Online Supplement for "Differential Attentional Costs of Encoding Specific and Gist Episodic Memory Representations"

Nathaniel R. Greene & Moshe Naveh-Benjamin

Racial Composition of the Sample

Participants responded to a ten-option multiple-choice question that asked, “What is your race?” Table S1 reports the proportion of participants in each attention condition who selected each possible option.

Table S1. *Racial Composition of the Sample in Each Attention Condition*

Racial Identity	FA	1-tone DA	2-tone DA	3-tone DA
American Indian or Alaska Native	0%	0%	0%	0%
Asian	6.52%	0%	2.38%	6.98%
Black/African American	4.35%	2.22%	0%	4.65%
Latinx/Hispanic	2.17%	4.44%	7.14%	4.65%
Middle Eastern	0%	0%	0%	0%
Pacific Islander	0%	0%	0%	0%
White/Caucasian	84.78%	91.11%	90.47%	81.40%
Multiracial	2.17%	2.22%	0%	2.33%
Other	0%	0%	0%	0%
Prefer not to say	0%	0%	0%	0%

Note. FA = full attention; DA = divided attention.

Sample Size Determination

Sample sizes were chosen to be on par with those of previous studies using the associative specificity recognition procedure (Greene & Naveh-Benjamin, 2020) including to assess effects of divided attention (DA) on specific and gist memory (Greene & Naveh-

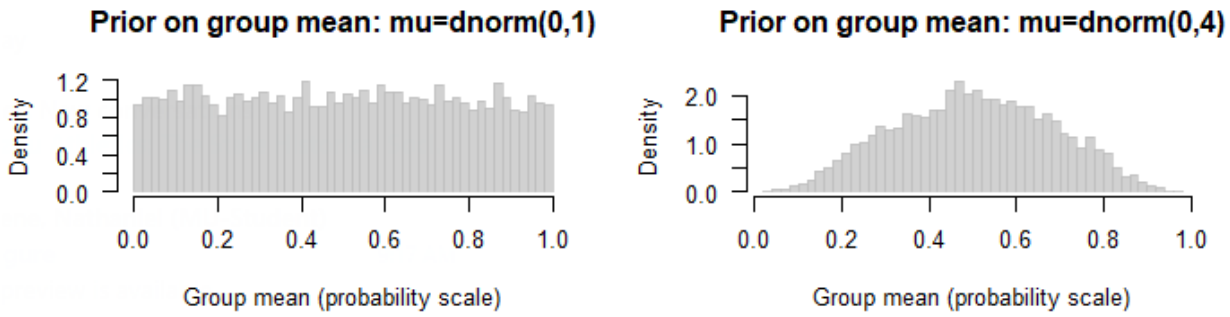
Benjamin, 2022a, 2022b; Greene et al., 2022). Specifically, these studies have traditionally relied on sample sizes of between 40 and 55 participants per condition, which are sufficiently powered to detect a medium sized between-group effect ($d = 0.70$, with variability ranging from $d = 0.60$ to $d = 0.80$) according to a Bayes Factor Design Analysis (BFDA) using optional stopping rules with a minimum sample size of 30 per group and a maximum sample size of 60 per group (see Greene et al., 2022 for full details). Using a BFDA with the aforementioned optional stopping rules and relying on a decision boundary of a Bayes Factor ($BF \geq 10$) as evidence for a between-group effect, Greene et al. (2022) showed that a median sample size of 40 participants per condition is sufficient to detect the presence of an effect of this size with an 89% true discovery rate and a 0% false negative rate. This BFDA also showed that, under the assumption that the null hypothesis was true (i.e., $d = 0$), there would be only a 1% false positive rate (erroneously detecting a credible between-group difference with a $BF \geq 3$). Thus, we aimed for samples of ~40 per group, though our sample sizes were slightly larger due to the variability in the number of participants who signed up online.

