The Morph-Flanker and the Brightness-Illusion Task: Two New Tasks to Assess Inhibition?

Michelle C. Donzallaz

Research Master's Psychology

Major: Psychological Methods

University of Amsterdam

Author Note

Supervisors: Dr. Julia M. Haaf and Dr. Jeffrey N. Rouder

Second assessor: Dr. Dora Matzke

This thesis was written in R Markdown. All analyses are integrated into the text. The R Markdown source code including links to the datasets is available at https://github.com/PerceptionAndCognitionLab/ctx-contrast.

Correspondence concerning this article should be addressed to Michelle C. Donzallaz, . E-mail: michelle.donzallaz@student.uva.nl

Abstract

Despite many attempts, examining individual differences in cognitive inhibition has proven difficult: performance across tasks purportedly assessing inhibition correlates weakly at best. However, it remains unclear whether the task performances truly are unrelated. Simulation studies suggest that correlations are obscured by measurement error and latent individual differences may be too small to be detected. To examine the psychometric structure of cognitive inhibition, tasks with a higher resolution to detect correlations are needed. One proposed solution is to design new inhibition tasks with more individual variation. In practice, this requires developing paradigms with sizable overall effects. In this thesis, we introduce two new potential inhibition tasks: the morph-flanker task and the brightness-illusion task. Both tasks are designed to assess the ability to inhibit contrastive information. First, we present the results of a small pilot study that we conducted online among university students. We analyzed the data using a hierarchical trial-level model that explicitly models the covariation between the two tasks. The results showed that only the morph-flanker but not the brightness-illusion task yielded a substantial overall effect. Yet, both tasks led to sizable individual differences and performance among the two tasks was highly positively correlated. We discuss this unexpected yet promising finding and how the task designs may be improved. In the second part of this thesis, we propose a large-scale correlation study to further assess the proposed tasks. Specifically, we suggest to explore how they are related to other established cognitive tasks, namely a regular flanker, a working memory, and a visual inspection time task.

Keywords: inhibition, individual differences, cognitive tasks, Bayesian hierarchical modeling

The Morph-Flanker and the Brightness-Illusion Task: Two New Tasks to Assess Inhibition?

The ability to ignore irrelevant information in one's environment is essential for everyday cognitive functioning (e.g., Cowan, 1995). Moreover, this ability to inhibit has been linked to psychologically relevant outcomes, such as impulsive behaviors (Sharma, Markon, & Clark, 2014) and psychopathology (e.g., Joormann, Yoon, & Zetsche, 2007; Nigg, 2000; von Bastian et al., 2020).

Cognitive inhibition is commonly assessed using experimental tasks. Two prominent examples of cognitive inhibition tasks are the flanker (Eriksen & Eriksen, 1974) and the Stroop task (Stroop, 1935). In the letter-flanker task, a common version of the flanker paradigm, individuals are asked to identify a centrally located letter (i.e., the target) surrounded by other letters (i.e., the distractors). There are two types of trials: in congruent trials, the target-letter and the distractor-letters are the same, and in incongruent trials, they differ. The assumption is that it takes longer for individuals to respond in incongruent than in congruent trials and that this difference in response time (RT) captures one's ability to inhibit. Accordingly, the flanker effect is the RT difference between trials where the target- and distractor-letters differ (i.e., incongruent) and trials where they are identical (i.e., congruent; Figure 1A). In the classic Stroop task, the color in which a color-word (e.g., "blue") is printed needs to be indicated. Here, the target is the print color, and the distractor is the meaning of the color-word. The Stroop effect then is the RT difference between trials where the print color and the meaning of the word are different (i.e., incongruent; e.g., "red" written in blue) and trials where they match (i.e., congruent; e.g., "red" written in red).

There are good reasons why researchers use these paradigms to assess inhibition (Rouder & Haaf, 2019). For example, they generally yield well-replicable effects — the Stroop effect is even seen as universal (Haaf & Rouder, 2017; MacLeod, 1991). Implicitly assuming that robust experimental effects are also well suited to assess individual differences, inhibition tasks are often used in correlational studies (Hedge, Powell, & Sumner, 2018). It

seems reasonable to assume that someone who does well in the Stroop task is also good at the flanker task — after all, they aim to assess the same concept. However, performance in inhibition tasks has generally been found to correlate weakly (e.g., Keye, Wilhelm, Oberauer, & van Ravenzwaaij, 2009; Paap & Greenberg, 2013; Pettigrew & Martin, 2014; Rey-Mermet, Gade, & Oberauer, 2018). This suggests that the robustness of experimental effects does not transfer as easily to individual difference studies as one might expect.

Explanations for the low correlations have been substantive as well as statistical (Rouder & Haaf, 2019). A substantive account is that the tasks capture unrelated mental processes (e.g., Rey-Mermet et al., 2018). Accordingly, inhibition should not be treated as a unified construct and performance not generalized beyond the specific task. Ultimately, if cognitive performance across standard inhibition tasks truly is unrelated, it is questionable whether cognitive inhibition as a general ability even exists (Rey-Mermet et al., 2018).

A statistical explanation for the low task correlations is that correlations are obscured by trial noise and latent individual differences too small to be detected. It is well established that observed correlations are attenuated by measurement error (e.g., Matzke et al., 2017; Spearman, 1904). In inhibition tasks, individuals only complete a certain, finite number of trials. Even if there are individual differences and the tasks truly are related, the correlations may not be detectable due to a large degree of trial noise (Rouder & Haaf, 2019). In fact, there seems to be around seven times more trial noise than true individual variation in inhibition tasks (Rouder et al., 2019).

The extent of true individual variation may be restricted by the size of the overall inhibition effects which tend to be rather small. For example, the average Stroop effect is around 50 milliseconds (Rouder & Haaf, 2019). It can be assumed that everybody is slower in incongruent than congruent trials – nobody truly has a negative Stroop effect (Haaf & Rouder, 2017). That is, we can assume that everyone requires more time to name the print color when the color and the semantic meaning of the word are different than when they match. If this assumption holds, the degree of possible true individual variation is dependent

on the size of the overall effect. With only little latent variation, there cannot be much covariation across different tasks. Rouder et al. (2019) showed that current experimental designs when using a realistic number of trials do not have the resolution to detect this small extent of latent variation and covariation; even when employing advanced statistical methods such as hierarchical modeling which take trial noise into account. Simulations revealed that to recover correlations between standard inhibition tasks such as the Stroop and flanker, over a thousand trials per task would be needed (Rouder et al., 2019).

Assuming inhibition as a distinct cognitive ability exists, a fundamental question is whether it is a unitary phenomenon or whether it comprises different kinds of phenomena. Often inhibition has been conceptualized as a unified construct (e.g., Hasher & Zacks, 1988). But various researchers have suggested distinct conceptual processes, comprising two or more domain general factors (e.g., Friedman & Miyake, 2004; Stahl et al., 2014; for a review, see Rey-Mermet et al., 2018). The question then is whether there is shared variance among those processes, that is whether an overarching cognitive inhibition construct can be established. A natural approach to answer this question is to examine whether performance across tasks assessing different inhibitory processes correlates. At this point, findings are inconclusive, and no substantive conclusions regarding the psychometric structure of inhibition can be drawn (Draheim, Mashburn, Martin, & Engle, 2019).

Given the difficulties in obtaining correlations and the uncertainty about the construct "inhibition", several researchers have called for new tasks assessing inhibitory processes (e.g., Draheim et al., 2019; Friedman & Miyake, 2004). Particularly tasks with a resolution high enough to detect correlations among them – and consequently individual variation – are needed to answer long-standing questions about the concept of inhibition. Assuming that everybody has inhibition effects in the same direction as in tasks such as the Stroop and flanker, a better inhibition task has to be one that shows a large overall effect and a high signal-to-noise ratio (Rouder et al., 2019).

The Morph-Flanker and the Brightness-Illusion Task

Here we present two new tasks that we think have the potential to yield large inhibition effects and individual differences. Both tasks aim to assess the ability to inhibit contrastive information. The first one is based on a paradigm introduced by Snyder, Rafferty, Haaf, and Rouder (2019; see also Rouder and King, 2003). These authors adapted the letter-flanker task such that the targets were A-to-H-morphs instead of ordinary A and H letters (Figure 1B). The effect assessed in this task was perceptual and is best described as a contrastive context effect – the distractors nudge the response in the opposite direction: participants were more likely to perceive a morph as "A" when the distractors were "H"s but more likely to perceive it as "H" when the distractors were "A"s (Figure 1B; Snyder et al., 2019). This paradigm yielded larger overall effects than usual and substantial variation across participants.

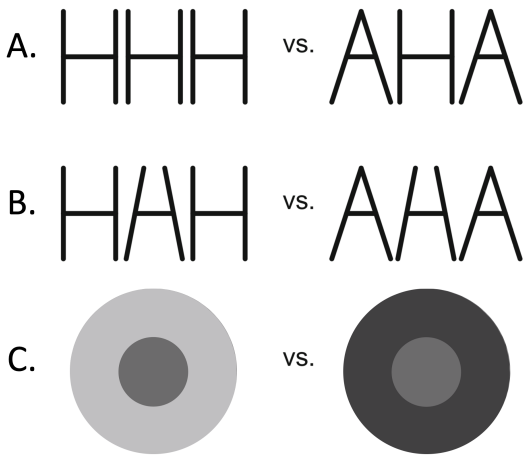

Figure 1. Original and modified flanker paradigms (adapted from Snyder et al., 2019, p. 1945). **A.** reflects the letter-flanker by Eriksen and Eriksen (1974). **B.** reflects the paradigm assessing contrast effects based on Snyder and colleagues (2019) and Rouder and King (2003). **C.** reflects the brightness-illusion paradigm.

However, Snyder et al. (2019) critically noted that demand characteristics could have influenced their results. There were no objectively right or wrong answers regarding the

identity of the target-morphs. This could have led some participants to more or less explicitly use the distractors to inform their responses — even though they were instructed not to — as the distractors were the only available information that they could base their judgement on.

Addressing this issue, we advanced the paradigm developed by Snyder et al. (2019) to what we call the *morph-flanker task*. Perhaps the biggest change is that in our version, participants receive feedback on their performance based on task criteria that determines which morphs are "A"s and which "H"s. Using a range of A-to-H-morphs, participants are trained and instructed to respond with "A" to one half of the morphs and with "H" to the other. One can then examine how effectively individuals ignore contrastive information by comparing their responses across distractor-contexts.

The second task builds on the same logic as the morph-flanker, but we use brightness stimuli consisting of an inner circle, the target, and an outer ring, the distractor (Figure 1C). We call this task the *brightness-illusion task*. The distractors are either of a particularly light or dark grey (Figure 1C), and the targets differ in their brightness. Participants first learn whether the targets fall into a "dark" or "bright" category. They then judge whether the targets are "bright" or "dark" with the distractors around them. By using two different types of stimuli in this adapted version of the flanker task, we can examine how the ability to inhibit contrastive information compares across (morph)-letter identification and brightness judgements.

There are two main aspects in which the morph-flanker and the brightness-illusion task differ from standard ones (e.g., Stroop and flanker task) and that lead us to believe that they will be more discriminative of individuals' inhibitory abilities. The first concerns the nature of the inhibitory effect we are trying to isolate and the second the metric used to quantify it.

First, the morph-flanker and the brightness-illusion task evoke the exact opposite response patterns from standard inhibition tasks, and these have shown to yield more individual variation (Snyder et al., 2019). In classic inhibition tasks, individuals are tempted to align their response with the identity of the distractors, a mechanism described as

assimilation (Snyder et al., 2019). Consequently, it generally takes longer to respond in incongurent than in congruent trials. For example, in the Stroop task, people are inclined to name the meaning of the color word (i.e., the distractor) instead of the print color. And in the letter-flanker task, the temptation is to use the letters surrounding the target letter to identify the target. In the tasks we introduce here, on the other hand, the assumption is that individuals are inclined to respond contrary to the distractors' identity. Besides the contrast task, Snyder et al. (2019) also developed an assimilation task involving morphs. In this task, the morph-targets ("A"-to-"H") were surrounded either by the distractor "C_T" or by "T_E" (the underscore is a placeholder for the target). Participants were asked to indicate whether they perceived the morph as "A" or as "H". This task yielded substantial assimilation effects and performance in the two tasks was substantially correlated, which suggests that there were individual differences. Because the contrast task yielded greater individual variation than the assimilation task, we believe that the former generally yields more individual differences than the latter.

The second major difference between classic inhibition tasks and the novel tasks introduced here is that the morph-flanker and the brightness-illusion tasks are based on accuracy data whereas the most prominent inhibition tasks quantify inhibition based on RT. RTs, specifically RT difference scores have been criticized for being susceptible to unreliability (e.g., Draheim et al., 2019). Indeed, as already mentioned, the signal-to-noise-ratio in standard inhibition tasks is rather small ($\approx 1/7$; Rouder et al., 2019). This ratio may be more favorable for accuracy data, and therefore for the introduced tasks.

Overview of Studies

This thesis examines the morph-flanker and the brightness-illusion task as two novel tasks for assessing the ability to inhibit contrastive information. To this aim, we first present results of a pilot study examining the morph-flanker and the brightness-illusion task. We asked whether the morph-flanker and/or the brightness-illusion task yield(s) substantial

inhibition effects and individual variation. We hypothesized that there are sizable overall effects (hypotheses 1 and 2) and individual differences (hypotheses 3 and 4) in both proposed tasks. Since both tasks were designed to assess the same concept, we also expected the contrast effects to be positively correlated (hypothesis 5).

In the second part of this thesis, we propose a correlation study involving the task that yields more individual variation. This study has yet to be conducted. We plan to only include one of the new tasks because we will administer more trials per task than are commonly used to decrease trial noise, as recommended by Rouder et al. (2019). The aim of this study will be to explore whether the novel inhibition task correlates with other established paradigms assessing cognitive ability. Specifically, we will include two tasks known to yield large individual differences: a working memory task and a visual inspection time task. Besides these two tasks, we will also include the regular letter-flanker task as established inhibition task in the battery. Performance in this task will serve as a reference point for the extent of correlations and the size of the inhibition effect commonly found in inhibition research.

(Pilot) Study 1

Procedure

Participants completed the morph-flanker and the brightness-illusion task online on their own computer. Before the start of the experiment, they gave informed consent and answered a few demographic questions. Each task comprised four blocks of eighty trials and took approximately thirty minutes to complete. Participants were encouraged to take breaks between different blocks and after the first task. All participants completed first the brightness-illusion task followed by the morph-flanker task. The tasks were presented in this order because we wanted to ensure the data quality for the brightness-illusion task in the event of decreasing motivation over time. This task was more novel than the morph-flanker

in the sense that it used new stimuli that had never been administered before. Whereas the stimuli used in the morph-flanker were a subset of the stimuli used in the paper by Snyder et al. (2019). Data were born open (Rouder, 2016), which means that data were automatically archived on GitHub the night they had been collected[1].

Participants

Participants were undergraduate psychology students at the University of California, Irvine. They received course credit for their participation. A total of N =27 completed the brightness-illusion task and N = 25 completed the morph-flanker task which means that two participants decided to stop participating after the brightness-illusion task.

Materials

Demographics. Participants indicated their age, gender, handedness, and ethnicity.

Morph-flanker and brightness-illusion task. The tasks were programed using lab.js, a free, html-based, online study builder (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2019). Both tasks had a 2x8 within-subjects design. The first factor refers to the background surrounding the target. Each target was neighbored by either "H"s or "A"s in the morph-flanker and surrounded by a "dark" or "bright" outer ring in the brightness-illusion task. To make the tasks more efficient, we did not use a neutral background condition — task reliability majorly depends on the number of trials per condition (Rouder et al., 2019), and omitting neutral trials allowed us to administer more trials for the relevant condition. The second factor denotes the target. Both tasks had eight

[1] The raw data can be retrieved using the following links: https://raw.githubusercontent.com/PerceptionCognitionLab/data3/master/morph2/data for the morph-flanker and https://raw.githubusercontent.com/PerceptionCognitionLab/data3/master/bright2/data for the brightness-illusion task.

different targets. In the morph-flanker task, these more or less resembled the letters "H" and "A" and in the brightness-illusion task, circles in eight shades of grey, ranging from "dark" to "bright" (see Figure 2 for an overview of the targets). At the beginning of each task, participants completed two training blocks (eight trials each). First, they practiced the targets' identity without the distractors surrounding them. Then, they practiced the experimental part. The experimental part of each task comprised a total of five blocks of 64 trials each. Participants were instructed to identify the target as either "A" or "H" or "bright" or "dark", respectively, while ignoring the distractors. The target centrally appeared for 167 milliseconds on the screen, surrounded by either "A"s or "H"s or a "bright" or "dark" ring, respectively. Thereafter, the screen went blank again until participants pressed either the key "A" for an "A"/"bright" response or until they pressed the key "H" for an "H"/ "dark" response. Each response was followed by written feedback indicating whether the response had been correct or not. In the brightness-illusion task, the stimuli appeared on top of randomly created white noise.

We recorded participants' responses and RTs. By comparing their proportion of "A" and "dark" responses, respectively, in trials where a target was surrounded by "H"s/a bright ring to the ones where it was surrounded by "A"s/a dark ring, we assessed participants' ability to inhibit contrastive information.

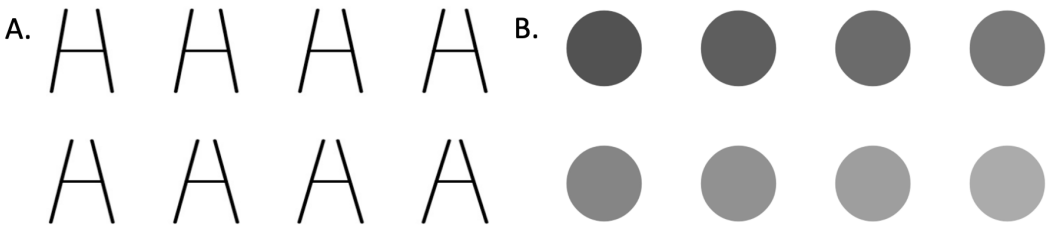

Figure 2. Target stimuli used in the morph-flanker (**A.**) and the brightness-illusion task (**B.**). The upper four stimuli belong to the H/dark category and the lower four to the A/bright category.

Results

We excluded all 16 warm-up trials from the analyses[2]. Two participants (partly) completed the tasks twice, which is why we excluded their second runs. Additionally, one of these participants had time-stamps that went back in time in their data. Therefore, we excluded this participant from the analyses. We further cleaned the data based on accuracy in the easiest conditions in that we excluded the data from those participants who did not respond correctly in at least 90 percent of the trials in the easiest task conditions in at least one task. In the morph-flanker task, the least difficult conditions are the ones where the target is the most A- or H-like morph and the distractor a grid of "H"s or "A"s, respectively. In the brightness-illusion task, the easiest conditions involve the brightest and darkest target surrounded by a "dark" or "bright" outer ring, respectively. Besides indicating that participants responded carefully, accurate responses in these conditions also reflect a proper understanding of the task instructions as they suggest that participants responded to the target as opposed to the distractor. It turned out that almost half of the participants did not meet this criterion and were therefore excluded from further analyses (i.e., 15 out of 27 participants). Given that identifying the target morphs in the easiest condition should not be challenging, we do not think that this exclusion criterion was too strong. Rather, we

[2] For all analyses, we used R (Version 3.6.1; R Core Team, 2019) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *brms* (Version 2.10.0; Bürkner, 2017, 2018), *coda* (Version 0.19.3; Plummer, Best, Cowles, & Vines, 2006), *corrplot2017* (Wei & Simko, 2017), *curl* (Version 4.2; Ooms, 2019), *dplyr* (Version 1.0.0; Wickham, François, Henry, & Müller, 2020), *ggplot2* (Version 3.2.1; Wickham, 2016), *gridExtra* (Version 2.3; Auguie, 2017), *invgamma* (Version 1.1; Kahle & Stamey, 2017), *MASS* (Version 7.3.51.4; Venables & Ripley, 2002), *Matrix* (Version 1.2.17; Bates & Maechler, 2019), *MCMCpack* (Version 1.4.5; Martin, Quinn, & Park, 2011), *msm* (Version 1.6.7; Jackson, 2011), *mvtnorm* (Version 1.0.11; Genz & Bretz, 2009), *papaja* (Version 0.1.0.9842; Aust & Barth, 2018), *Rcpp* (Version 1.0.2; Eddelbuettel & François, 2011; Eddelbuettel & Balamuta, 2017), *rstan* (Version 2.19.2; Stan Development Team, 2019a), *scales* (Version 1.0.0; Wickham, 2018), *StanHeaders* (Version 2.19.0; Stan Development Team, 2019b), and *tidyr* (Version 1.0.0; Wickham & Henry, 2019).

believe that the online setting of the experiment somehow led to more careless responses and participants who at some point during the experiment gave up. As a final step, we excluded all trials with unreasonably long RTs (> 2500 ms; a total of 28 trials)[3].

