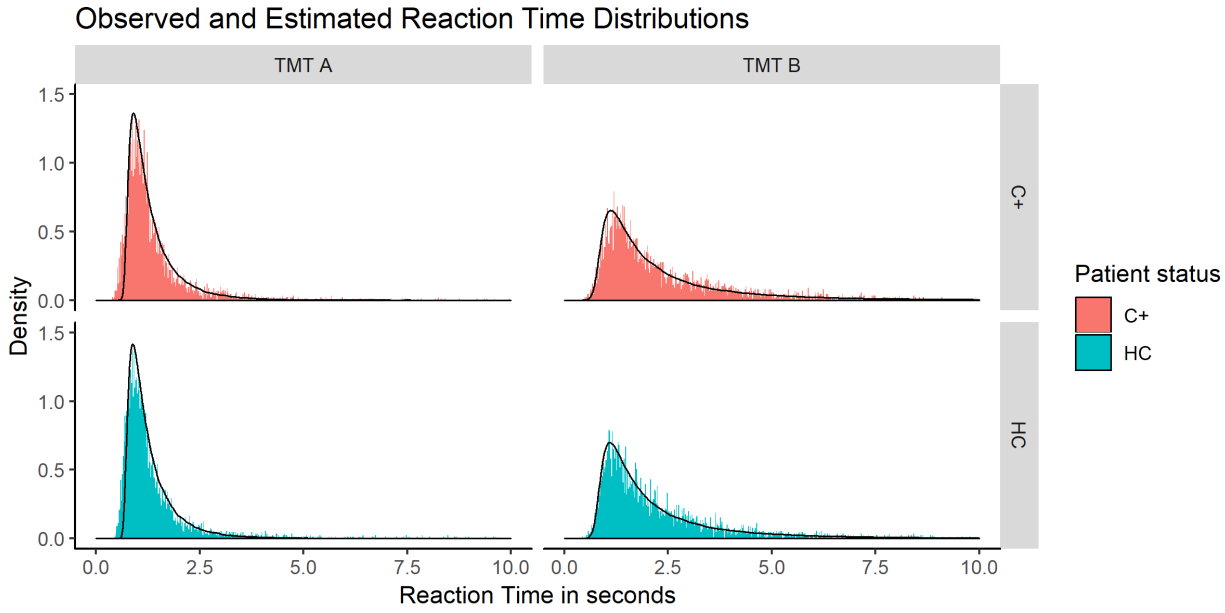
Supplementary Table 1. Effect Sizes, on Three Parameters and Parts A and B of the TMT. *Notes: TMT = Trail Making Test*

Part	Parameter	Mean Cohen's d	95% Highest Posterior Density Interval of Cohen's d
TMT A	Evidence accumulation	-0.033	-0.159 - 0.095
TMT B	Evidence accumulation	-0.162	-0.298 - -0.028
TMT A	Threshold	0.154	0.006 - 0.303
TMT B	Threshold	0.168	0.0460 - 0.290
TMT A	Non-decision time	-0.006	-0.063 - 0.051
TMT B	Non-decision time	-0.024	-0.129 - 0.080

Supplementary Table 2. Parameter Estimates, on Three Parameters and Parts A and B of the TMT. *Notes: TMT = Trail Making Test, C+ = non-CNS cancer patients, HC = controls.*

Group	Part	Parameter	Mean Estimate	95% Highest Posterior Density Interval of Estimate
HC	TMT A	Evidence accumulation	1.346	1.293 - 1.397
C+	TMT A	Evidence accumulation	1.332	1.282 - 1.384
HC	TMT B	Evidence accumulation	0.834	0.803 - 0.865
C+	TMT B	Evidence accumulation	0.804	0.775 - 0.834
HC	TMT A	Threshold	0.926	0.890 - 0.962
C+	TMT A	Threshold	0.958	0.921 - 0.998
HC	TMT B	Threshold	1.324	1.274 - 1.373
C+	TMT B	Threshold	1.380	1.329 - 1.434
HC	TMT A	Non-decision time	0.630	0.620 - 0.640
C+	TMT A	Non-decision time	0.629	0.619 - 0.639
HC	TMT B	Non-decision time	0.660	0.638 - 0.681
C+	TMT B	Non-decision time	0.655	0.631 - 0.679

Supplementary Figures



Supplementary Figure 10. Observed Reaction Time Distribution (Histogram of 18432 Reaction Times, 4608 per Panel) and Estimated Reaction Time Distribution (Density Line, Based on Median Parameter Estimates), Pooled over All Participants. *Abbreviations: TMT = Trail Making Test, C+ = non-CNS cancer patients, HC = controls.*

Note: The graph is cut off at a reaction time of 10 seconds. A very small minority of observed and estimated reaction times exceeded this cutoff.