In addition, we also conducted a Bayesian prior sensitivity analysis of the multinomial processing tree (MPT) model, which assesses whether, at a given sample size (the sample sizes reported in the manuscript), if we adjusted the amount of information conveyed in the prior distribution, if this would influence the posterior estimates of the obtained parameters. This provides a robustness test of the parameter estimates from the MPT models reported in the main text because it enables us to assess whether comparable parameter estimates were obtained both under weakly informative prior specifications and under more strongly informative prior specifications. If so, then this tells us that, even if we included a lot more certainty in our *prior belief* as to what the true value of a parameter should be once conditioned on the data, the data

clearly played an important role in determining what those values were. This is not to say that the prior sensitivity analysis indicates the sample size was sensitive to detecting a difference in a given parameter, per se, but rather it can show that the obtained parameter estimates at our selected sample size were robust to different levels of informativeness of the prior distribution.

To conduct the prior sensitivity analysis, we ran the MPT models with more informative priors specified on the group means of the parameters (more information about prior specifications is listed in the technical section that follows). For each parameter *s*, the weakly informative prior (reported in the main text) was a standard normal distribution on each group mean μ_s, which implies a uniform distribution in probability space (Rouder & Lu, 2005). For our prior sensitivity analysis, we also specified our models with a more informative Normal(0, 4) prior specified for each μ_s, which draws parameters closer to the mid-point of the probability scale and places lower prior expectation on extreme values (see Figure S1).

Figure S1. *Different Prior Specifications on Group Means and Their Effects on the Probability Scale*



Note. Left side: Assuming each μ_s follows a standard normal prior, as done in the models reported in the text. Right side: Assuming each μ_s follows a more informative Normal(0,4) prior, which places lower

expectation on extreme values of a parameter (near 0 or 1) and higher prior belief in values of the parameter nearer the midpoint of the probability scale.

Table S2 lists the MPT estimates obtained under the more informative prior specification. As listed, these estimates are almost identical to those obtained under the weakly informative specification reported in the main text (see Table 2 in the main text), showing that the parameter estimates obtained were robust to how much certainty was conveyed in the priors.

Table S2. *Population-level parameter estimates [95% credible intervals] of the MPT model with More Informative Prior Distribution*

Parameter	FA	DA 1-tone	DA 2-tone	DA 3-tone
V_i	0.61 [0.49, 0.70]	0.53 [0.45, 0.60]	0.43 [0.32, 0.53]	0.42 [0.30, 0.54]
V_r	0.12 [0.04, 0.22]	0.08 [0.03, 0.16]	0*	0.05 [0.01, 0.11]
G_i	0.48 [0.39, 0.56]	0.52 [0.44, 0.59]	0.53 [0.45, 0.60]	0.48 [0.39, 0.56]
G_r	0.69 [0.60, 0.77]	0.62 [0.53, 0.70]	0.63 [0.55, 0.70]	0.58 [0.49, 0.66]
F	0.04 [0.01, 0.09]	0.07 [0.02, 0.12]	0.07 [0.02, 0.12]	0.08 [0.03, 0.13]
a	0.38 [0.33, 0.43]	0.32 [0.28, 0.37]	0.43 [0.37, 0.51]	0.42 [0.38, 0.47]
ab	**	**	0.28 [0.21, 0.35]	**
b	0.26 [0.19, 0.33]	0.28 [0.24, 0.33]	0.28 [0.23, 0.33]	0.33 [0.26, 0.41]

Note. FA = full attention; DA = divided attention. V_i = probability of specific/verbatim retrieval given an Intact probe; V_r = probability of specific/verbatim retrieval given a Related probe; G_i = probability of gist retrieval given an Intact probe; G_r = probability of gist retrieval given a Related probe; F = probability of fuzzy retrieval given an Unrelated-Within probe. a = probability of guessing "intact" when gist is retrieved. a_b = probability of guessing "intact" in non-gist retrieval states; b = probability of responding "intact/related" when there is no verbatim or gist information. *V_r was constrained to 0 in the 2-tone DA condition due to initial model misfit for model reported in main text. **Parameters a and a_b were set to

equality in the FA, 1-tone DA, and 3-tone DA conditions due to initial model misfit for model reported in main text.