Overall contrast effects. Figure 3 shows the proportion of "A" and "bright" responses, respectively, across targets, distractor conditions, and participants as well as the contrast effects. The proportion of "A"/"bright" responses overall increased, the more A-like/the brighter the target was, depicting a sigmoid curve. We defined the contrast effect as the difference between the proportion of "A"/"bright" responses when the distractors are "H"/"dark" and when they are "A"/"bright". The plots in Figure 3 show considerable overall effects in the morph-flanker task but practically no effect in the brightness-illusion task. Plots showing the contrast effects with accuracy plotted on the y-axis and plots based on the full dataset can be found in Appendix A. Both plots vastly depict the same pattern of overall effects although the effects appear slightly smaller in the full dataset. Overall accuracy across participants, targets, and conditions in the cleaned data was 0.86 for the morph-flanker and 0.88 for the brightness-illusion task.

To quantify the evidence for the presence of contrast effects, we conducted two one-sided paired Bayesian t-tests using a noninformative Jeffreys prior for the variance parameter and a cauchy prior with a scale of $\sqrt{2}/2$ as specified by default in the BayesFactor package[4]. The data supported hypothesis 1 that participants responded more often with "A" in the morph-flanker task when the distractors were "H" rather than "A"s, $M = 0.13$ 95%

[3] There was an issue with the RT recording, as some latencies were recorded as negative. This was the case for 271 negative values in the morph-flanker task and 79 concerned data from 5 participants in the morph-flanker and 4 in the brightness-illusion task (in the raw data). We suspect that this had something to do with the browser type and/or computer system that participants used. We therefore decided not to exclude any participants based on fast responses.

[4] Note that in the thesis proposal, we had initially planned to test hypotheses 1 and 2 using a hierarchical probit model as we did for hypotheses 3 to 5. Later on, we decided that it was more straightforward to use paired t-tests to test hypotheses 1 and 2.

HDI [0.06, 0.19] over the null hypothesis of no effect by a Bayes factor of 141.72. A Bayes factor is a continuous measure of evidence comparing the predictive accuracy of two competing hypotheses/models (e.g., Rouder, Haaf, & Aust, 2018). A Bayes factor of 141.72 means that the data are 141.72 times more likely to have occurred under hypothesis 1 than under the null hypothesis. For the second hypothesis, that participants responded more with "bright" in the brightness-illusion task when the distractors were dark rather than bright, $M = 0.04$ 95% HDI [0.00, 0.08], $\text{BF}_{10} = 1.23$ there is much less evidence for an overall contrast effect. The data were only 1.23 times more likely to have occurred under hypothesis 2 than under the null hypothesis of no effect, suggesting neither evidence for nor against hypothesis 2.

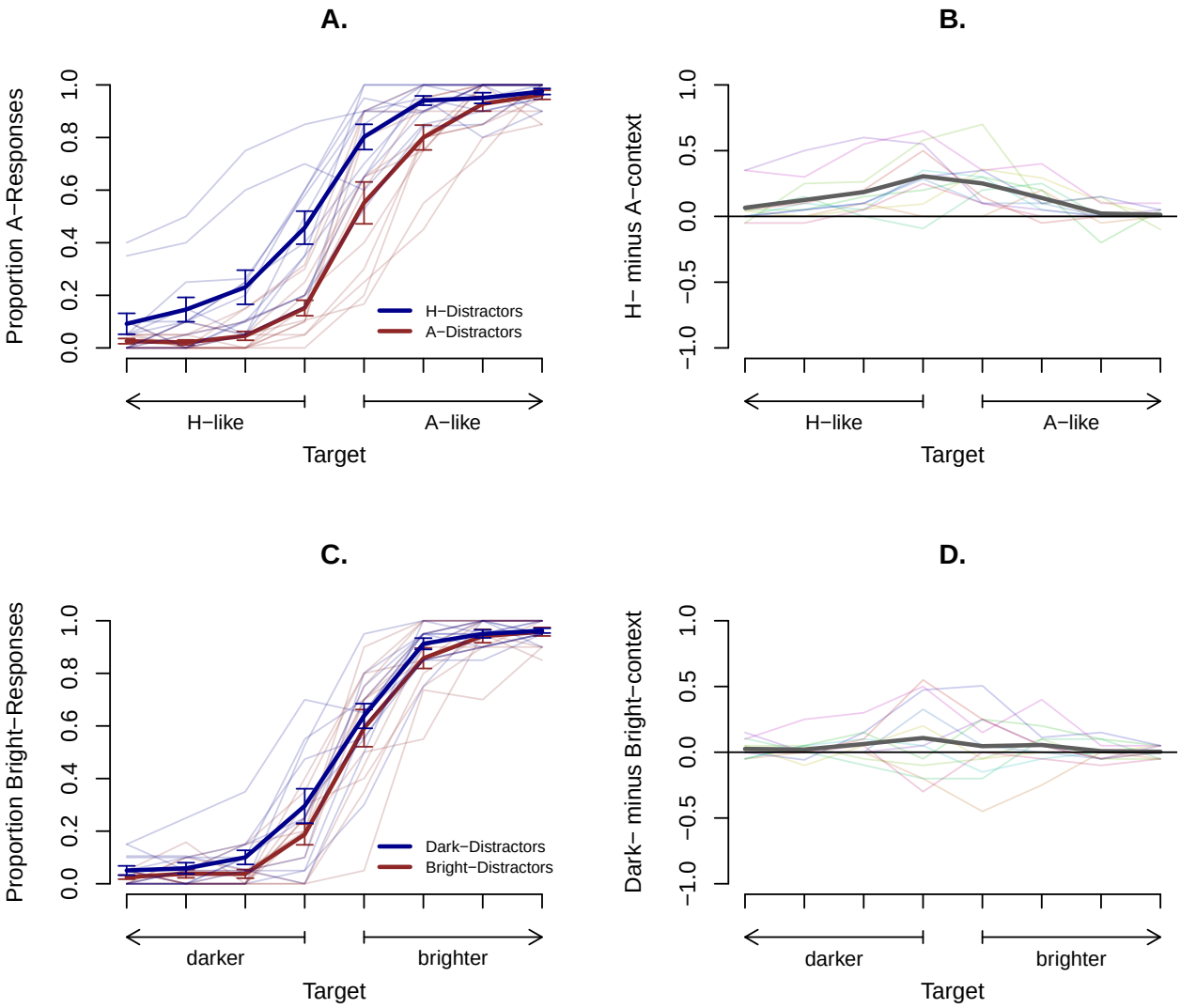

Figure 3. **A.** and **C.** show the proportion of 'A' responses in the morph-flanker task and the proportion of 'bright' responses in the brightness-illusion task, respectively, for each target. Each line represents a participant's average response in either of the distractor conditions. The thick lines denote the average proportion per condition and across participants. The error bars denote the standard errors. The contrast effect for each target is reflected in the vertical distance between the blue and red lines. **B.** and **D.** show the inhibition effects across the two tasks. For each target, we subtracted the proportion of 'A' or 'bright' responses, respectively, in one condition from the other. Each line represents a participant and the thick grey lines shows the average contrast effects.

Individual differences in contrast effects. To estimate the extent of true individual variation, we estimated a hierarchical trial-level model. Such a model allowed us to quantify contrast effects across targets and to estimate the correlation between the tasks while taking trial noise into account. Because participants only completed a finite number of trials, their observed effects are confounded by measurement error. Measurement error is known to attenuate correlations (e.g., Matzke et al., 2017; Spearman, 1904) and to overestimate individual differences. In theory and when the trial number is large enough, hierarchical models can estimate effects in the large trial limit, thereby disattenuating correlations (but see Rouder et al., 2019). They also provide more conservative estimates of individual differences by shrinking the individual effects towards the population mean (e.g., Efron & Morris, 1977; Haaf & Rouder, 2018).

Model specification. We describe the hierarchical model using random variable notation. Similar models were described in Snyder et al. (2019), Rouder and Lu (2005), and in Rouder and Haaf (2019). Let $Y = 0, 1$ denote whether a response is "H" or "A" in the morph-flanker task, or "dark" or "bright" in the brightness-illusion task. Further, let Y_{ijklm} denote a response for the ith individual, $i = 1, \ldots, 12$, in the jth task, $j = 1, \ldots, 2$ (for the morph-flanker and brightness illusion task, respectively), the kth target, $k = 1, \ldots, 8$ the lth distractor condition, $l = 1, 2$ ("H" or "A" or "dark" or "bright", respectively) and the mth replicate, $m = 1, ..., 20$.

We model the dichotomous responses Y_{ijklm} as independent Bernoulli-trials with a probit link transforming probabilities into z scores $\in (-\infty, \infty)$:

$$Y_{ijklm} \overset{ind}{\sim} \text{Bernoulli}\left[\Phi(\eta_{ijkl})\right].$$

Here Φ denotes the cumulative distribution function of the standard normal, and η_{ijkl} is the combined effect of participants, tasks, targets, and background conditions on the propensity to respond with "A" or "bright", depending on the task j. An η_{ijkl} of zero corresponds to a probability of 0.5 of responding with "A"/"bright".

To model individual inhibition effects across tasks, we decompose η_{ijkl} as follows,

$$\eta_{ijkl} = \alpha_{ij} + u_k\beta_{ij} + x_l\gamma_{ij}.$$

u_k indicates the background condition, $u_k = \{-0.5, 0.5\}$, where "H"-distractors and "dark"-distractors were coded as 0.5 and "A" and "bright"-distractors as -0.5. β_{ij} is our main parameter of interest, the contrast effect for individual i in task j. x_l denotes the target, ranging from -3.5 to 3.5 with the most ambiguous ones centered around 0. γ_{ij} denotes participant i's target effect in task j. This parameter captures the notion that the propensity to respond with "A"/"bright" increases, the more A-like/the brighter the target is. Lastly, α_{ij} is the intercept for individual i in task j. It is the average propensity to respond with "A"/"bright" for a middle target. Consequently, it can be understood as a parameter that captures the bias towards either response.

We needed prior distributions for α_{ij}, β_{ij}, and γ_{ij}. Because we used Gibbs sampling (Geman & Geman, 1984) to estimate the model, normal priors for the population means and inverse-gamma priors for the variances are a convenient choice. We start with the intercept α_{ij},

$$\alpha_{ij} \sim \text{Normal}(\mu_{\alpha_j}, \sigma^2_{\alpha_j}),$$

on which we placed a normal prior distribution with mean μ_{α_j} and variance σ^2_j. We used the following hyperpriors:

$$\mu_{\alpha_j} \sim \text{Normal}(0, 0.15^2)$$

$$\sigma^2_{\alpha_j} \sim \text{Inverse-Gamma}(3, 0.1).$$

The prior on μ_{α_j} reflects the assumption that the overall bias towards either response lies somewhere between 0.4 and 0.6 in the probability space and that values outside of that range are increasingly implausible. The prior on $\sigma^2_{\alpha_j}$ reflects the expectation that the variance lies below 0.1 and that greater variability parameters are implausible. For the inhibition effects β_{ij}, the parameter of primary interest, we used a multivariate-normal prior,

$$\beta_{ij} \sim \text{Multivariate-Normal}(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}),$$

where $\boldsymbol{\mu_\beta}$ is a vector of length two containing the population means μ_{β_j}, one for each of the two tasks, and $\boldsymbol{\Sigma_\beta}$ is a 2x2 variance-covariance matrix. We placed again a normal distribution on the population means,

$$\mu_{\beta_j} \sim \text{Normal}(0, 0.7^2).$$

This prior is again centered on zero and restricts the population means to a plausible range. As hyperprior for the variance-covariance matrix $\boldsymbol{\Sigma_\beta}$ we used an inverse-wishart distribution,

$$\boldsymbol{\Sigma_\beta} \sim \text{Inverse-Wishart}(3, \Omega).$$

Here, Ω is a scale matrix, $\Omega = \begin{vmatrix} 0.04 & 0 \\ 0 & 0.04 \end{vmatrix}$ and three is the number of degrees of freedom. With two estimated variances and one covariance, three degrees of freedom reflect a uniform prior on the size and direction of the correlation. The diagonal in Ω contains the expected variances, 0.04. They are weakly informative and reflect our expectations regarding the ratio of latent individual variation to trial noise. This signal-to-noise ratio can be computed by dividing the standard deviation of the contrast effect (i.e., the interindividual variability) by the standard deviation of the observations (i.e., trial-by-trial variation; Rouder et al., 2019). Based on the findings of Snyder et al. (2019), we expected a slightly higher signal-to-noise ratio than what is commonly found in standard inhibition tasks. A variance of .04 corresponds to a signal-to-noise ratio of 1/5 (i.e., $\sqrt{0.04}/1 = 1/5$) since in our hierarchical probit model, the trial-by-trial variation is set to 1 (see Appendix B). Whereas in standard inhibition tasks such as the Stroop and flanker task, this ratio has shown to lie between 1/7 and 1/12 (Rouder et al., 2019). The off-diagonal elements of 0 in Ω reflect a lack of information regarding the magnitude and direction of the covariation between the inhibition effects across tasks (e.g., Rouder et al., 2007). The values we place on the diagonal are equal to each other, reflecting no prior information regarding the relative size of the individual variability in the two tasks. The covariance between the two inhibition effects across tasks is defined as $\Sigma_{1,2} = \rho\sigma_{\beta_1}\sigma_{\beta_2}$, and so the task correlation can be computed as follows: $\rho = \frac{\Sigma_{1,2}}{\sigma_{\beta_1}\sigma_{\beta_2}}$.

Lastly, for the target effects, we used the same prior settings as for the contrast effects,

$$\gamma_{ij} \sim \text{Multivariate-Normal}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}).$$

On the population mean, we place the same normal distribution,

$$\mu_{\gamma_j} \sim \text{Normal}(0, 0.7^2),$$

and on the the variance-covariance matrix the inverse-wishart prior,

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \sim \text{Inverse-Wishart}(3, \Omega),$$

with three degrees of freedom and scale matrix $\Omega = \begin{vmatrix} 0.04 & 0 \\ 0 & 0.04 \end{vmatrix}$.

To check whether the prior settings were reasonable, we generated 4000 data sets from the prior predictive distribution. Figure 4A and B show the overall contrast effects across the eight targets. The generated data cover all reasonable sizes of the overall effect with most datasets centered around zero. For example, highly implausible outcomes, such as overall effects as high as 0.8, are not covered by the generated datasets. Note that the effects are more extreme for the middle targets. This pattern is in line with our expectations and caused by the coding of the target effects from -3.5 to 3.5 with the most ambiguous ones in the middle. When predicting the propensity to respond "A"/"bright", the contrast effect is relatively more accentuated for the ambiguous targets than for the easier ones.

Model estimation. We estimated the model using Gibbs sampling (Geman & Geman, 1984). We ran 20000 iterations and discarded 2000 as burnin. Consequently, inference is based on 18000 samples[5]. A summary overview of the estimated population means and variances as estimated using Gibbs sampling is shown in Tables 1 and 2.

[5] The model estimation procedure is further described in Appendix B. To cross-check the Gibbs sampling results, we also estimated the specified model in Stan. This estimation yielded largely the same result. Perhaps the greatest difference between the two estimation approaches lies in the size of the correlation

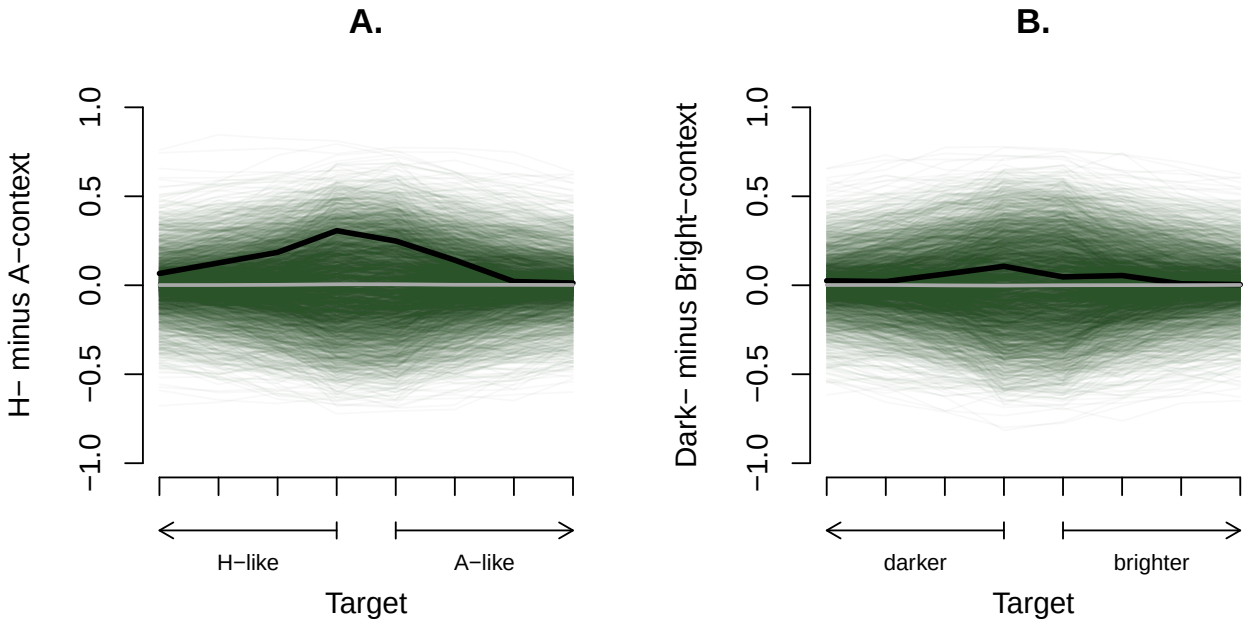

Figure 4. Contrast effects (as shown in Figure 3C and D) from four thousand datasets generated according to the prior predictive distribution for the morph-flanker (**A.**) and the brightness-illusion task (**B.**). Each horizontal line denotes a dataset. The black line denotes the observed effects in the collected data and the lightgrey line the contrast effect for each target averaged across all generated datasets.

Model fit assessment. Before presenting the estimation results, we describe the model fit. We first checked to what extent we could reproduce the data using point estimates of the regression model parameters. We thereby compared the average proportion of "A"/"Bright" responses in each distractor condition as observed in the data to the corresponding average proportions as computed using the posterior medians of the model parameters. We did this for each of the sixteen target-distractor conditions and for both tasks. As shown in Figure 5, the observed and estimated proportions across conditions are more or less aligned. The biggest misfit in both tasks seems to be at the fourth target where, according to the model, the proportion "A"/"Bright" is slightly higher than observed. But

coefficient which was estimated slightly higher in Stan. An overview of those results can be found in Appendix B.

overall, there are no concerning differences.

To inspect the individual-level model fit, we computed the average probit-transformed contrast effects per participant and task using again the point estimates of the model parameters. We then compared them to the corresponding "observed" contrast effect on the probability scale. For this purpose, we only used data from the four middle targets as the contrast effect is mostly reflected in those. For some individuals, the contrast effect could be better reproduced than for others and the model-based average contrast effect seemed to be slightly overestimated (see Figure 6A and B). However, overall, there was no dramatic misfit. We also inspected to what extent the model could reproduce the intercepts. We turned the probit-transformed intercepts into probabilities and compared them to the observed overall proportion of "A"/"bright" responses (see Figure 6C and D). Here, the model somewhat overestimated the extent of individual variation for the morph-flanker but less so for the brightness-illusion task. We explored whether a more restrictive prior variability parameter would reduce this, but none of the prior settings we tried out improved the fit.

One possibility is that the misfits observed were caused by some peculiarities in the small dataset. We therefore fitted the model on simulated data with more participants. The model assessment plots from this simulation can be found in Appendix D. Essentially, we observed the same patterns in the simulated data as in the pilot data, suggesting that the patterns did not occur due to the small sample size. To further assess the fit of the model, we conducted posterior predictive checks which are shown in Appendix D. In sum, apart from a couple of misfits at the individual-level, the model could reproduce the general patterns in the data.

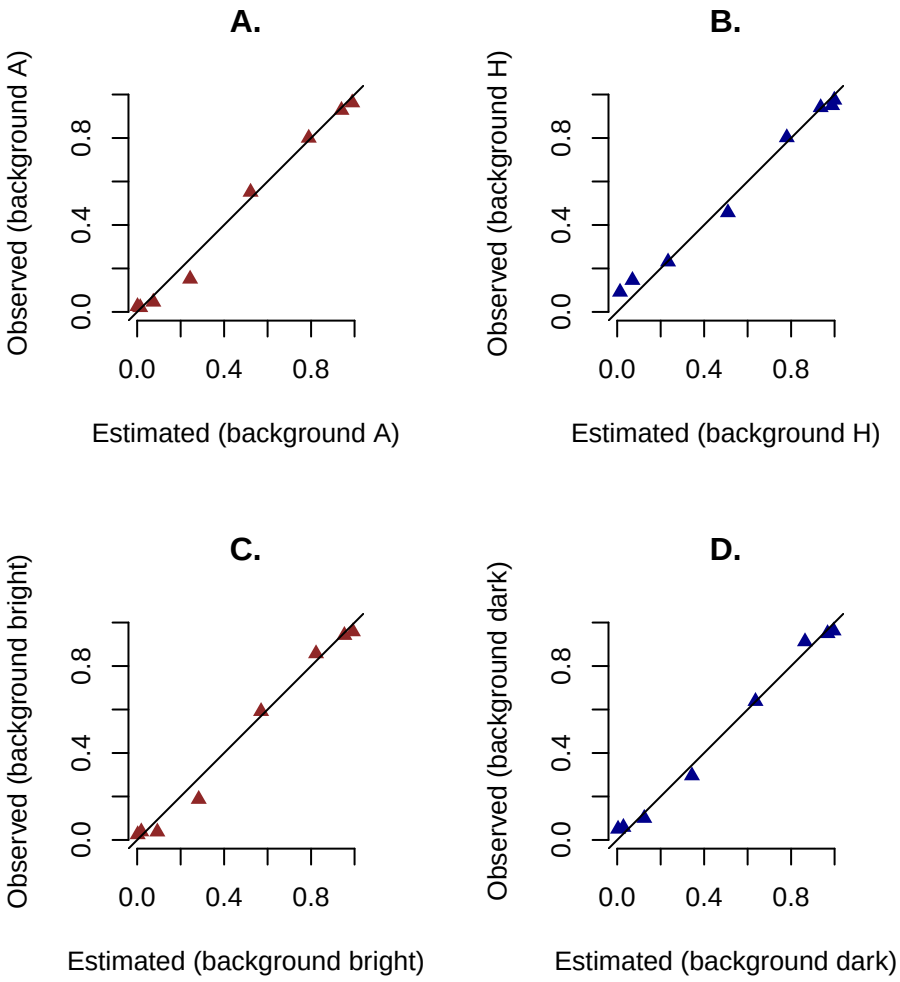

Figure 5. **A.** and **B.** show the model-based and observed proportions of A-responses for the background conditions 'A' and 'H', respectively. **C.** and **D.** show the proportion of 'bright' responses for the background conditions 'bright' and 'dark', respectively. Each triangle represents a target. The model-based and observed proportions are highly correlated, suggesting that the model represents the general pattern in the data well.

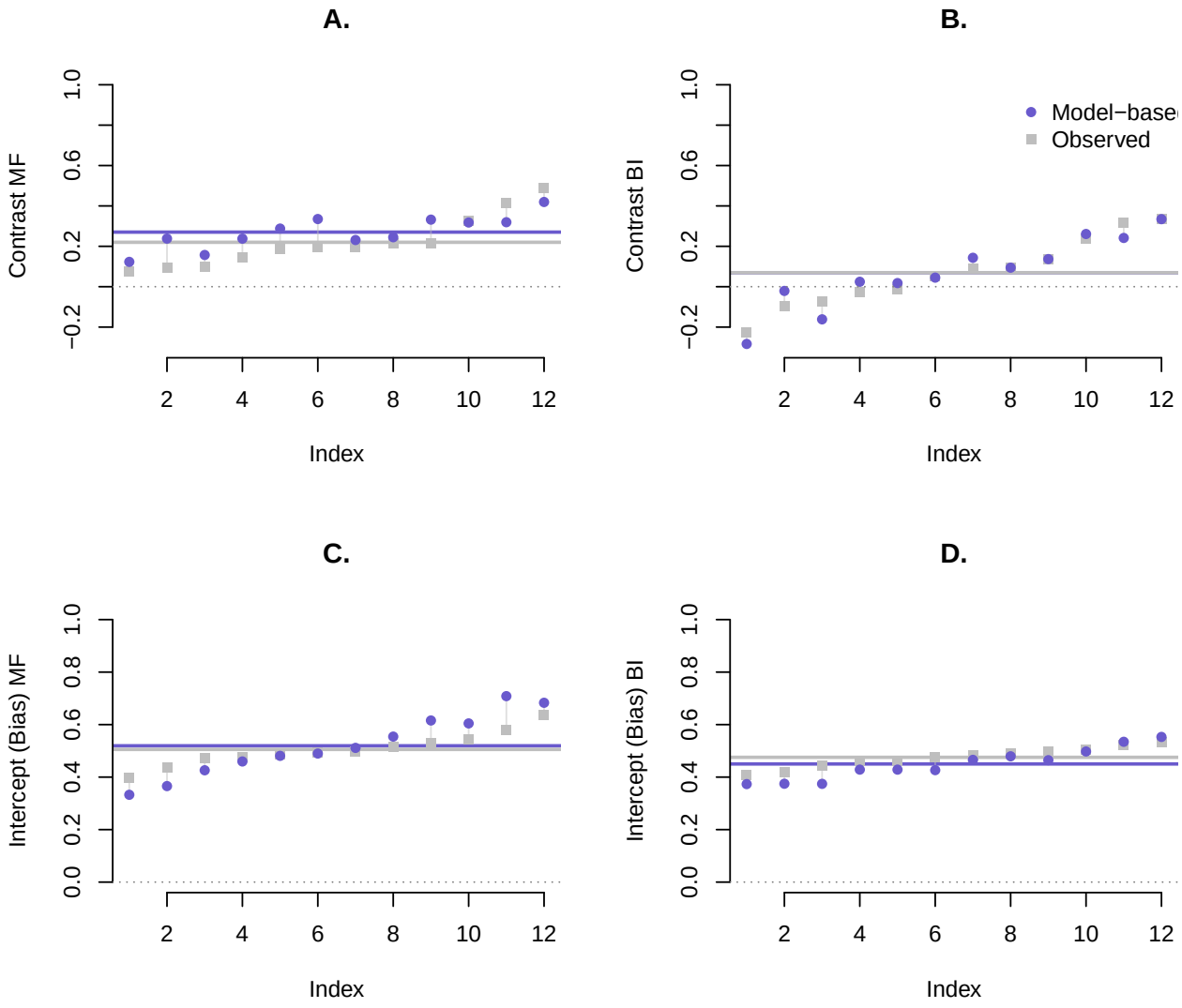

Figure 6. Fit assessment of the hierarchical trial-level model. **A.** and **B.** show the contrast effects on the probability scale for the morph-flanker and the brightness-illusion task, respectively. **C.** and **D.** show the observed and estimated intercepts per participant (see text for details). Each dot and rectangle denotes a participant. The blue dots are computed based on model-based estimates and the grey rectangles are observed effects. The horizontal lines indicate the corresponding averages.

Model estimation results. Figure 7A and B show the posterior medians of the overall contrast effects of the two tasks, μ_{β_1} and μ_{β_2}, and the posterior medians of the hierarchical estimates, β_{i1} and β_{i2}, in increasing order including their credible intervals. Both

plots suggest that there are considerable individual differences in the sample since some credible intervals do not overlap with each other.

In the morph-flanker task, all but one 95% credible intervals exclude zero. Furthermore, all posterior medians lie on the positive side, suggesting that more or less everyone has a positive contrast effect. To quantify the evidence for this constraint[6], we used the encompassing prior approach (Klugkist & Hoijtink, 2007; Klugkist, Laudy, & Hoijtink, 2005). In the encompassing prior approach, one counts the evidential iterations that fulfill a constraint in the posterior samples, as well as in the prior samples. The Bayes factor can then be approximated by the ratio of evidential iterations in the posterior and prior samples. This Bayes factor was 2.89 in favor of a "positive only" effect as opposed to an unconstrained effect in the morph-flanker task.

In the brightness-illusion task, some participants seem to show an effect in the expected direction, some an effect around zero, and a few even an effect in the opposite direction, suggesting that they judged a target more often as "bright" when it was surrounded by a bright outer ring rather than a dark outer ring. In line with this observation, the Bayes factor comparing the relative predictive accuracy of an unconstrained model that allows positive, negative, and no effects to a positive-only model was infinitely high in favor of the unconstrained model.

---

[6] The analyses involving the encompassing prior approach had not been previously specified in the thesis proposal and corresponding response letter. For the sake of readability, they are reported here and not in the "exploratory analysis" section.

Table 1

Model estimation results for the morph-flanker task

	Median	SD	Lower Bound	Upper Bound
μ_{α_1}	0.039	0.077	-0.114	0.189
μ_{β_1}	0.718	0.113	0.486	0.938
μ_{γ_1}	0.749	0.082	0.589	0.912
$\sigma^2_{\alpha_1}$	0.080	0.038	0.039	0.183
$\sigma^2_{\beta_1}$	0.065	0.054	0.016	0.216
$\sigma^2_{\gamma_1}$	0.066	0.040	0.031	0.177

Note. Posterior medians, standard deviations, and 95% credible intervals of the population means and variances for the morph-flanker task (indicated by the subscript 1) as estimated using Gibbs sampling.

Table 2

Model estimation results for the brightness-illusion task

	Median	SD	Lower Bound	Upper Bound
μ_{α_2}	-0.113	0.058	-0.225	0.004
μ_{β_2}	0.172	0.156	-0.140	0.476
μ_{γ_2}	0.750	0.070	0.615	0.891
$\sigma^2_{\alpha_2}$	0.033	0.017	0.016	0.078
$\sigma^2_{\beta_2}$	0.219	0.140	0.084	0.605
$\sigma^2_{\gamma_2}$	0.047	0.030	0.020	0.129

Note. Posterior medians, standard deviations, and 95% credible intervals of the population means and variances for the brightness-illusion task (indicated by the subscript 2) as estimated using Gibbs sampling.

Comparing the extent of individual differences among the contrast effects, one can see that the brightness-illusion task yielded more individual variation than the morph-flanker task. This is reflected in the posterior medians of the variances $\sigma^2_{\beta_1}$ (0.06) for the morph-flanker and $\sigma^2_{\beta 2}$ (0.22) for the brightness-illusion task. Moreover, a Bayes factor of 35.22 suggested that it is 35.22 times more plausible that the brightness-illusion task yields greater individual variation than the morph-flanker as opposed to the other way around.

Correlation between contrast effects. To assess our fifth and last hypothesis that the inhibition effects in the morph-flanker and the brightness-illusion task are positively correlated, we computed a Bayes factor using the Savage-Dickey approach (Dickey, 1976). In this method, the Bayes factor is approximated by computing a ratio of the prior and

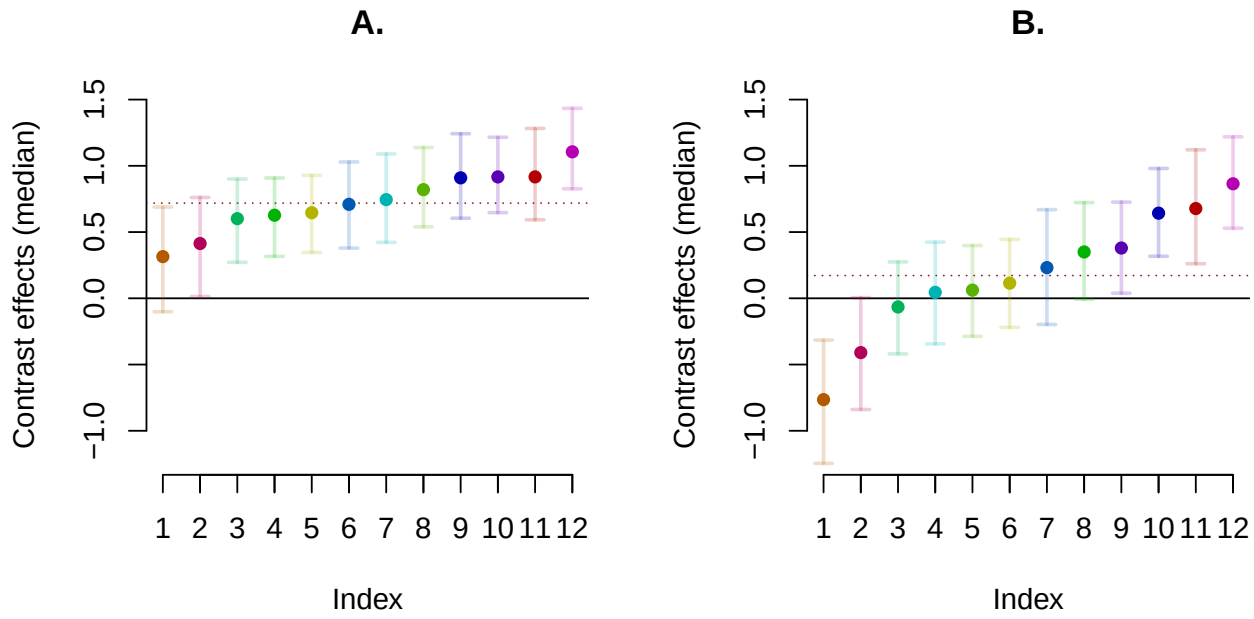

Figure 7. **A.** shows the individual contrast effects (posterior medians) in increasing order for the morph-flanker task and **B.** for the brightness-illusion task. Each dot represents a participant. The horizontal lines denote the 95% credible intervals. The dotted line shows the posterior median of the population mean of the contrast effect.

posterior density at point zero. The resulting Bayes factor was 10.01, suggesting that the data were 10.01 times more likely to have occurred under hypothesis 5 than under the null hypothesis of no correlation. Note that this Bayes factor reflects an undirected test whereas our hypothesis was a directed one. We realized that we cannot test the directed hypotheses in the intended manner because we did not place a truncated prior on the variance-covariance matrix $\mathbf{\Sigma}_\beta$. However, since we used a uniform prior on the correlation, the Bayes factor in favor of a positive correlation as opposed to no correlation is double as high, $BF = 20.02$. The posterior median of the correlation was 0.79, $SD = 0.21$, 95% CrI[0.18, 0.96] suggesting that the contrast effects across tasks were highly positively correlated: the higher the contrast effect in the morph-flanker, the higher the contrast effect in the brightness-illusion task. Figure 8C visualizes the correlation coefficient in terms of its prior and posterior distribution and Figure 8B in terms of a scatterplot of participants'

estimated contrast effects.

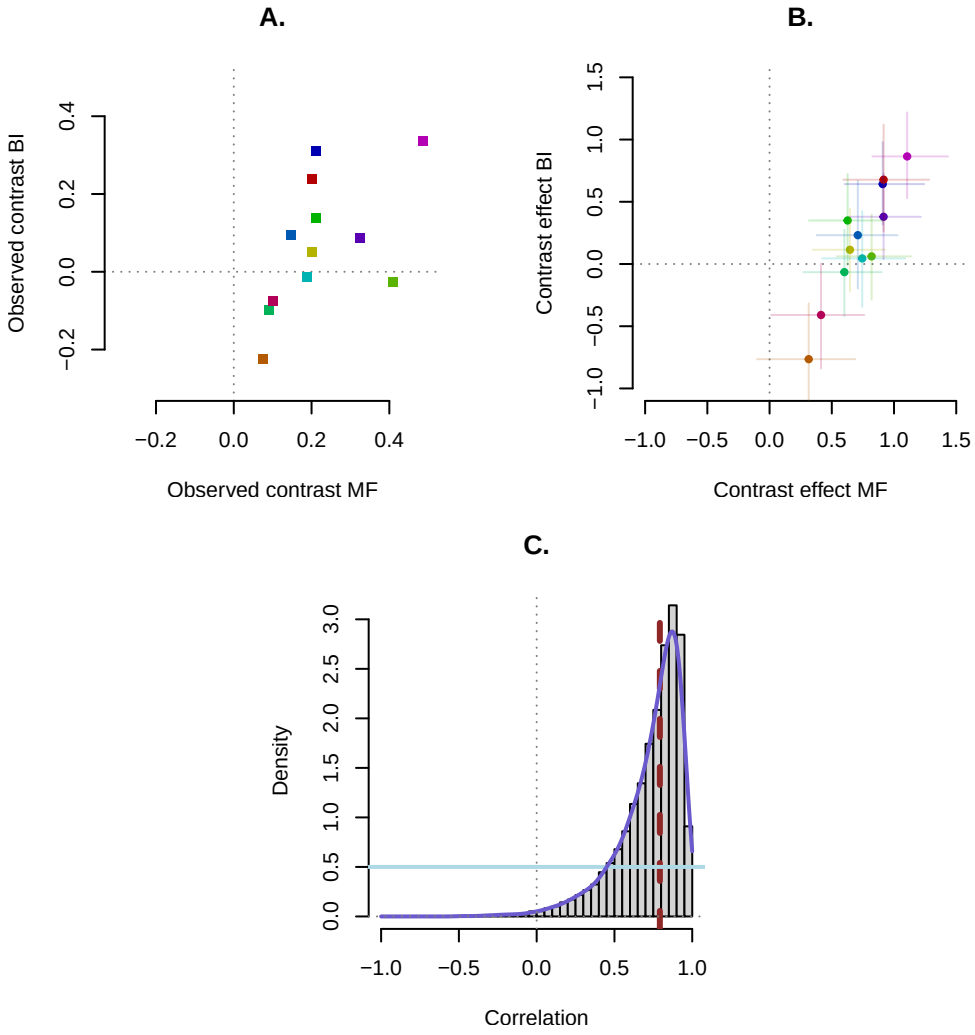

Figure 8. Correlation between the contrast effect in the morph-flanker and the brightness-illusion task. **A.** shows the relationship between the observed contrast effects based on the middle four targets in the two tasks. Each rectangle represents a participant. **B.** shows the individual estimates (medians) of the contrast effect in the morph-flanker and the brightness-illusion task. Each dot represents a participant. The horizontal and vertical lines denote the 95% credible intervals. **C.** shows the posterior and prior distribution of the correlation. The dotted vertical line denotes the median and serves as a point estimate. MF indicates morph-flanker and BI brightness-illusion.