Multinomial Processing Tree (MPT) Model Sampling Routines and Diagnostics

The MPT model was fitted to the data in each Attention condition separately using a hierarchical Bayesian estimation method with the TreeBUGS package for R (Heck et al., 2018; R Core Team, 2022). We used the latent-trait specification (see Klauer, 2010 for details), retaining the program's default (weakly informative) priors. Under the latent-trait specification, the parameter vector of participant *j* is inverse-normal transformed as $\Phi^{-1}(\theta_j)$ and follows a multivariate normal distribution, with hyperpriors on the group-level distribution for the mean for each parameter *s*, $\boldsymbol{\mu}_s$, and covariance matrix $\boldsymbol{\Sigma}$**.** Using the default priors of the TreeBUGS package, each $\boldsymbol{\mu}_s$ had a Normal(0,1) prior, which corresponds to a uniform distribution in probability space. For $\boldsymbol{\Sigma}$, a scaled inverse Wishart prior was specified, with an identity scale matrix of size *s* x *s* with *s* + 1 degrees of freedom, with *s* corresponding to the number of parameters in the model (8). Scaling parameters ξ_s with a uniform prior on the interval [0, 10] were placed on the standard deviations of each parameter to ensure the inverse Wishart prior was only weakly informative.

The posterior distributions of the model parameters were estimated from 3 independent Markov chain Monte Carlo (MCMC) chains, each with 65,000 iterations. The first 15,000 iterations were adaptation iterations, and the next 10,000 were burn-in iterations. Thus, the first 25,000 iterations per chain were discarded. Every 10^{th} iteration was retained thereafter to reduce autocorrelation and enhance the sampling efficiency (i.e., processing time to run the model), resulting in a total of 19,500 iterations from the posterior distribution in the final estimation of each parameter. Chain convergence was assessed via the $\hat{R}$ statistic, which was < 1.03 for all

parameters in the final model that best fit the data in each condition (see further details below), indicating the chains converged on a stable posterior distribution. We also considered the effective sample size (ESS) of each parameter, which indicates the number of iterations that were actually retained in the estimation of each parameter. Given a high degree of covariation in the model parameters, which is an assumption of the latent-trait specification, there is no clear cut rule for what determines a good ESS, but we considered ESS < 200 to indicate insufficient estimation of a parameter as ESS < 200 typically results in large credible intervals of a given parameter that indicate the model was unable to find the most likely estimate of the parameter under such a small ESS.

Model fit was evaluated via posterior predictive checks, in which the posterior distribution of the model parameters was used to simulate 1,000 new datasets, and the correspondence between posterior-predicted and observed means and covariances was computed via the T_1 and T_2 statistics, respectively (Klauer, 2010). A posterior predictive *p* (PPP) value was computed for each statistic. Model fit is considered satisfactory when PPP > .05. Initially, in each condition, the models misfit the data (multiple PPP < .05) and resulted in untenable parameter estimates (several instances where parameters' ESS < 200) and high values of the $\hat{R}$ statistic exceeding 1.05 for several parameters, suggesting the chains could not converge on a stable posterior distribution. In the FA, 1-tone DA, and 3-tone DA conditions, this was primarily driven by unstable parameter estimates of parameters a and a_b (ESS < 200, $\hat{R}$ > 1.05), the two guessing "intact" parameters that model, respectively, the tendency to guess "intact" when gist memory is retrieved (a) versus when gist memory is not retrieved but the individual still elects to guess either "intact" or "related" (a_b). Essentially, the estimates of these parameters overlapped substantially in each of these three Attention conditions and their estimation was imprecise

(resulting in large credible intervals that spanned most of the probability range from 0 to 1). This suggests that assuming two separate guessing "intact" values was unnecessary, so in these conditions, we fit a constrained model in which $a = a_b$ (i.e., assuming only one tendency to guess "intact" that is common both when gist, but not specific, representations are retrieved and when neither type of representation is retrieved). This adjustment substantially improved model fit (all $PPP \geq .389$) and resulted in stable effective sample sizes (all ESS > 400) and $\hat{R}$ values (all < 1.03).

In the 2-tone DA condition, constraining the two guessing "intact" parameters to equality still resulted in a poor fit to the data ($PPP < .05$). An examination of posterior-predicted compared to observed frequencies of responses indicated that the model *overestimated* the proportion of correct "related" responses to Related probes in the 2-tone DA condition. Estimates of parameter V_r (specific retrieval of the original pair when shown a Related pair) were highly concentrated near 0 in this condition, which can be problematic for the model (see Greene & Naveh-Benjamin, 2022c) because the model attempts to estimate some slightly nonzero value of this parameter, which in turn results in an overestimation of correct "related" responses to Related probes. To adjust for this, in a final fit of the model to the 2-tone condition, we constrained parameter V_r to 0 but retained the two separate guessing "intact" parameters (i.e., we allowed parameters a and a_b to vary). Doing so substantially improved the fit of the model ($PPP \geq .350$) and resulted in stable parameter estimates (all ESS > 500, all $\hat{R} < 1.03$).

Effects of Divided Attention on Response Bias Parameters

Table S3 lists the difference scores comparing each DA condition to the FA condition (FA minus DA) for the response bias parameters a (probability of guessing "intact") and b (probability of guessing either "intact" or "related"). The only parameter that credibly differed in

>95% of posterior samples between FA and DA was parameter *b* for the 3-tone DA compared to the FA condition. Estimates of *b* were higher under the 3-tone DA condition, indicating that participants in this condition were more inclined to guess that a probe was "intact" or "related" even in the absence of specific or gist memory retrieval.

Table S3. *Difference Scores for the Response Bias Parameters Obtained from Subtracting Posterior Samples of the DA Conditions from the FA Condition*

Parameter	FA – DA 1-tone	FA – DA 2-tone	FA – DA 3-tone
a	0.05 [-0.02, 0.13]	-0.07 [-0.15, 0.02]	-0.05 [-0.13, 0.02]
b	-0.03 [-0.11, 0.05]	-0.02 [-0.10, 0.07]	**-0.08 [-0.19, 0.02]**