The correlation coefficient as estimated by the hierarchical model seems extraordinarily high. Therefore, even if attenuation due to trial noise is at play, we should also see a correlation in the observed contrast effects. For example, Rouder et al. (2019) suggested that in most designs, an attenuation factor of around two can be expected. This would lead us to expect an "observed" correlation as high as $\rho = 0.40$. As a plausibility check, we examined the correlation between the "observed" contrast effects of the tasks. We computed the average proportion of "A"/"bright" responses across the four middle targets in each task and for each distractor condition separately. The contrast effect reflects the difference between these two averages. In line with the modeling results, 8A depicts a positive relationship between the observed contrast effects across tasks posterior mean = 0.36, $SD = 0.22$, $r = 0.333$, $\text{BF}_{10} = 2.00$. This is a sizable correlation given the small sample size. And it is approximately what we would expect under the assumption that observed correlations are about half as big as latent ones.

Exploratory Analysis. We also explored to what extent there were individual differences in the intercepts and in the target effects. There was substantial variation in both. Furthermore, the target effects were positively correlated, posterior median = 0.69, $SD = 0.19$, 95% CrI[0.16, 0.91]. This finding suggests that the more individuals discriminated between the different morph-levels, the more they discriminated between the different brightness-levels and vice versa. Plots of the hierarchical estimates for the intercepts and target effects across participants and the correlation between the target effects can be found in Appendix C.

Discussion (Pilot) Study 1

This pilot study examined the morph-flanker and the brightness-illusion task as two paradigms for assessing the ability to inhibit contrastive information. The morph-flanker task worked as expected and yielded a sizable inhibition effect in that participants were more

likely to respond with "A" when the (morph-)target was surrounded by H's rather than A's. The results of the brightness-illusion task, on the other hand, were different than expected. Some participants even had a negative effect, meaning that they were more likely to respond with "bright" when the distractor was bright rather than dark. However, both tasks yielded substantial individual variation. Despite the small overall effect, the brightness-illusion task even led to more individual differences than the morph-flanker, and the contrast effects across the two tasks were highly positively correlated.

The results are initial evidence that the two tasks can distinguish between individuals. They may even have a better resolution to detect individual variation in the presence of trial noise than standard ones such as the Stroop or flanker task. Both tasks yielded a greater signal-to-noise ratio than what is commonly found. In the morph-flanker, this ratio was 1/3.59 and in the brightness-illusion task, this was 1/2.00. Whereas the signal-to-noise ratio in standard inhibition paradigms commonly lies between 1/7 and 1/12 (Rouder et al., 2019).

An unexpected finding was that the brightness-illusion task yielded substantial individual differences even though the data provided very little evidence for an overall effect. The brightness-illusion task even led to more individual variation than the morph-flanker task. This finding is unexpected because we started this research assuming that a sizable overall effect is needed to find true individual variation. The reasoning was that true individual variation would be bounded at zero (and therefore restricted by the size of the overall effect) because negative contrast effects are implausible (see e.g., Haaf & Rouder, 2018). Yet, two participants showed such a negative contrast effect. A negative contrast effect is the opposite of the visual illusion that the task is thought to evoke. Essentially, it is an assimilation effect where individuals align their response with the distractors. It means responding with "bright" more often when the target is surrounded by a bright rather than a dark outer ring. Note that the outer rings were always more extreme than the inner circles such that the distractor was brighter than the brightest target. This raises the question as to how credible these results are. An alternative explanation for the negative effects is that the

participants suppressed their natural responses and deliberately showed the opposite effect. On the other hand, the RTs in question were not noticeably different from the rest. We would expect that consciously suppressing one's initial reaction increases RTs. As we proceed with this research, we will increase task difficulty of the brightness-illusion task by making the targets more similar and the outer rings more extreme. This way we hope to increase the overall contrast effect. It is is possible that the whole range of individual estimates will then be shifted towards the positive side, such that negative contrast effects no longer persist.

While the results of this pilot study are encouraging, they should be interpreted with some caution. One reason is the small sample size ($n = 12$) that the results are based on. We excluded over half of the initial participants due to careless responding. We suspect that the online setting of the study contributed to the low quality of the data. In the future, we will take several measures to increase the quality of the data. By welcoming participants to future studies personally, for example via online video call, we hope to increase a sense of accountability. Furthermore, we will remind participants of the task instructions if they fall below a certain accuracy level and increase stimulus presentation time for participants who answer incorrectly in the easier trials. This might discourage rushing through the task.

We also observed some misfit in the model. While the hierarchical probit model overall seemed to provide an acceptable account of the data, the fit was worse at the individual level. One explanation for this misfit might be the model's restrictiveness. For example, regarding the effects of the target, the model assumes a strictly linear trend in the probit space. This assumption might not hold in reality. Hopefully, more extensive simulations and more data both at the individual as well as the trial-level will show whether certain model assumptions or aspects of the model in general need to be adjusted.

Initially, our plan was to use the results of this pilot study to decide with which of the two tasks to proceed. The decision criterion was the extent of latent individual variation. However, because of the unexpected result that the task with the smaller overall effect yielded more individual differences and the small sample size, we will postpone this decision.

We will make the described adjustments to the brightness-illusion and collect new data before deciding which of the two proposed tasks we will use in the correlational study. Nonetheless, we describe the planned correlational study below.

Proposed Correlational Study

Given we have established that at least one of the proposed tasks works as expected and leads to substantial individual variation, the next question we aim to address is: does the proposed task truly assess individual differences in inhibition? We will answer this question by examining what performance in the proposed task correlates with. For example, we can explore how contrast effects relate to cognitive abilities that have shown to be associated with inhibitory processes, thereby gaining more insight into the psychometric structure of the task and inhibition as a construct.

Since substantial individual variation is needed to detect covariation, we plan to administer tasks that are known to be well-discriminative: a working memory capacity measure and a visual inspection time task. First, from an executive-attention account of working memory (WM; Engle, 2002, 2018), one of the driving forces behind WM capacity is the ability to keep information retrievable in the face of interference. Accordingly, studies suggest that individuals with better inhibitory abilities have greater WM capacity (e.g., Chuderski, 2014; McVay & Kane, 2009, 2012; Unsworth, Fukuda, Awh, & Vogel, 2014). Therefore, if the proposed task truly assesses an inhibitory process, we would expect a negative correlation between the contrast effects in our proposed task and a WM capacity score in that people who are better at inhibiting contrastive information presumably have a higher WM capacity score.

Second, visual inspection time (VIT) reflects the time needed to process a single visual input (Vickers, Nettelbeck, & Willson, 1972). VIT tasks are among the most popular psychometric measures for general cognitive ability. They capture individual differences in basic information processing, specifically in perceptual discrimination. Including a VIT task

in the correlational study will allow us to explore how individuals' basic perceptual ability is associated with their performance in the proposed inhibition task.

A useful property of the proposed and standard inhibition tasks is that inhibition can be isolated from general cognitive performance by computing difference scores between congruent and incongruent conditions. Therefore, if the contrast effect in the proposed task truly assesses an inhibitory process, we would expect it to not be closely related to VIT. In fact, a positive correlation would suggest that the effect may not (only) be a function of one's ability to inhibit contrastive information but (also) to visually discriminate briefly presented target-morphs (i.e., the A-to-H morphs or different shades of grey, respectively). Moreover, there may only be a correlation because some people process visual stimuli faster than others (see processing-speed theory; Salthouse, 1996). Since the correlations may then not tell us anything about inhibitory processes, this would be an important insight. We would, however, expect VIT to be positively correlated with the target effects since those are thought to represent individuals' ability to distinguish between morphs.

Besides WM capacity and VIT, we will also include a regular letter-flanker task. This task will serve as a reference point for the extent of individual (co-)variation commonly found in inhibition research. Due to the methodological challenges, we are skeptical about finding a correlation between the proposed task and the regular letter-flanker task. It is likely that the resolution will not be high enough to obtain correlations. Simulations by Rouder et al. (2019) suggest that over a thousand trials per task and participant may be needed to find correlations across inhibition paradigms such as the letter-flanker. This is not feasible within the scope of this project. But even if we do not find correlations, it will still be interesting to directly compare the tasks, for example, regarding their signal-to-noise ratio as the morph-flanker and the brightness-illusion tasks were inspired by the design of the standard letter-flanker task.

This study will be exploratory in that we will not specify any strong hypotheses regarding the presence of correlations between the proposed task and the remaining tasks.

As outlined above, different patterns of correlations seem plausible. The goal is to explore whether the contrast effects go hand in hand with performance in other established cognitive tasks. If we find correlations, this would suggest that the proposed task does indeed assess some cognitive mechanisms that go beyond those underlying the paradigm itself. If we do not find correlations, we might be merely assessing task-specific performance with little meaning for the concept "inhibition" (see Rey-Mermet et al., 2018). Furthermore, if we find that the signal-to-noise ratio is considerably smaller for the proposed task than for the letter-flanker, a standard inhibition task, we will have some initial direct evidence that the proposed task may be better suited for assessing individual differences in cognitive inhibition.

Planned Procedure

Participants will complete the four cognitive tasks on different days in random order. For the proposed inhibition task, the letter-flanker task, and the working memory task, we will administer two sessions on two different days. This will allow us to increase the number of trials per participant. Accordingly, the study will take place over seven days. At the beginning of the first session, participants will answer a few demographic questions. They will complete the sessions online on their own computer, except for the VIT task. Because of the perceptual nature of this task, it is particularly important to have a controlled environment. Before completion of each task, participants will receive oral as well as written instructions. For the tasks administered online, we will give participants personal instructions via a video-conferencing software. Each session will take approximately thirty minutes to complete.

Participants

Participants will be undergraduate students at the University of California, Irvine. They will receive course credit for their participation.

Materials

All tasks will be programed using lab.js (Henninger et al., 2019).

Demographics. Participants will be asked the same questions as in the pilot study.

Proposed inhibition task. Depending on the results of future pilot studies, participants will either complete the morph-flanker or the brightness-illusion task. They will complete sixteen warm-up trials, eight only showing the targets and eight mimicking the experiment. The number of experimental trials per session will be 200, resulting in 400 trials in total.

Rotation span. We will use the rotation span task (e.g., Kane et al., 2004) to measure WM capacity. This task has shown to be well-discriminative of high-ability individuals such as university students (Draheim, Harrison, Embretson, & Engle, 2018). Participants need to recall a series of three arrows pointing in one of eight directions. They must keep the arrow sequence in memory while, before each arrow, judging whether a letter is mirrored or not. After three letter-arrow pairs, participants are asked to recall the sequence of arrows in the right order (see Figure 9A). The rotation span score will be computed by taking the average proportion of correctly recalled arrows with respect to order and orientation (Kane et al., 2004; Wilhelm, Hildebrandt, & Oberauer, 2013). As letter stimuli, we will use the letters R, G, and F. Each trial will consist of three letter-arrow pairs. While researchers commonly administer several set sizes (e.g., set sizes ranging from two to five; Wilhelm et al., 2013), we plan to only use one set size of three to maximize the number of trials per condition, which we think will reduce trial noise. Participants will complete two practice and 36 experimental trials per session. They will be given auditory feedback during the letter-task and visual feedback after the recall task.

Letter-flanker. In the letter-flanker task (Eriksen & Eriksen, 1974), participants are

asked to identify a centrally located letter-target surrounded by other distractor-letters. They will first see a fixation cross followed by a target distractors (see Figure 9B). In congruent trials, the target and the distractors are the same and in incongruent trials, they differ. We will use the letters "A" and "H" as stimulus targets and distractors. The flanker effect will be computed as the RT difference between the incongruent and congruent conditions. Participants will complete four practice trials and 200 experimental trials per session, half of which will be congruent and half of which will be incongruent. They will be instructed to answer as fast and as accurately as possible and will be given auditory and visual feedback after each trial.

Visual inspection time. In the VIT task (Vickers et al., 1972), participants are shown a stimulus consisting of two vertical parallel lines that differ in length and that are horizontally connected at the top. Their task is to indicate which line is longer. We will use the same task set-up as Garaas and Pomplun (2008). The distance between participants eyes and the screen will approximately be 60 cm, with a horizontal viewing angle of 31.5 degrees and a vertical viewing angle of 24.6 degrees. In each trial, participants will see a fixation cross followed by a brief stimulus presentation and a subsequent backward mask. The backward mask will consist of randomly overlapping stimuli and will prevent further processing of the stimulus (see Figure 9C). Participants can take as long as they want to respond as their VIT is computed based on the required presentation time to achieve a certain accuracy. We will use an accuracy level of 75 percent. The line lengths will be 3.4 degrees for the horizontal, and 5.1 and 6.8 degrees for the two vertical lines. To make the task more efficient, we will use an adaptive staircase algorithm as described in Garaas and Pomplun (2008). In the first experimental trial, the stimulus will be presented for 75 ms. Subsequently, presentation time will be increased by 8.3 ms after every two trials if either of the two trials was answered incorrectly and decreased if both trial were answered correctly. The task is completed once the following two presentation times can be identified: the presentation time where the accuracy is $\leq$ 75 percent in at least 36 trials and the

presentation time where the accuracy is $\geq$ 75 percent in at least 36 trials. Using these two times, we will estimate the presentation time for which participant respond accurately in exactly 75 percent (i.e., the VIT). Before the start of the experiment, participants will complete eight practice trials to familiarize themselves with the task.

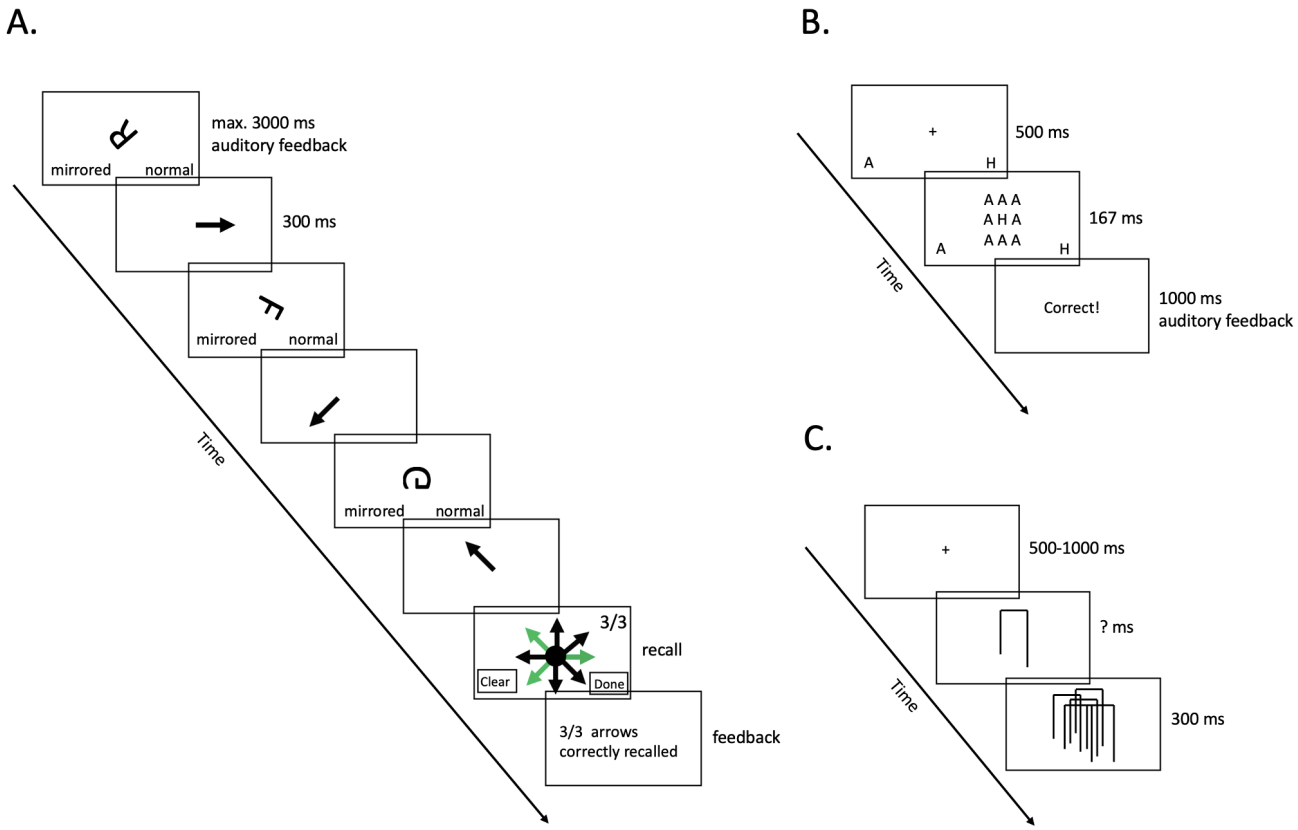

Figure 9. Example trials for **A.** the rotation-span task, **B.** the letter-flanker task, and **C.** the visual inspection time task.

Planned analysis

To explore correlations across the different tasks, we will estimate a multivariate hierarchical trial-level model using Stan (Carpenter et al., 2017) that explicitly models the covariation between the task parameters of interest. The model will contain the proposed inhibition task, the rotation span task, and the letter-flanker task. No trial-level model can be estimated for the VIT task since for this task, we only have one observation per

participant. However, we will correlate the VIT estimate with the individual hierarchical estimates from the other tasks.

Model specification. We specify the model using random variable notation. Let i, denote the individuals, $i = 1, ..., I$, j the trials, $j = 1, ..., J$, k the experimental conditions, $k = 1, ..., K$, and l the tasks, $l = 1, ..., L$. Furthermore, let $\boldsymbol{\theta}$ denote the effects of interests such that θ_{il} indicates participant i's effect in task l. For the proposed inhibition task, the effect of interest is the contrast effect, for the rotation span task, the probability of recalling an arrow correctly, and for the letter-flanker task, it is the congruency effect.

We start by describing the different likelihoods of the tasks. To model the proposed inhibition task, we again use a probit model,

$$y_{ijk1} \sim \text{Bernoulli}\left[\Phi(\alpha_i + \theta_{i1}u_k + \gamma_i x_l)\right],$$

where y_{ijk1} denotes participant i's response in the jth trial and kth condition, α_i denotes the individual intercepts, x_k the condition variable that is either -0.5 or 0.5, and θ_{i1} denotes the contrast effect. Finally, x_l is the target variable, ranging from -3.5 to 3.5, and γ_i is the individual target effect.

For the rotation span task, y_{ij2} denotes the number of successfully recalled arrows for participant i, in trial j. We model y_{ij2} using a binomial distribution,

$$y_{ij2} \sim \text{Binomial}(3, \Phi(\theta_{i2})),$$

with three subtrials and probit-transformed success probability θ_{i2}.

Finally, for the letter-flanker task, y_{ijk3} denotes individual i's response time in the jth trial and kth condition,

$$y_{ijk3} \sim \text{Normal}(\zeta_i + \theta_{i3}x_k, \tau^2).$$

Here, ζ_i is the intercept, θ_{i3} the flanker effect, x_k is the condition variable coded as -0.5 in the case of a congruent trial and as 0.5 in the case of incongruent trials, and τ^2 is the

trial-by-trial variation.

We model the parameters that are of primary interests, θ_{i1}, θ_{i2}, and θ_{i3}, using a multivariate-normal distribution,

$$\theta_{i1}, \theta_{i2}, \theta_{i3} \sim \text{Multivariate-Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ contains the population means $[\mu_1\ \mu_2\ \mu_3]^T$ and $\boldsymbol{\Sigma}$ is a 3x3 variance-covariance matrix.

To estimate the model, we need prior distributions for all model parameters. Our strategy is to use mildly informative priors that restrict the parameters to plausible ranges given their scales. For the population means $\boldsymbol{\mu}$, we will use the following normal priors:

$$\mu_1 \sim \text{Normal}(0, 0.7^2),$$

$$\mu_2 \sim \text{Normal}(0, 0.9^2),$$

$$\mu_3 \sim \text{Normal}(0, 0.1^2).$$

The prior on μ_1 anticipates the population mean of the contrast effect to lie somewhere between -2 and 2 with most weight assigned to values around zero. The prior on μ_2, the population mean of the probability of a correctly recalled arrow, assigns less prior mass to extreme probabilities such as zero and one. Lastly, the prior that we place on the population mean of the flanker effect, μ_3, reflects the finding that flanker effects typically lie somewhere between zero and 100 milliseconds (Rouder et al., 2019).

To estimate the variance-covariance matrix $\boldsymbol{\Sigma}$ efficiently in Stan, we do not directly place a prior on $\boldsymbol{\Sigma}$ but place separate priors on the standard deviations and on the cholesky-factor of the respective correlation matrix. We thereby use the fact that $\boldsymbol{\Sigma}$ can be decomposed into a correlation matrix and a matrix with the standard deviations of the parameters of interest on the diagonal,

$$\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma})\boldsymbol{\Omega}\text{diag}(\boldsymbol{\sigma}).$$

On the standard deviations $\boldsymbol{\sigma}$, we place a truncated multivariate-normal prior,

$$\boldsymbol{\sigma} \sim \text{Normal}_3^+ \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.7^2 & 0 & 0 \\ 0 & 0.7^2 & 0 \\ 0 & 0 & 0.7^2 \end{bmatrix} \right),$$

which allows each variance to be estimated freely within a plausible range.

We then further decompose $\boldsymbol{\Omega}$, such that $\boldsymbol{\Omega} = \text{LL}'$ and consequently,

$$\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma})\text{LL}'\text{diag}(\boldsymbol{\sigma}),$$

where L is the Cholesky factor of the correlation matrix and a lower triangular matrix with the same dimensions as $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$. We then place an LKJ-Cholesky prior distribution on L,

$$\text{L} \sim \text{LKJ-Cholesky}(3).$$

This parameterization implies $\text{LL}' \sim \text{LKJ}(3)$, and the shape parameter of 3 denotes a mildly informative prior that is skeptical regarding high correlations.

For the proposed inhibition task, we further need to specify priors for the intercept α_i, and the target effect γ_i. We place hierarchical priors on both parameters, $\alpha_i \sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2)$ and $\gamma_i \sim \text{Normal}(\mu_\gamma, \sigma_\gamma^2)$. The normal prior on the population mean of the intercept, $\mu_\alpha \sim \text{Normal}(0, 0.2^2)$, reflects our belief that on average, the bias towards either response option lies around 0.5 and that an average bias that is greater than 0.7 and smaller than 0.3 is rather implausible. On σ_α, we place a truncated normal $\sigma_\alpha \sim \text{Normal}^+(0, 0.3^2)$. This prior reflects the expectation that individual variability in the bias lies somewhere below a standard deviation of one with smaller values being more probable. The prior that we place on the population mean of the target effect, $\mu_\gamma \sim \text{Normal}(0, 0.7^2)$ anticipates μ_γ to lie somewhere between -2 and 2 with most probable values centered around 0. Lastly, we again use a truncated normal for the interindividual variation in the target effects, $\sigma_\gamma \sim \text{Normal}^+(0, 0.3^2)$, expecting the standard deviation to lie somewhere below a standard deviation of one.

The remaining model parameters that need a prior distribution are the intercept, ζ_i, and the trial-by-trial variation, τ^2, of the letter-flanker task. ζ_i represents the average response time, and we place a normal prior on it that constrains average response times to be around 900 milliseconds plus minus 1 second, $\zeta_i \sim \text{Normal}(0.9, 0.5^2)$ (in seconds). Lastly, we again use a truncated normal distribution as a prior for the trial-by-trial variation, $\tau \sim \text{Normal}^+(0, 0.4^2)$.

Intended Results

To examine whether the multivariate hierarchical model works as expected, we created a correlated dataset containing simulated data for the proposed inhibition, the rotation span, the letter-flanker, and the VIT task. We simulated data for 100 participants. For the proposed inhibition task, we thereby used the posterior medians of the morph-flanker parameters from the pilot study. To simulate data for the rotation span task, we sampled individual scores from a population distribution with a mean proportion correct of 0.68 as reported in Wilhelm et al. (2013) and a standard deviation of 0.1. For the letter-flanker task, we used the same values as Rouder et al. (2019) in their simulations. The individual intercepts were drawn from a population distribution with mean 0.80 (in seconds) and a standard deviation of 0.20 and the individual congruency effects from a population distribution with mean 0.05 (in seconds) and a standard deviation of 0.02. As trial noise, we specified a standard deviation of 0.18. These settings reflect a signal-to-noise ratio of 1 to 7.00, as commonly found in inhibition studies (Rouder et al., 2019). Finally, we sampled individual VIT scores from a normal distribution with mean 80.10 and standard deviation 23.40 (in milliseconds). These are values reported in the study by Garaas and Pomplun (2008). We further set the size of all correlations to 0.5. The correlations involving the WM task were further set to negative since higher scores indicate greater WM capacity. Note that this size of the correlation does not reflect our expectations regarding the absolute or relative sizes of the correlations that we will find. We merely used it to demonstrate the analyses

that we are planning to do and to check whether the proposed model manages to recover the true parameter values.

Table 3

Model estimation results for simulated data

	True Parameter	Posterior Median	SD	Lower Bound	Upper Bound
μ_1	0.718	0.679	0.032	0.617	0.743
μ_2	0.468	0.491	0.030	0.432	0.551
μ_3	0.050	0.046	0.003	0.039	0.053
σ_1	0.255	0.247	0.029	0.195	0.307
σ_2	0.303	0.289	0.023	0.248	0.340
σ_3	0.025	0.029	0.003	0.024	0.035

Note. True parameters, posterior medians, standard deviations, and 95% credible intervals of the population mean and variances for the proposed inhibition task (subscript 1), the rotation span task (subscript 2), and the letter-flanker task (subscript 3) as estimated using Stan. Note that the population means and standard deviations for the inhibition task and the rotation span task show the probit-transformed parameters.

We estimated the model using Stan (Carpenter et al., 2017). We ran 4 chains with 4000 iterations each, of which 1500 were used as warm-up to calibrate the sampler. Consequently, inference is based on 10000 samples. There were no signs of non-convergence: there were no divergent transitions, the largest *R-hat* value was 1.00 and the smallest number of estimated effective sample size was 1,439.02. The "true" population mean and variances and the estimated posterior medians including standard deviation and credible intervals are shown in Table 3. There are no severe differences between the posterior medians and the true parameter values. This suggests that the model worked as expected.

Importantly, the model also managed to recover the task correlations. Figure 10 shows the true correlations, the correlations based on the observed effects and the ones based on the individual hierarchical estimates of the effects. The correlations between the observed effects are considerably smaller than the true correlations. This indicates that the observed ones are attenuated by trial noise. The hierarchical model takes the trial noise into account and thereby disattenuates the correlations. Notably, the model even slightly overestimates the correlation between the WM score and the contrast effect and the correlation between the flanker effect and the contrast effect. An explanation for these higher values is that the correlations are computed based on single individual point estimates without directly taking the uncertainty associated with each estimate into account. Figure 12 therefore shows the posterior distribution from the population-level correlation parameter. Here the model could recover the correlations more precisely, except the posterior median of the correlation between the working memory score and the flanker effect was slightly underestimated. Overall, all posterior distributions mostly lie on one side of zero, suggesting that the model was able to detect the presence of the correlations.

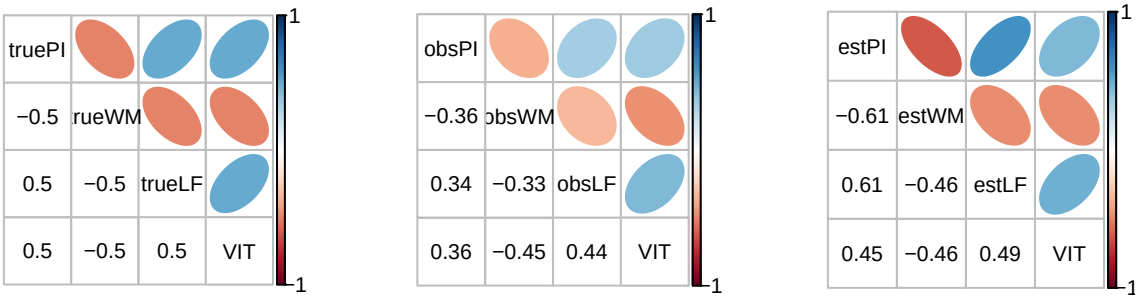

Figure 10. True, observed, and model-based task correlations based on simulated data. PI = proposed inhibition task, LF = letter-flanker task, WM = working memory task, VIT = visual inspection time task. The true correlations were set to 0.5 and -0.5, respectively. Note that for visual inspection time, there was only one observation. Therefore, no trial-level hierarchical model was estimated.

Figure 12 shows the observed and estimated effects in increasing order, sorted

according to the observed effects. For the proposed inhibition task, we again computed the observed and model-based contrast effects on the probability scale and based on the middle four targets. For the proposed inhibition task and the flanker task, the extent of individual variation in the individual estimates is smaller than the extent of individual variation in the observed estimates. The estimates are drawn towards the population mean. This indicates hierarchical shrinkage and suggests that a large extent of the variation in the observed effects was due to trial noise. For the rotation span task, there is less shrinkage, which was to be expected given the large true individual differences.

While the results of this small simulation seem encouraging, we should keep in mind that the data were simulated under highly ideal conditions. For example, we used the same likelihoods as the model to create the data. With real data, the model might not perform that well because real data usually are noisier. Therefore, a way to further advance the presented model is to extend it to a factor model. An advantage of a factor model is that it is possible to constrain the covariance matrix by a priori determining the number of factors. This decreases model flexibility. In noisy data, the correlations may therefore be better localized (Rouder et al., 2019). In the further course of this project, we plan to develop a one-factor model and to compare its performance to the pure estimation model introduced here. A one-factor model implies that the cognitive tasks are solely correlated via one latent variable which may reflect general task performance. If we find that a latent factor loads onto the proposed inhibition task as well as on the other three cognitive tasks in the task battery, we can conclude that the tasks are indeed correlated.

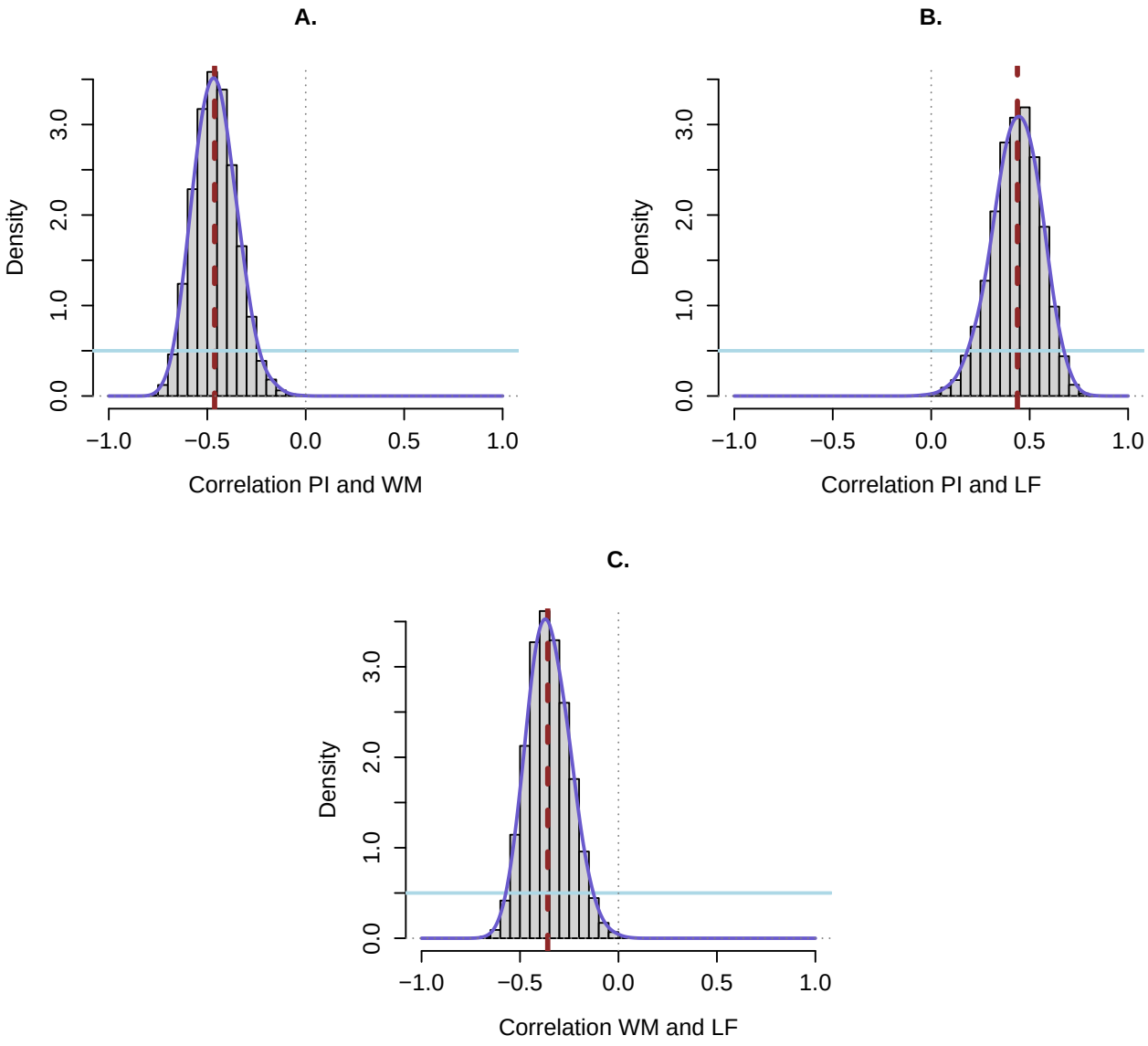

Figure 11. Posterior and prior distributions of the correlation between the simulated effects as estimated by the hierarchical trial-level model. The true correlations were set to 0.5 and -0.5, respectively. The dotted vertical lines denote the medians and serve as point estimates. PI = proposed inhibition task, WM = working memory task, LF = letter-flanker task. **A.** shows the correlation between the contrast effect and the rotation-span score. **B.** shows the correlation between the contrast effect and the flanker/assimilation effect. **C.** shows the correlation between the rotation-span score and the flanker effect.

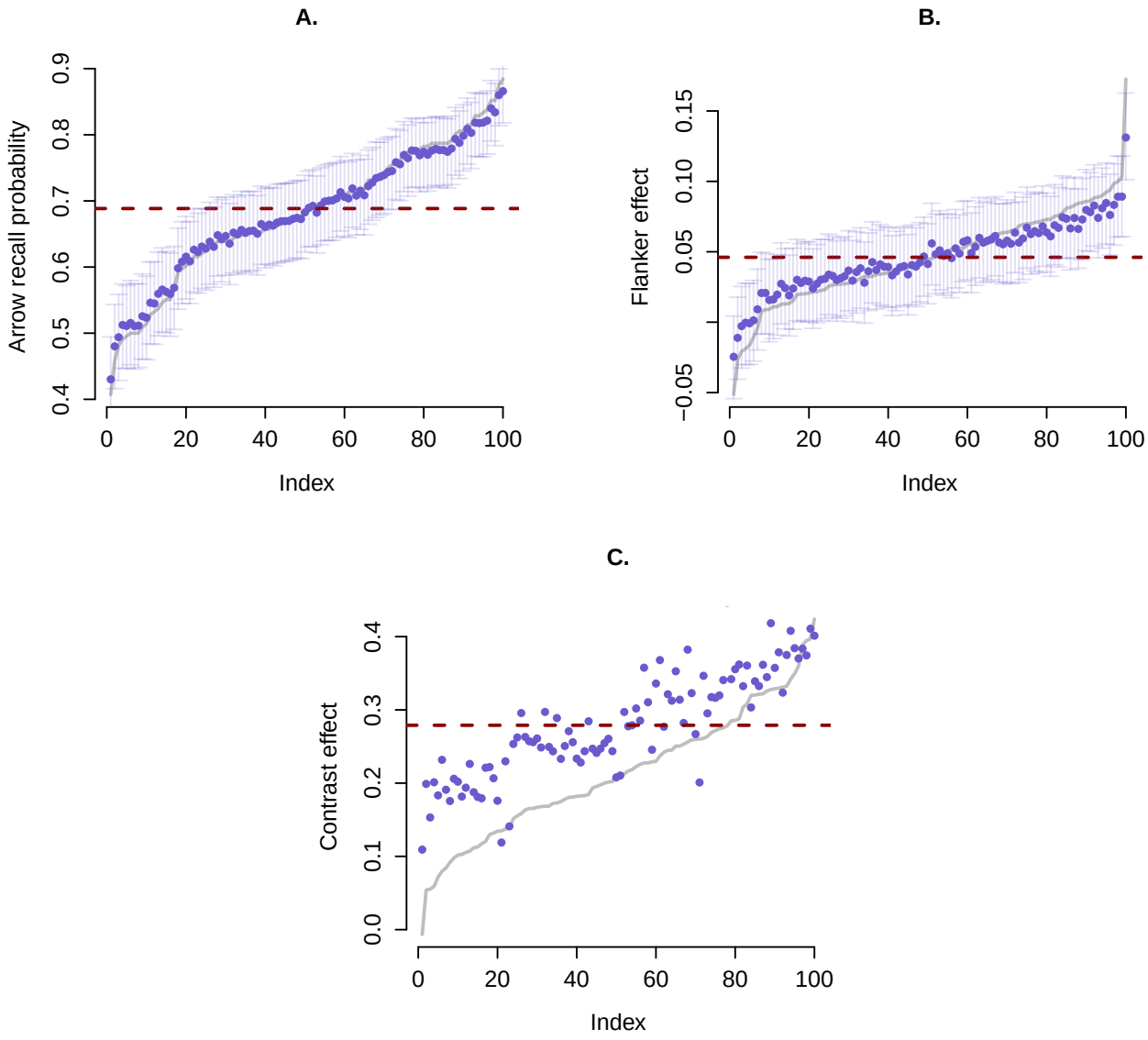

Figure 12. Estimated and 'observed' effects for the simulated data. The blue dots denote individuals' posterior medians and the vertical lines denote the 95% credible intervals. The estimates are ordered according to the observed effects shown in grey. The dashed lines denote the posterior medians. The difference between the observed and model-based effects indicates hierarchical shrinkage. **A.** shows the individual success probabilities in the rotation-span task, **B.** the assimilation effects in the letter-flanker task, and **C.**, the contrast effects in the proposed inhibition task. Note that for the proposed inhibition task, the contrast effect is not an immediate reflection of a single model parameter. It is computed based on the intercept, the contrast and target effects and based on the middle four targets. Here, the dashed line denotes the average of those effects.