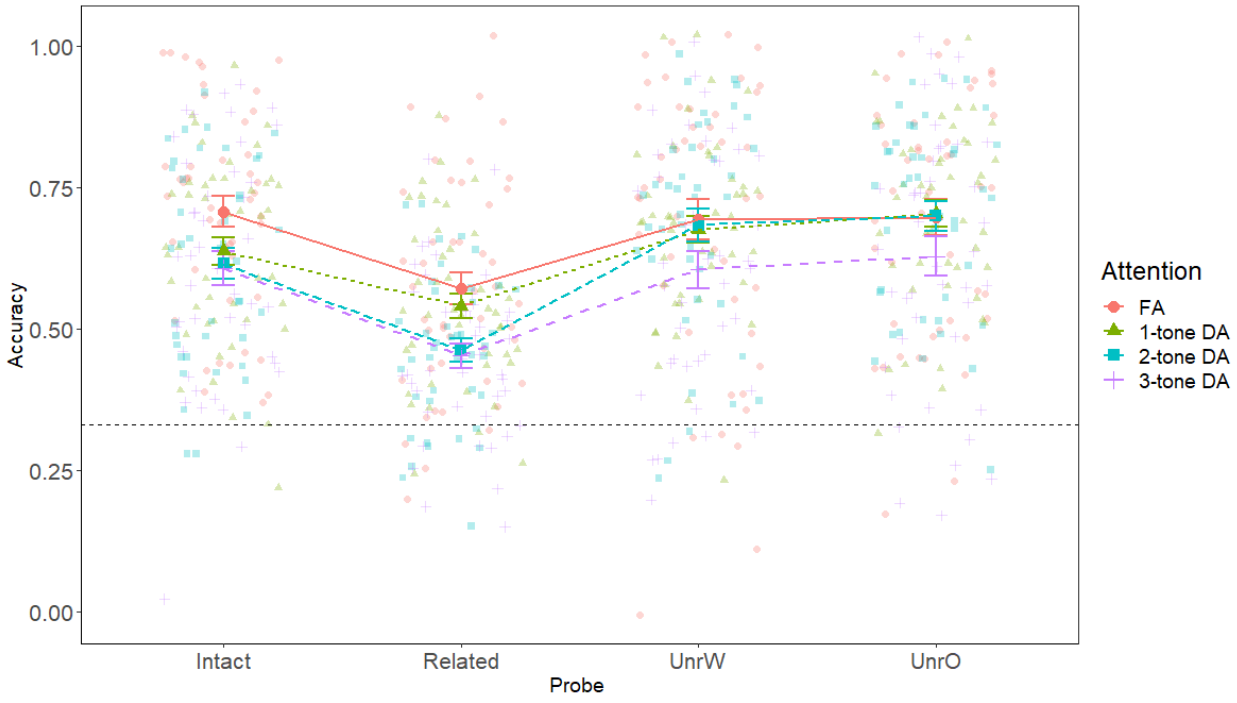
Note. Difference scores obtained by subtracting posterior samples of each divided attention (DA) condition from the full attention (FA) condition. Bolded difference scores correspond to a credible difference between FA and DA in >95% of posterior samples. *a* = probability of guessing "intact" (in the full model, a separate parameter, *ab*, denotes the probability of guessing "intact" in cognitive state *b*, i.e., when gist memory is not retrieved, but these response bias parameters were set to equality in each condition with the exception of the 2-tone DA condition). *b* = probability of guessing either "intact" or "related" when no specific or gist memory is retrieved or the probe does not match specific or gist representations in memory

ANOVA on Proportion of Correct Responses

The proportion of correct responses (correct "intact" responses to Intact probes, "related" responses to Related probes, and "unrelated" responses to Unrelated-Within and Unrelated-Opposite probes) as a function of attention condition is depicted in Figure S2. These data were treated to a 4 (Probe) x 4 (Attention) mixed ANOVA, using both frequentist and Bayesian alternatives in JASP (JASP Team, 2020). In the Bayesian statistical framework, the strength of

evidence for or against each main effect and interaction is given by a Bayes factor, where BF_{10} describes the Bayes factor in favor of an effect. Using the nomenclature of van Doorn et al. (2021), we deemed the evidence for an effect to be weak when BF_{10} fell between 1 and 3, moderate when BF_{10} was between 3 and 10, strong when BF_{10} was between 10 and 30, and decisive when BF_{10} exceeded 30. Conversely, we deemed the evidence *against* an effect or interaction (i.e., in favor of the null) to be weak when BF_{10} was between 1/3 and 1, moderate when BF_{10} was between 1/10 and 1/3, strong when BF_{10} was between 1/30 and 1/10, and decisive when BF_{10} was less than 1/30. Default priors based on recommendations from Rouder et al. (2012) were used for the Bayesian ANOVA.

Figure S2. *Proportion of Correct Responses to Each Probe*



Note. Group means are illustrated as bolded red circles (for full attention (FA) condition), bolded green triangles (for 1-tone divided attention (DA) condition), bolded blue squares (for 2-tone DA condition), and bolded purple crosses (for 3-tone DA condition), along with standard error (+/- 1 SE). Individual

participants' data are shown as jittered circles, triangles, squares, or crosses. Dashed horizontal line indicates chance level accuracy (33% correct), but because both "intact" or "related" responses to Intact, Related, and Unrelated-Within probes could reflect judgments made on the basis of retrieving some level of representation about an original episode, even performance near 33% correct can still be informative in this paradigm. UnrW = Unrelated-Within; UnrO = Unrelated-Opposite.