Conclusion

In this thesis, we have introduced the morph-flanker and the brightness-illusion task as two new potential inhibition tasks. Our hope was that those paradigms would yield larger overall effects and more individual differences than what is commonly found in inhibition studies. The results of a small pilot study indeed suggested that the two tasks yielded relatively more individual differences than standard inhibition tasks such as the Stroop and flanker. Furthermore, performance in the two tasks was highly correlated. However, an unexpected finding was that only the morph-flanker and not the brightness-illusion task yielded a large overall effect. The brightness-illusion task yielded substantial individual variation even though there was practically no overall effect. As we proceed with this project, we will therefore make the brightness-illusion task more difficult. Perhaps, this will increase the overall contrast effect and still yield substantial individual variation. We will then collect more data to examine whether we can corroborate the promising findings from the pilot study.

Overall, the results from the pilot study are initial evidence that tasks designed to assess visual contrast effects, such as the morph-flanker and the brightness-illusion task, may be better at distinguishing between indivuals than standard inhibition tasks. The proposed large-scale correlation study will further show whether the tasks truly are well-suited for assessing individual differences in cognitive inhibition and, ultimately, whether they can be used to address the question whether inhibition is a unified concept or whether it comprises distinct conceptual phenomena.

References

Albert, J. H., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, *82*, 747–759.

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. doi:10.32614/RJ-2018-017

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). doi:10.18637/jss.v076.i01

Chuderski, A. (2014). The relational integration task explains fluid reasoning above and beyond other working memory tasks. *Memory & Cognition*, *42*(3), 448–463. doi:10.3758/s13421-013-0366-x

Cowan, N. (1995). *Attention and memory: An integrated framework*. Oxford University Press.

Dickey, J. M. (1976). Approximate posterior distributions. *Journal of the American*

Statistical Association, *71*(355), 680–689. Retrieved from http://www.jstor.org/stable/2285601

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What item response theory can tell us about the complex span tasks. *Psychological Assessment*, *30*(1), 116–129. doi:10.1037/pas0000444

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508.

Eddelbuettel, D., & Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, *5*, e3188v1. doi:10.7287/peerj.preprints.3188v1

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. doi:10.18637/jss.v040.i08

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19–23. doi:10.1111/1467-8721.00160

Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psychological Science*, *13*(2), 190–193. doi:10.1177/1745691617720478

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology:*

General, *133*, 101–135.

Garaas, T. W., & Pomplun, M. (2008). Inspection time and visualPerceptual processing. *Vision Research*, *48*(4), 523–537. doi:10.1016/j.visres.2007.11.011

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Heidelberg: Springer-Verlag.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798.

Haaf, J. M., & Rouder, J. N. (2018). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-018-1522-x

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 22, pp. 193–225). Academic Press.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavioral Research Methods*.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2019). *Lab.Js: A free, open, online study builder* (Preprint). PsyArXiv. doi:10.31234/osf.io/fqr49

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*(8), 1–29. Retrieved from http://www.jstatsoft.org/v38/i08/

Joormann, J., Yoon, K. L., & Zetsche, U. (2007). Cognitive inhibition in depression. *Applied and Preventive Psychology*, *12*(3), 128–139. doi:10.1016/j.appsy.2007.09.002

Kahle, D., & Stamey, J. (2017). *Invgamma: The inverse gamma distribution.* Retrieved from https://CRAN.R-project.org/package=invgamma

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. E. (2004). The generality of working-memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189–217.

Keye, D., Wilhelm, O., Oberauer, K., & van Ravenzwaaij, D. (2009). Individual differences in conflict-monitoring: Testing means and covariance hypothesis about the Simon and the Eriksen Flanker task. *Psychological Research*, *73*, 762–776. doi:10.1007/s00426-008-0188-9

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*(4), 477.

MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, *42*(9), 22. Retrieved from http://www.jstatsoft.org/v42/i09/

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of

measurement error and estimation uncertainty. *Collabra: Psychology*, *3*(1).

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 196–204. doi:10.1037/a0014104

McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology: General*, *141*(2), 302–320. doi:10.1037/a0025250

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin*, *126*(2), 220–246. doi:10.1037//0033-2909.126.2.220

Ooms, J. (2019). *Curl: A modern and flexible web client for r*. Retrieved from https://CRAN.R-project.org/package=curl

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*, 232–258. doi:10.1016/j.cogpsych.2012.12.002

Pettigrew, C., & Martin, R. C. (2014). Cognitive declines in healthy aging: Evidence from multiple aspects of interference resolution. *Psychology and Aging*, *29*(2), 187.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Retrieved from http://dx.doi.org/10.1037/xlm0000450

Roberts, G. O., & Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B, Methodological*, *59*, 291–317.

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavioral Research Methods*, *48*, 1062–1069. Retrieved from 10.3758/s13428-015-0630-z

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*, 452–467. doi:10.3758/s13423-018-1558-y

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*, 41–56. Retrieved from https://doi.org/10.1080/03637751.2017.1394581

Rouder, J. N., & King, J. W. (2003). Flanker and negative flanker effects in letter identification. *Perception & Psychophysics*, *65*(2), 287–297.

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. *PsyArXiv*. doi:10.31234/osf.io/3cjr5

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an

application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal Detection Models with Random Participant and Item Effects. *Psychometrika*, *72*(4), 621–642. doi:10.1007/s11336-005-1350-6

Salthouse, T. A. (1996). The processing speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403–428.

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*, 374–408. doi:10.1037/a0034418

Singmann, Henrik. (2018, January). Diffusion/Wiener Model Analysis with brms : Model Diagnostics and Model Fit.

Snyder, H. K., Rafferty, S. M., Haaf, J. M., & Rouder, J. N. (2019). Common or distinct attention mechanisms for contrast and assimilation? *Attention, Perception, & Psychophysics*, *81*, 1944–1950. doi:10.3758/s13414-019-01713-8

Spearman, C. (1904). 'General intelligence,' obectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Stahl, C., Voss, A., Schmitz, F., Nuszbaum, M., Tüscher, O., Lieb, K., & Klauer, K. C. (2014). Behavioral components of impulsivity. *Journal of Experimental Psychology: General*, *143*(2), 850.

Stan Development Team. (2019a). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Stan Development Team. (2019b). StanHeaders: Headers for the R interface to Stan.

Retrieved from http://mc-stan.org/

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*, 1701–1728.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26. doi:10.1016/j.cogpsych.2014.01.003

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Vickers, D., Nettelbeck, T., & Willson, R. J. (1972). Perceptual Indices of Performance: The Measurement of "Inspection Time" and "Noise" in the Visual System. *Perception*, *1*(3), 263–295. doi:10.1068/p010263

von Bastian, C. C., Blais, C., Brewer, G. A., Gyurkovics, M., Hedge, C., Kałamała, P., ... Wiemers, E. A. (2020). *Advancing the understanding of individual differences in attentional control: Theoretical, methodological, and analytical considerations* (Preprint). PsyArXiv. doi:10.31234/osf.io/x3b9k

Wei, T., & Simko, V. (2017). *R package "corrplot": Visualization of a correlation matrix*. Retrieved from https://github.com/taiyun/corrplot

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2018). *Scales: Scale functions for visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00433

Appendices

Appendix A

Accuracy based on cleaned data. Figure 13 shows accuracy as a function of the target and distractor condition. The plots show that the contrast effects based on the propensity of "A"/"bright" responses and the contrast effects based on accuracy are comparable in size.

Descriptive plots based on uncleaned data. Figure 14 shows the proportion of "A"/"bright" responses as a function of target and distractor in the full sample. Warm-up trials and double runs have been excluded.

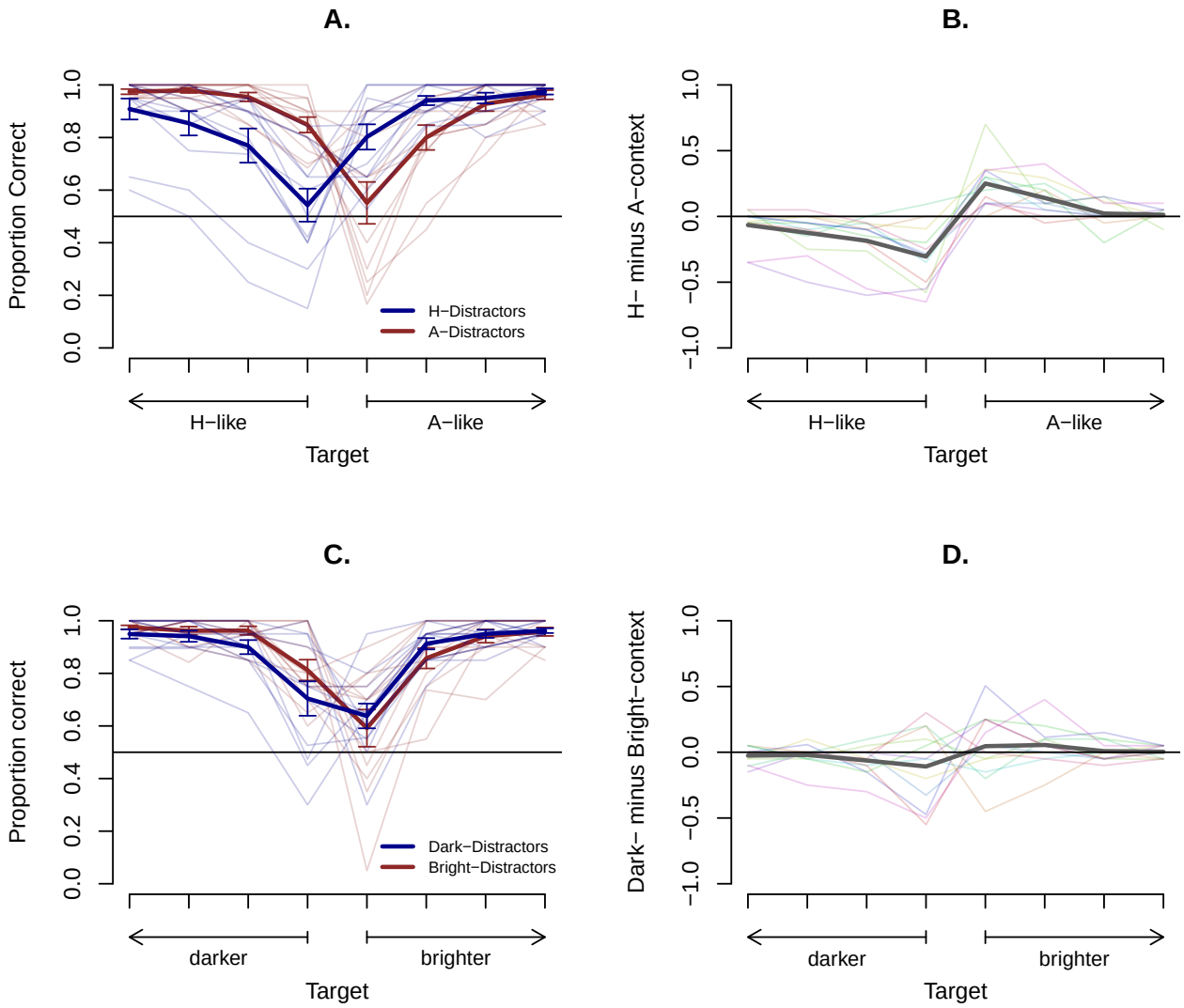

Figure 13. The plots show the proportion of accurate responses as a function of the target and the distractor condition. **A.** and **C.** show the proportion of correct responses in the morph-flanker and the brightness-illusion task, respectively, for each target. Each line represents a participant's average accuracy in either of the distractor conditions. The thick lines denote the mean accuracy per condition and across participants. The error bars denote the standard errors. The contrast effect for each target is reflected in the vertical distance between the blue and red lines. **B.** and **D.** show the inhibition effects across the two tasks. For each target and task, we subtracted the proportion of correct responses in one condition from the other. Each line represents a participant and the thick grey lines shows the mean contrast effects.

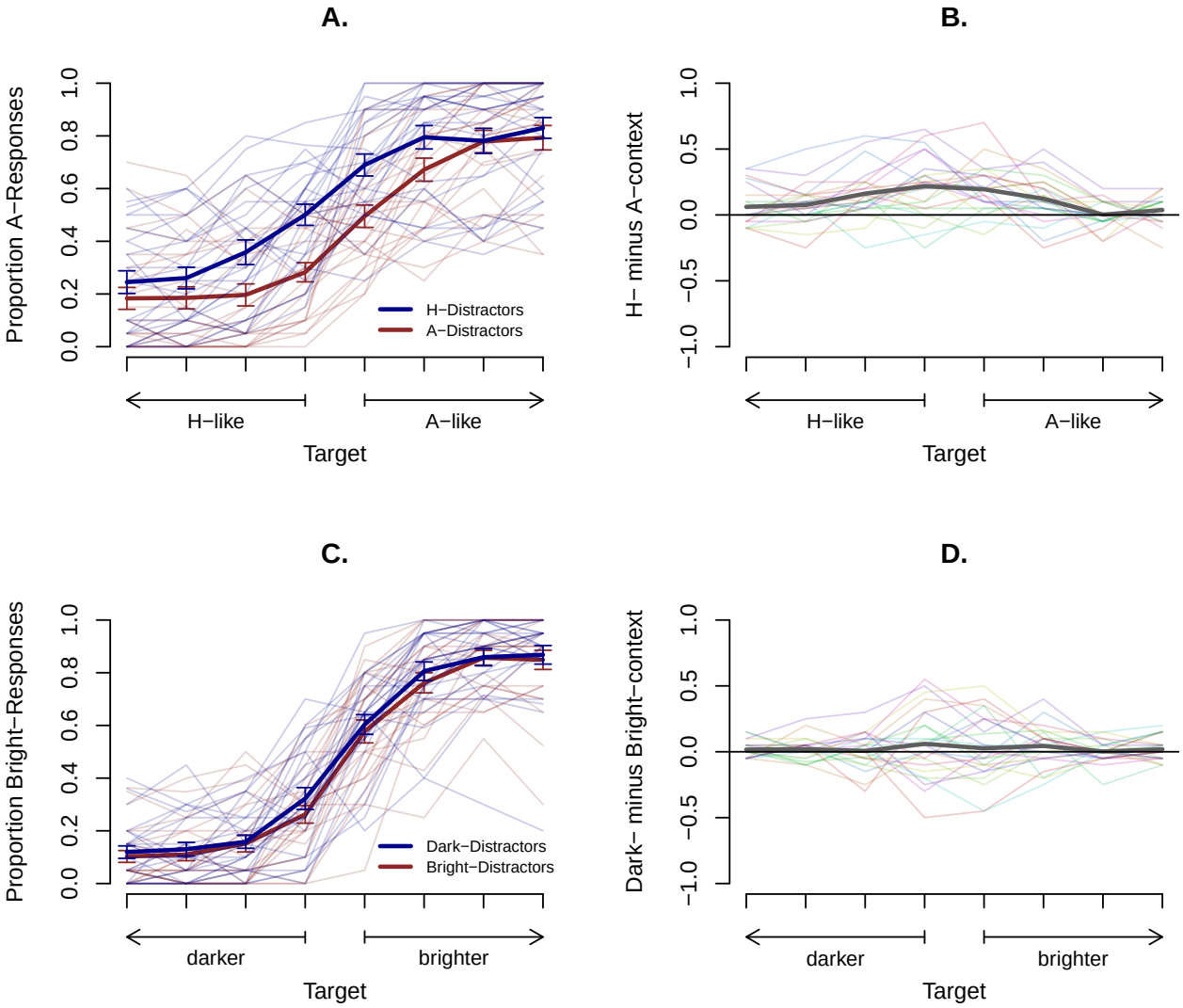

Figure 14. **A.** and **C.** show the proportion of 'A' responses in the morph-flanker task and the proportion of 'bright' responses in the brightness-illusion task, respectively, for each target. Each line represents a participant's average response in either of the distractor conditions. The thick lines denote the mean proportion per condition and across participants. The error bars denote the standard errors. The contrast effect for each target is reflected in the vertical distance between the blue and red lines. **B.** and **D.** show the inhibition effects across the two tasks. For each target, we subtracted the proportion of 'A' or 'bright' responses, respectively, in one condition from the other. Each line represents a participant and the thick grey lines shows the mean contrast effects.

Appendix B

In this Appendix, we first describe the Gibbs sampling estimation procedure. We then present the results of the model re-estimation using Stan.

Model estimation using Gibbs sampling. We estimated the hierarchical trial-level probit model using Gibbs sampling (Geman & Geman, 1984). Gibbs sampling is a Markov-Chain-Monte-Carlo (MCMC) algorithm that can be used to draw samples from the posterior distribution of model parameters of interest. These samples approximate the true marginal posterior and can be used for estimation and inference (Rouder & Lu, 2005).

To implement a Gibbs sampling algorithm, conjugate priors need to be chosen and the conditional posteriors need to be known for each model parameter (i.e., the posterior distributions of each parameter conditional on the other model parameters). Having specified the conditionals, one chooses random start values for each parameter. These start values are used to initiate the sampler. The algorithm is set up such that it iteratively takes random samples from the conditional posteriors. For example, say we want to estimate the marginal posteriors of a mean μ and of a variance σ^2 of normally distributed data $\boldsymbol{Y}$. We first define the conditionals $f(\mu|\sigma^2, \boldsymbol{Y})$ and $f(\sigma^2|\mu, \boldsymbol{Y})$, and then pick a random start value for σ^2. Conditional on this start value and the data, the algorithm draws a random sample for μ from the conditional posterior $f(\mu|\sigma^2, \boldsymbol{Y})$. The resulting value for μ is in turn used to sample from the conditional posterior $f(\sigma^2|\mu, \boldsymbol{Y})$, and so on. Perhaps unsurprisingly, the first samples will be majorly influenced by the start value. But after a certain number of iterations, the Gibbs sampler will draw from the true marginal posterior distributions of μ and σ^2 (Tierney, 1994). Once the sampler has reached this high-density regions of the marginal distributions, it will stay there and continue to draw samples from that region.

We are interested in estimating the marginal posterior distribution of all specified parameters in our model. Those are the intercepts α_{ij}, the contrast effects β_{ij}, the effects of the targets on the response γ_{ij}, the population means μ_{α_j} and μ_{β_j}, μ_{γ_j}, and the

variance-covariance matrix $\mathbf{\Sigma}_\beta$. To derive the posterior conditionals, it is convenient and in most cases computationally necessary to use the proportional form of the Bayes theorem, which states that the posterior distribution of a parameter is proportional to its prior distribution times the likelihood. To estimate the parameters that constitute η_{ijkl} (i.e., α_{ij}, β_{ij}, and γ_{ij}.), we used an analysis approach by Albert and Chib (1995) for estimating a probit-transformed probability. This approach can circumvent the problem that in the case of our bernoulli-trial probit model, the product of the prior distribution and the likelihood is intractable. The idea of Albert and Chib (1995)'s approach is that for each trial Y_{ijklm}, a new latent variable, w_{ijklm}, is defined as follows,

$$w_{ijklm} > 0 \qquad \text{if } Y_{ijklm} = 1$$

$$w_{ijklm} < 0 \qquad \text{if } Y_{ijklm} = 0.$$

With no loss of generality, $\boldsymbol{w}$ is assumed to be normally distributed with mean $\boldsymbol{\eta}$ and variance 1, $w_{ijklm} \sim \text{Normal}(\eta_{ijkl}, 1)$. We are aiming for the conditional distributions $f(\boldsymbol{w}|\boldsymbol{\eta}, \boldsymbol{Y})$ and $f(\boldsymbol{\eta}|\boldsymbol{w}, \boldsymbol{Y})$ to then approximate the marginals using Gibbs sampling. We know that independent of Y_{ijklm}, w_{ijklm} is normally distributed and that its mean is η_{ijkl}. When w_{ijklm} is conditioned on Y_{ijklm}, we also know whether w_{ijklm} is positive or negative (Rouder & Lu, 2005). Therefore, the conditional posterior $f(w_{ijklm}|\eta_{ijkl}, Y_{ijklm})$ follows a truncated normal distribution as follows,

$$f(w_{ijklm}|\eta_{ijkl}, Y_{ijklm}) \sim \text{Normal}^{+}(\eta_{ijkl}, 1) \qquad \text{if } Y_{ijklm} = 1,$$

$$f(w_{ijklm}|\eta_{ijkl}, Y_{ijklm}) \sim \text{Normal}^{-}(\eta_{ijkl}, 1) \qquad \text{if } Y_{ijklm} = 0.$$

To sample from the conditional posteriors of the variables that constitute $\boldsymbol{\eta}$, we used blocked sampling, a more efficient form of Gibbs sampling (Roberts & Sahu, 1997; Rouder & Lu, 2005). In blocked sampling, all parameters are updated at once instead of separately. Let $\boldsymbol{\theta}$ be a vector containing the individual and task specific intercepts, contrast effects, and task-specific target effects, $\boldsymbol{\theta} = \{\alpha_{ij}, \beta_{ij}, \gamma_{ij}\}$, such that $\boldsymbol{\eta} = \boldsymbol{\theta}\boldsymbol{X}$, whereby $\boldsymbol{X}$ is the design

matrix. $\boldsymbol{\theta}$ follows a multivariate normal distribution,

$$f(\boldsymbol{\theta}) \sim \text{Multivariate-Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ is a vector containing all population means of the parameters and $\boldsymbol{\Sigma}$ is the corresponding variance-covariance matrix. Note that the only non-zero-covariances in $\boldsymbol{\Sigma}$ are the ones that we estimated among the contrast effects β_{ij} and the one we estimated among the target effects γ_{ij}. The conditional posterior of $\boldsymbol{\theta}$, $f(\boldsymbol{\theta}|.)$ then follows a multivariate normal distribution,

$$f(\boldsymbol{\theta}|.) \sim \text{Multivariate-Normal}(\boldsymbol{CV}, \boldsymbol{V}),$$

where $\boldsymbol{C} = (\boldsymbol{X'w} + \boldsymbol{\Sigma}^{-1}) + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$ and $\boldsymbol{V} = (\boldsymbol{X'X} + \boldsymbol{\Sigma}^{-1})^{-1}$. We further define the conditional posterior distributions of the hyperparameters, the population means and variances. For μ_{α_j}, the conditional posterior distribution is, $f(\mu_{\alpha j}|.) \propto f(\mu_{\alpha_j}) f(\alpha_{ij}|\mu_{\alpha_j}, \sigma^2_{\alpha_j})$. If we insert the normal density function and the normal likelihood we obtain the following:

$$f(\mu_{\alpha_j}|.) \propto \frac{1}{\sqrt{2\pi c_2}} \exp\left(\frac{-(\mu_{\alpha_j} - c_1)^2}{2c_2}\right) \times \frac{1}{(2\pi\sigma^2_{\alpha_j})^{I/2}} \exp\left(\sum_{i=1}^{I} \frac{-(\alpha_{ij} - \mu_{\alpha_j})^2}{2\sigma^2_{\alpha_j}}\right),$$

where $c1$ and $c2$ are the prior values. By removing the terms that do not depend on μ_{α_j} as proportionality constants, rearranging, and completing the square we obtain that the posterior distribution $f(\mu_{\alpha j}|.)$ follows again a normal distribution,

$$f(\mu_{\alpha_j}|.) \sim \text{Normal}(cv, v),$$

where $c = \left[\frac{c_1}{c_2} + \frac{\Sigma_i^I \alpha_{ij}}{\sigma^2_{\alpha j}}\right]$ and $v = \left[\frac{I}{\sigma^2_{\alpha_j}} + \frac{1}{c_2}\right]^{-1}$. Using the same logic as above, the conditional posterior distribution for $\sigma^2_{\alpha_j}$ is proportional to $f(\sigma^2_{\alpha_j}|.) \propto f(\sigma^2_{\alpha j}) f(\alpha_{ij}|\mu_{\alpha j}, \sigma^2_{\alpha j})$. Again inserting the density of the gamma distribution and the normal likelihood, we obtain

$$f(\sigma^2_{\alpha_j}|.) \propto \frac{c_4^{c_3}}{\Gamma(c_3)\left(\sigma^2_{\alpha_j}\right)^{c_3+1}} \exp\left(\frac{-c_4}{\sigma^2_{\alpha_j}}\right) \times \frac{1}{(2\pi\sigma^2_{\alpha_j})^{I/2}} \exp\left(\sum_{i=1}^{I} \frac{-(\alpha_{ij} - \mu_{\alpha_j})^2}{2\sigma^2_{\alpha_j}}\right).$$

Dropping all terms that do not depend on $f(\sigma^2_{\alpha_j})$ and rearranging them, we obtain that $f(\sigma^2_{\alpha_j}|.)$ follows an inverse-gamma distribution,

$$f(\sigma^2_{\alpha_j}|.) \sim \text{Inverse-Gamma}(q, s),$$

where $q = c_3 + \left[\frac{I}{2}\right]$ is the updated shape parameter and $s = c_4 + 0.5\left[\sum_{i=1}^{I}(\alpha_{ij} - \mu_{\alpha j})^2\right]$ the updated rate parameter. c_3 is the prior value for the shape and c_4 the prior value for the rate.

For μ_{β_j}, the conditional posterior distribution is proportional to $f(\mu_{\beta j})f(\beta_{ij}|\mu_{\beta j}, \boldsymbol{\Sigma_\beta})$. If we again discard the terms that are independent of $\mu_{\beta j}$ into the proportionality constant, we can derive that μ_{β_j} follows a normal distribution,

$$f(\mu_{\beta j}|.) \sim \text{Normal}(cv, v),$$

with $c = \left[\frac{c_1}{c_2} + \frac{\sum_i^I \beta_{ij}}{\sigma^2_{\beta_j}}\right]$ and $v = \left[\frac{I}{\sigma^2_{\beta_j}} + \frac{1}{c_2}\right]^{-1}$ whereby c_1 and c_2 are again the specified prior values for the mean and variance, respectively.

Lastly, for the variance-covariance parameter, the conditional posterior distribution $f(\boldsymbol{\Sigma_\beta}|.)$ is proportional to $f(\boldsymbol{\Sigma_\beta})f(\beta_{ij}|\mu_{\beta_j}, \boldsymbol{\Sigma_\beta})$. By again rearranging and discarding terms that are independent of $\boldsymbol{\Sigma_\beta}$, we can obtain that the conditional posterior distribution of the variance-covariance matrix follows again an inverse-wishart distribution,

$$f(\boldsymbol{\Sigma_\beta}|.) \sim \text{Inverse-Wishart}(d, \boldsymbol{M}),$$

whereby $d = I + c_5$ is the updated degrees of freedom and $M = \boldsymbol{\Omega} + \sum_{i=1}^{I}(\boldsymbol{\beta_i} - \boldsymbol{\mu_\beta})(\boldsymbol{\beta_i} - \boldsymbol{\mu_\beta})^T$ the scale matrix, with with the prior parameters $\boldsymbol{\Omega}$ for the scale matrix and c_5 for the degrees of freedom. $\boldsymbol{\beta_i}$ is a vector of length $J = 2$ that contains a β-parameter for each task.

For μ_{γ_j}, the derivation is the same as for μ_{β_j}. The conditional posterior distribution is proportional to $f(\mu_{\gamma_j})f(\gamma_{ij}|\mu_{\gamma j}, \boldsymbol{\Sigma_\gamma})$. Discarding the terms independent of $\mu_{\gamma j}$ into the proportionality constant, we can derive that μ_{γ_j} follows a normal distribution,

$$f(\mu_{\gamma j}|.) \sim \text{Normal}(cv, v),$$

with $c = \left[\frac{c_1}{c_2} + \frac{\sum_i^I \gamma_{ij}}{\sigma^2_{\gamma_j}}\right]$ and $v = \left[\frac{I}{\sigma^2_{\gamma_j}} + \frac{1}{c_2}\right]^{-1}$ whereby c_1 and c_2 are again the specified prior values for the mean and variance, respectively.

Lastly, for the variance-covariance parameter, the conditional posterior distribution $f(\mathbf{\Sigma_\gamma}|.)$ is proportional to $f(\mathbf{\Sigma_\gamma})f(\gamma_{ij}|\mu_{\gamma_j}, \mathbf{\Sigma_\gamma})$. By again rearranging and discarding terms that are independent of $\mathbf{\Sigma_\gamma}$, we can obtain that the conditional posterior distribution of the variance-covariance matrix follows again an inverse-wishart distribution,

$$f(\mathbf{\Sigma_\gamma}|.) \sim \text{Inverse-Wishart}(d, \boldsymbol{M}),$$

whereby $d = I + c_5$ is the updated degrees of freedom and $M = \mathbf{\Omega} + \sum_{i=1}^{I}(\boldsymbol{\gamma_i} - \boldsymbol{\mu_\gamma})(\boldsymbol{\gamma_i} - \boldsymbol{\mu_\gamma})^T$ is the scale matrix, with the prior parameters $\mathbf{\Omega}$ for the scale matrix and c_5 for the degrees of freedom. $\boldsymbol{\gamma_i}$ is a vector of length $J = 2$ that contains a γ-parameter for each task.

Model estimation using Stan. As a cross-check, we also estimated the hierarchical model in Stan, using the same prior settings. We estimated 4 chains with 7000 iterations each, of which 2000 were used as warm-up to calibrate the sampler. Consequently, inference is based on 20000 samples. Table 4 and 5 show summaries of the population means and variances. There were no signs of non-convergence: there were no divergent transitions, the largest *R-hat* value was 1.00 and the smallest number of estimated effective sample size was 5,657.21. The Stan results are comparable to the Gibbs sampling results (see Figure 15 and 16). One noticeable difference is that Stan estimated the correlation coefficient between the contrast effects slightly higher than the Gibbs sampler. The posterior median of ρ was 0.86, $SD = 0.17$, 95% CrI[0.32, 0.97].

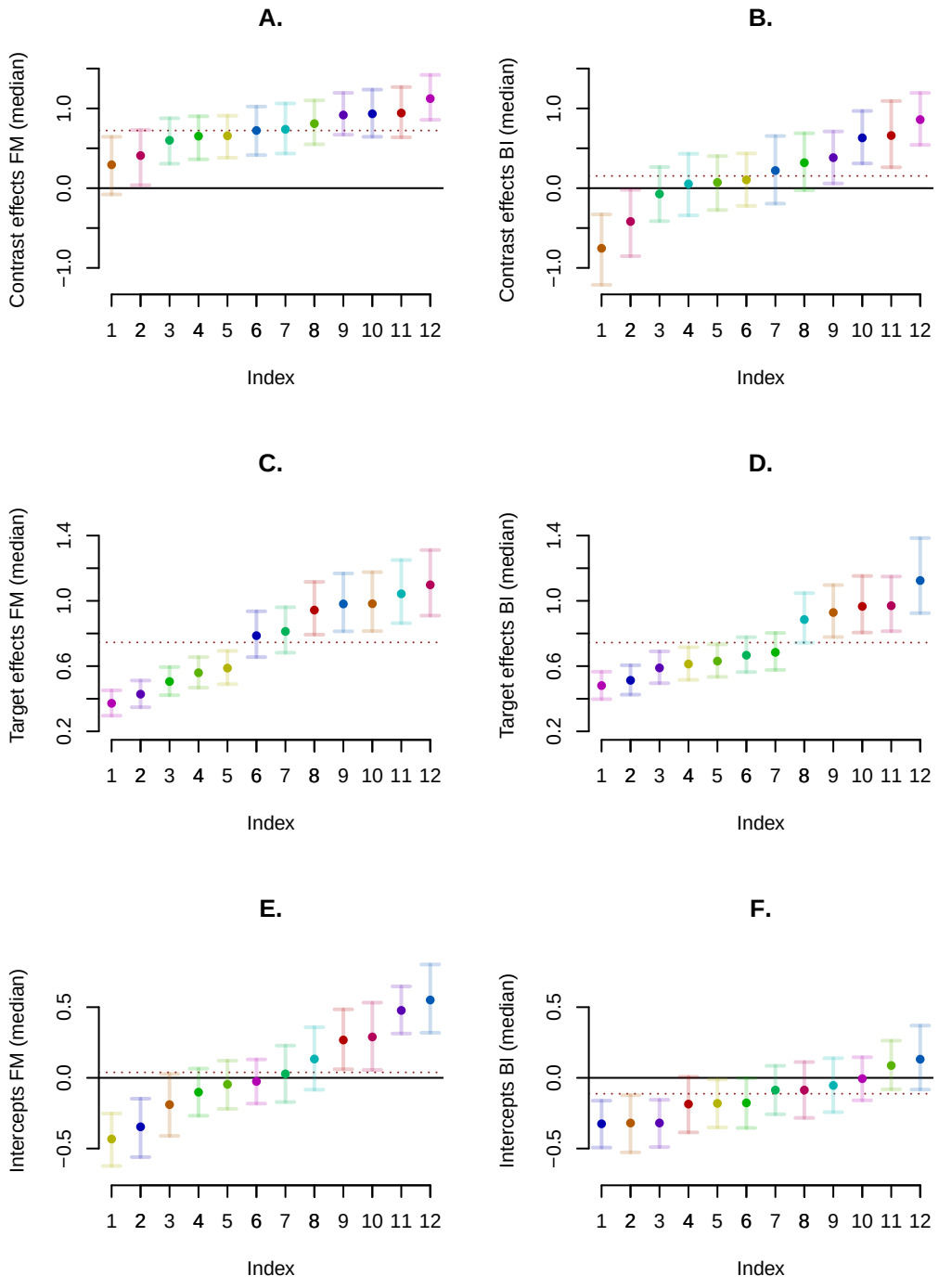

Figure 15. Figure **A.** shows the individual contrast effects (posterior medians) in increasing order for the morph-flanker task and **B.** for the brightness-illusion task as estimated by Stan. **C.** and **D.** show the target effects and **E.** and **F.** the intercepts. Each dot represents a participant. The horizontal lines denote the 95% credible intervals. The dotted line shows the posterior median of the population mean of the contrast effect.

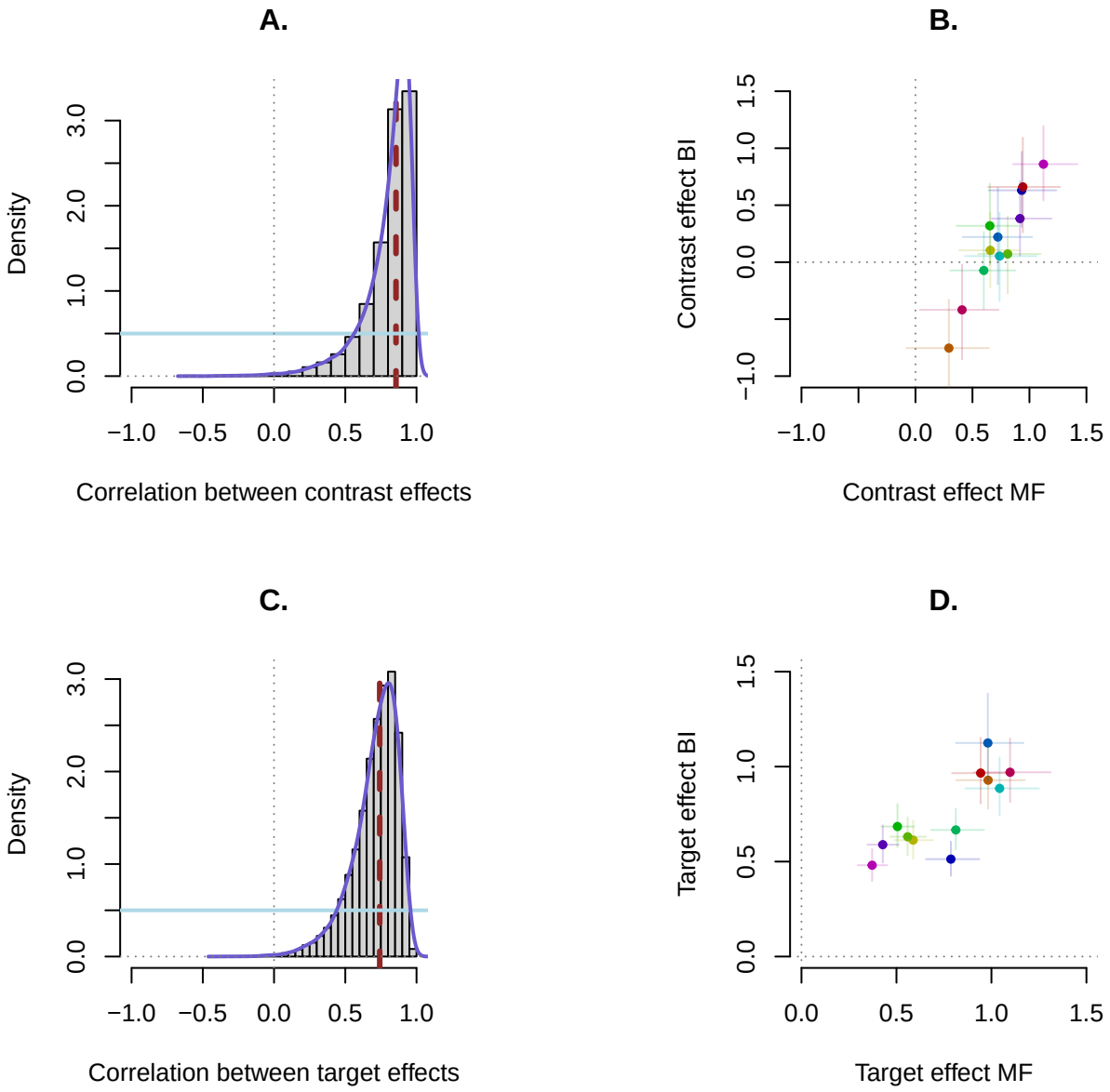

Figure 16. The plots show the correlation between the contrast and target effects as estimated by Stan. **A.** and **C.** show the prior and posterior distributions. MF indicates morph-flanker and BI brightness-illusion. The dotted vertical line denotes the median and serves as a point estimate. **B.** and **D.** show the individual estimates (median) of the effects in the morph-flanker and the brightness-illusion task. Each dot represents a participant. The horizontal and vertical lines denote the 95% credible intervals.

Table 4

Model estimation results for the morph-flanker task

	Median	SD	Lower Bound	Upper Bound
μ_{α_1}	0.039	0.076	-0.114	0.187
μ_{β_1}	0.723	0.102	0.511	0.915
μ_{γ_1}	0.745	0.082	0.580	0.907
$\sigma^2_{\alpha_1}$	0.081	0.038	0.039	0.184
$\sigma^2_{\beta_1}$	0.067	0.055	0.019	0.226
$\sigma^2_{\gamma_1}$	0.067	0.037	0.030	0.170

Note. Posterior medians, standard deviations, and 95% credible intervals of the population mean and variances for the morph-flanker task (indicated by the subscript 1) as estimated using Stan.