The ANOVA revealed decisive evidence for a main effect of Probe, $F(3, 516) = 73.82$, $p < .001$, $BF_{10} = \infty$. Post-hoc pairwise comparisons showed that proportion correct differed among all probes (all $p_{Holm} \leq .007$, all $BF_{10} \geq 6.91$), with two exceptions: there was no significant difference in proportion correct to Intact and Unrelated-Within probes, $t = -1.71$, $p_{Holm} = .177$, $BF_{10} = 0.23$; nor between Unrelated-Within and Unrelated-Opposite probes, $t = -1.36$, $p_{Holm} = .177$, $BF_{10} = 0.34$. However, accuracy was higher for Unrelated-Opposite than Intact probes, and for all probes relative to Related probes.

There was also a main effect of Attention, $F(3, 172) = 3.08$, $p = .029$, $BF_{10} = 1.23$, though the Bayes factor is only weakly in favor of the effect. Follow-up pairwise comparisons revealed that there was a significant difference in performance between the full attention (FA) and 3-tone DA conditions, $t = -2.94$, $p_{Holm} = .023$, $BF_{10} = 446.09$, with higher accuracy under FA. The Bayes factor also suggested a credible difference in performance between 1-tone DA and 3-tone DA conditions, though this difference was not significant under the frequentist approach, $t = 2.04$, $p_{Holm} = .213$, $BF_{10} = 18.79$.

Finally, the evidence for the Probe by Attention interaction differed under the two statistical frameworks, $F(9, 516) = 1.99$, $p = .039$, $BF_{10} = 0.60$, as the Bayes factor suggests weak evidence against the interaction. Nevertheless, we conducted separate one-way ANOVAs to assess the main effect of Attention for each Probe separately. For Intact probes, there was a

significant effect of Attention, $F(3, 172) = 2.88$, $p = .038$, $BF_{10} = 0.94$, though again the Bayes factor was inconclusive and slightly favored a null effect. Post-hoc tests revealed a significant difference in response accuracy to Intact probes between the FA and 3-tone DA conditions, $t = -2.62$, $p_{Tukey} = .046$, $BF_{10} = 3.11$. For Related probes, there was a significant effect of Attention, $F(3, 172) = 6.28$, $p < .001$, $BF_{10} = 54.90$. Post-hoc tests revealed significant or credible differences for all Attention conditions with the following exceptions: proportion correct to Related probes did not significantly differ between the FA and 1-tone DA conditions, $t = -0.97$ $p_{Tukey} = .767$, $BF_{10} = 0.31$, nor between the 2-tone and 3-tone DA conditions, $t = 0.30$, $p_{Tukey} = .991$, $BF_{10} = 0.24$. For Unrelated-Within probes, there was no main effect of Attention, $F(3, 172) = 1.66$, $p = .178$, $BF_{10} = 0.21$. However, an extreme group comparison (comparing just the FA to 3-tone DA conditions) revealed a marginally significant difference between these two conditions, $t(87) = -1.79$, $p = .077$, $BF_{10} = .90$, though the Bayes factor is weakly in favor of the null. Similarly, for Unrelated-Opposite probes, there was no main effect of Attention, $F(3, 172) = 1.48$, $p = .221$, $BF_{10} = 0.17$.

To summarize, response accuracy was generally poorest for Related probes, while performance to Intact, Unrelated-Within, and Unrelated-Opposite probes was mostly on par with some slight differences (i.e., performance to Unrelated-Opposite probes was superior to performance to Intact probes). Also, overall, effects of DA relative to FA were most pronounced under the most demanding level of the DA task (i.e., the 3-tone condition). However, there were some subtle differences in the effects of DA based on the type of probe. Whereas participants in the 3-tone DA condition performed significantly worse than participants in the FA condition on both Intact and Related probes, they did not significantly differ in correctly classifying Unrelated-Within and Unrelated-Opposite probes. Also, for Related probes, participants in the 2-

tone DA condition were less accurate at classifying these probes than were participants in the FA condition. There were no significant or credible differences in response accuracy between the FA and 1-tone DA conditions. Consequently, the effects of DA, relative to FA, on task performance in the associative specificity recognition task appear to emerge in a rather graded fashion, with no discernible effects under light load (1-tone DA); effects on those probes requiring individuals to remember the most specific information (i.e., Related probes) under intermediate loads (2-tone DA); and effects on both Related probes and Intact probes, which provide good retrieval cues for specific representations as they are an exact reproduction of an original pair, under heavy loads (3-tone DA).

Effects of DA compared to FA in correctly classifying the two types of Unrelated probes were, at best, weakly present (for Unrelated-Within category probes) and not significant (for Unrelated-Opposite category probes). To correctly classify these probes (especially Unrelated-Opposite probes), retrieval of any amount of specific or general representation of the original association may be sufficient, given the categorical mismatch with the originally encoded association at both specific and general/gist levels of representation. Thus, the lack of a significant effect of DA in classifying these probes may suggest that the disrupting effects of DA did not extend to gist representations. However, an analysis of proportion of correct responses, as we have done here, cannot unearth the underlying cognitive processes that may differ between FA and DA. An individual can correctly endorse a probe as being “intact,” “related,” or “unrelated” based on retrieval of an underlying memory representation or due to a guessing proclivity/bias to respond in a certain way (e.g., to endorse probes as being “unrelated”) even in the absence of specific or gist memory retrieval. ANOVA on proportion correct reveals nothing about these underlying processes, but the MPT results (reported in the main text) provide deeper

insights that there were effects of DA on specific representations (that emerged under the immediate 2-tone load) and on both specific and gist representations (that emerged under the difficult 3-tone load).

Divided Attention Concurrent Task Results

We also analyzed performance differences in terms of reaction time (RT) and accuracy (% correct) on the concurrent DA tone task in both the baseline phase and during the study blocks of the experiment (see Table S4 for RTs). For RT, we ran two one-way ANOVAs to test for a main effect of DA load (1-tone, 2-tone, or 3-tone) on (1) baseline RT and (2) study phase RT. For the baseline RT, there was a significant main effect of DA load, $F(2, 127) = 110.25$, $p < .001$, $BF_{10} = \infty$. Post-hoc paired-samples *t*-tests showed that RT during the baseline period significantly differed among all conditions, with RT increasing with each successive difficulty level of the concurrent task (all $p_{Tukey} < .001$, $BF_{10} \geq 1.57 \times 10^4$). A similar pattern of results was obtained for the study phase RT, with $F(2, 127) = 124.45$, $p < .001$, $BF_{10} = \infty$, all post-hoc $p_{Tukey} < .001$, $BF_{10} \geq 4.63 \times 10^7$.

Table S4. *Mean (SD) Reaction Times During Baseline and Study Phases for the Divided Attention (DA) Concurrent Tone Task*

	Baseline Period	Study Phase
1-tone DA	404.25 (117.10) ms	475.46 (137.51) ms
2-tone DA	528.35 (99.57) ms	691.15 (139.10) ms
3-tone DA	787.19 (147.39) ms	940.77 (138.54) ms

For accuracy, which was at ceiling for the 1-tone task (as this was a simple RT task), we did not include the 1-tone condition in the analysis. Instead, we compared the proportion of correct tone classifications in the choice RT tasks (2-tone vs 3-tone) with two independent samples *t*-tests comparing accuracy during baseline and during the study phases of the experiment. There was a significant effect of DA load on baseline accuracy, $t(83) = 3.52$, $p < .001$, $BF_{10} = 41.04$, which was higher in the 2-tone ($M = 0.97$, $SE = 0.01$) than the 3-tone ($M = 0.92$, $SE = 0.01$) condition. However, there was not a significant difference between the 2-tone ($M = 0.93$, $SE = 0.02$) and 3-tone ($M = 0.89$, $SE = 0.01$) conditions on study phase accuracy, $t(83) = 1.87$, $p = .064$, $BF_{10} = 1.04$.