Table 5

Model estimation results for the morph-flanker task

	Median	SD	Lower Bound	Upper Bound
μ_{α_2}	-0.112	0.058	-0.226	0.006
μ_{β_2}	0.153	0.152	-0.159	0.446
μ_{γ_2}	0.744	0.070	0.607	0.885
$\sigma^2_{\alpha_2}$	0.033	0.017	0.016	0.079
$\sigma^2_{\beta_2}$	0.216	0.138	0.083	0.608
$\sigma^2_{\gamma_2}$	0.046	0.028	0.020	0.126

Note. Posterior medians, standard deviations, and 95% credible intervals of the population mean and variances for the brightness-illusion task (indicated by the subscript 2) as estimated using Stan.

Appendix C

This appendix shows the hierarchical estimates of the intercepts and targets effects in the morph-flanker and the brightness-illusion task as estimated via Gibbs sampling.

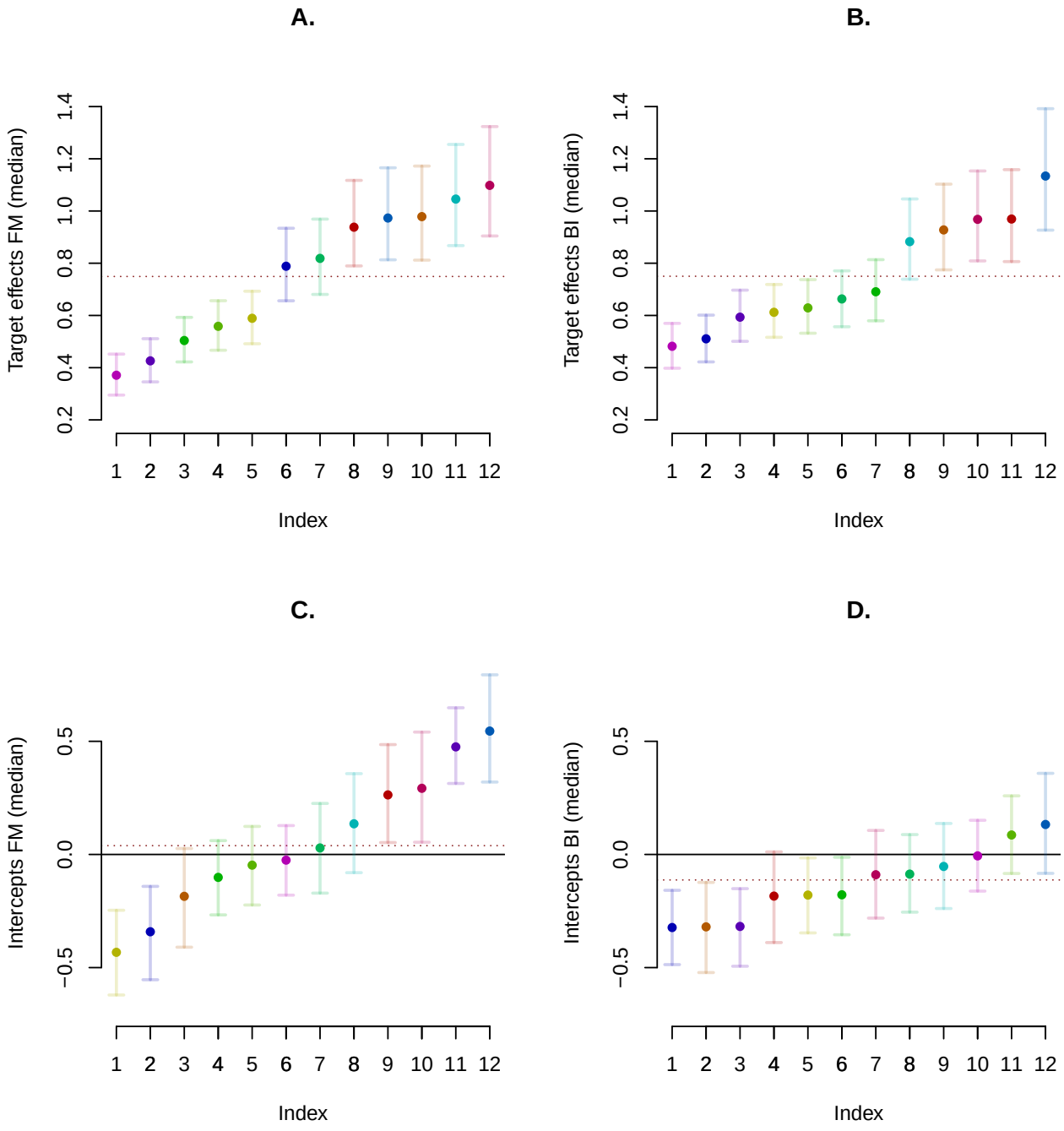

Figure 17. **A.** shows the individual target effects (posterior medians) in increasing order for the morph-flanker task and **B.** shows them for the brightness-illusion task. **C.** and **D.** depict the individual intercepts. Each dot represents a participant. The horizontal lines denote the 95% credible intervals. The dotted lines show the posterior medians of the population mean.

Appendix D

This appendix shows the results of the model fit assessment based on simulated data of 66 participants as well as posterior predictive checks for the hierarchical trial-level model.

Fit assessment on simulated data. We initially simulated data for 70 participants, but excluded 4 based on the 90 percent accuracy in the easiest condition criterion that is described in the main text. To generate the data, the posterior medians from the pilot study results were used. The model fit assessment based on the simulated dataset shows the same pattern as the one from the pilot data (see Figure 18): In the morph-flanker task, the model slightly overestimates the overall contrast effects and in both tasks the extent of individual differences in the intercept.

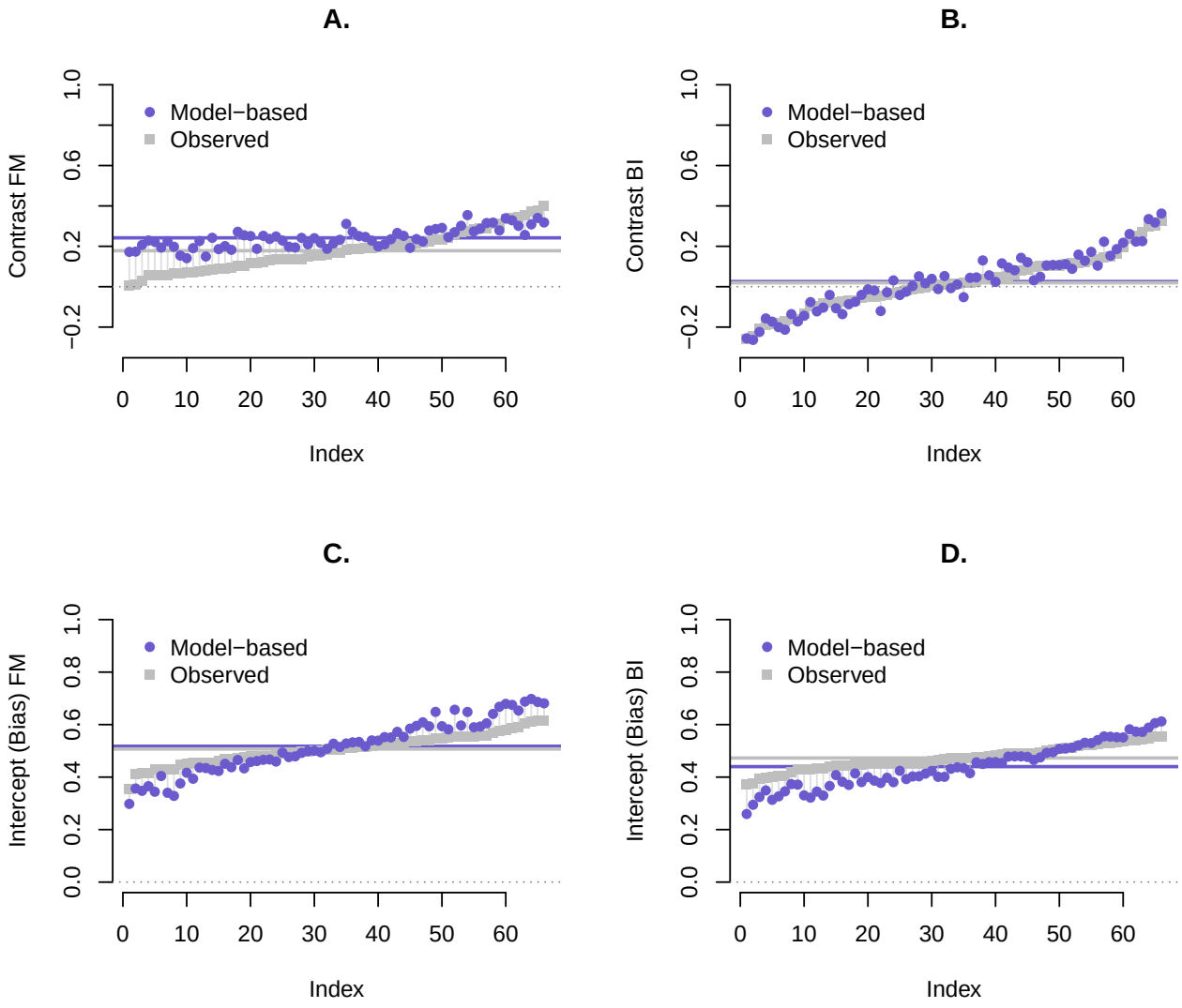

Figure 18. Fit assessment of the hierarchical trial-level model based on simulated data. **A.** and **B.** show the contrast effects on the probability scale for the morph-flanker and the brightness-illusion task, respectively. **C.** and **D.** shows the observed and estimated intercept/bias per participant (see text for details). Each dot and rectangle denotes a participant. The blue dots are model-based estimates and the grey rectangles are observed effects. The blue horizontal line indicates the average estimated effect and the grey line indicates the average observed effect.

Posterior predictive checks. To examine to what extent the model could account for the data, we took 20000 samples from the posterior predictive distribution and computed

both the contrast effects, as well as the proportion of "A"/"bright" responses per distractor condition and target for both the observed data and for each sample. For each computed proportion and across the posterior predictive samples, we then computed the median as well as 80% and 95% credible intervals and thereby partly followed the procedure described by (Singmann, Henrik, 2018).

Figure 19 shows the posterior predictive equivalent to Figure 4 depicting the result of the prior predictive checks. It shows the contrast effects aggregated across participants for each of the eight targets. Again, each vertical line represents a dataset. The model can reproduce the general pattern of the aggregated contrast effects reasonably well. Similarly, Figure 20 shows the aggregated proportion of "A" responses in the morph-flanker task for each condition separately and Figure 21 shows the same for the proportion of "Bright" responses from the brightness-illusion task. For these and the subsequent visual checks, we only used the first 1000 samples that had been drawn from the posterior predictive distribution to lower computation time. Overall, the model could reproduce the data at the aggregated level well: the observed proportions generally lie within the 80% credible intervals. In line with the model assessment reported in the main text, the greatest misfit seems to occur at the fourth target. Furthermore, across both tasks and conditions, the model tends to predict the first four targets slightly too high and the last four targets too low.

Using Figure 22 and Figure 23 we can visually examine the model fit at the individual level. The figures show the observed and predicted proportion of "A" and "bright" responses, respectively, for each participant and each condition. There appears to be some misfit for a couple of middle targets where the observed proportions lie outside of the credible intervals. It seems that the model fails to account for the data when the response pattern across targets deviates from the shape of a sigmoid curve and that this restrictiveness in the model leads to imprecise predictions regarding the data at hand. However, observed proportions that lie outside of the credible intervals are rather rare. We therefore conclude that the fit at the individual level is acceptable.

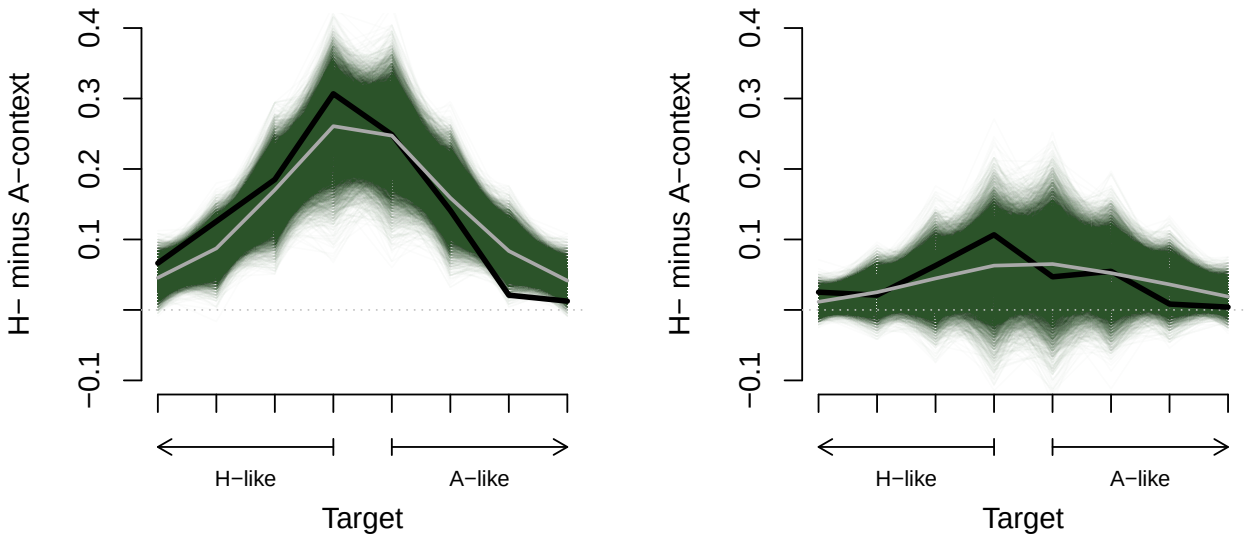

Figure 19. Contrast effects (as shown in Figure 3C and D) from the generated datasets generated according to the posterior predictive distribution for the morph-flanker (**A.**) and the brightness-illusion task (**B.**). Each horizontal line denotes a dataset. The black line denotes the observed effects in the collected data and the lightgrey line the contrast effect for each target averaged across all generated datasets.

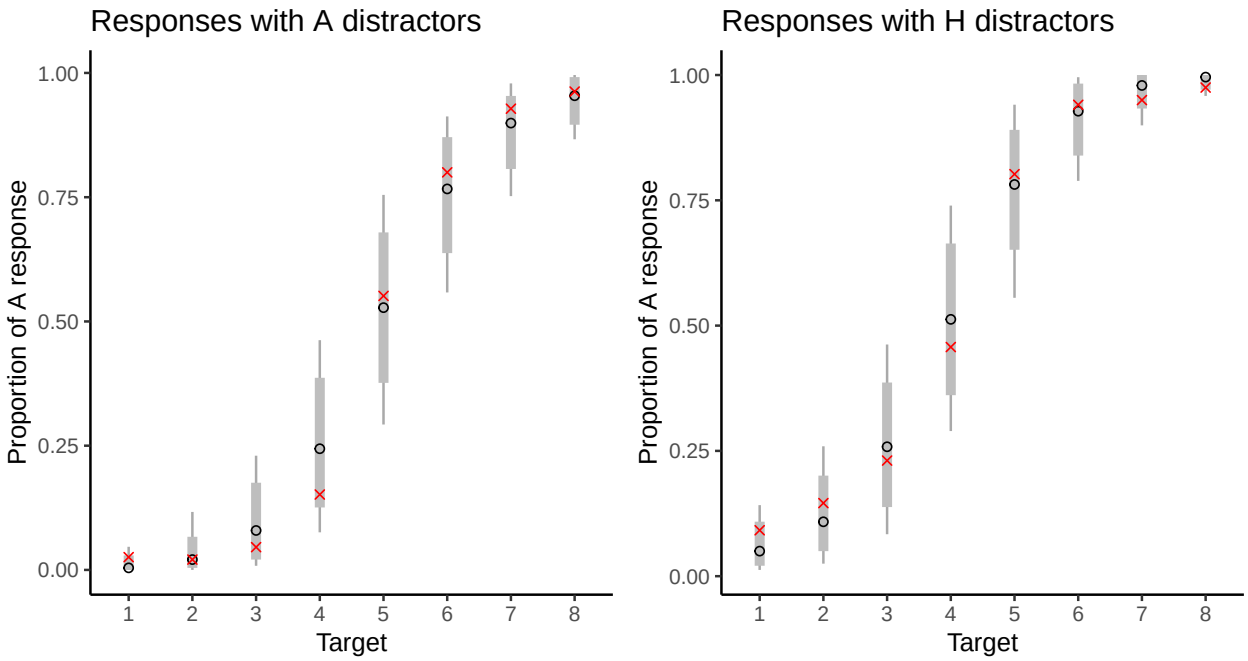

Figure 20. Aggregated proportion of "A" responses in the morph-flanker task for each condition separately based on the posterior predictive samples. The red crosses denote the observed data and the black circle the predictions (the medians). The grey bars denote 80% credible intervals and the vertical lines 95% credible intervals.

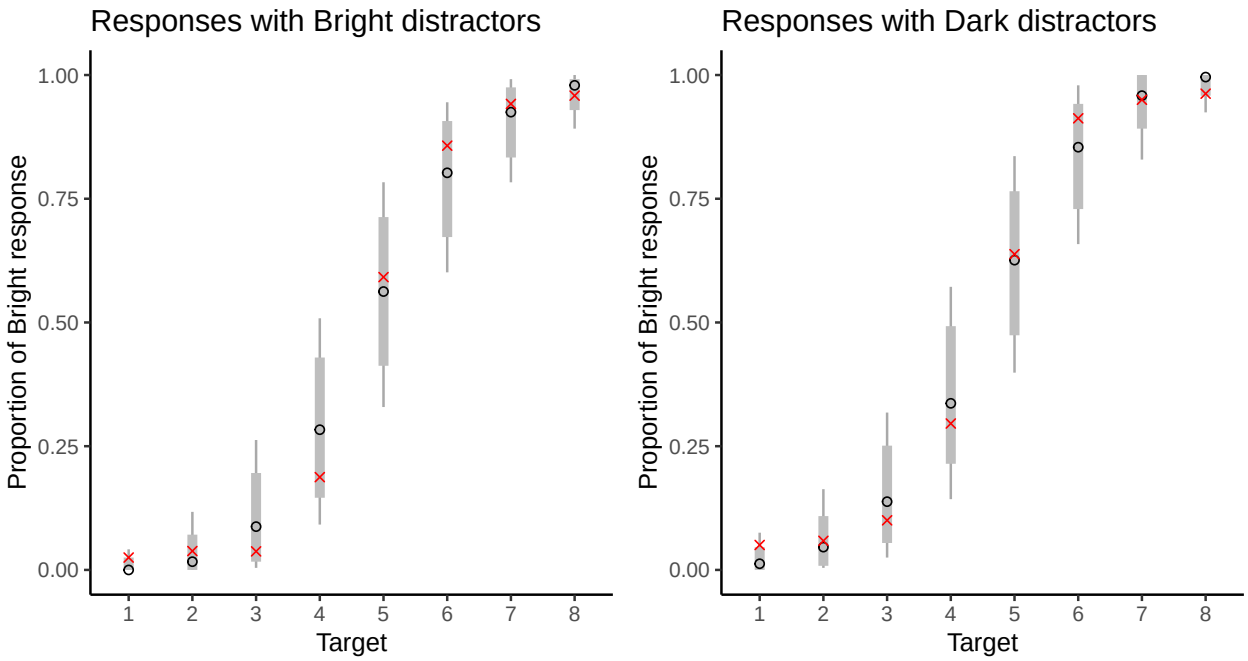

Figure 21. Aggregated proportion of "Bright" responses in the brightness-illusion task for each condition separately based on the posterior predictive samples. The red crosses denote the observed data and the black circle the predictions (the medians). The grey bars denote 80% credible intervals and the vertical lines 95% credible intervals.

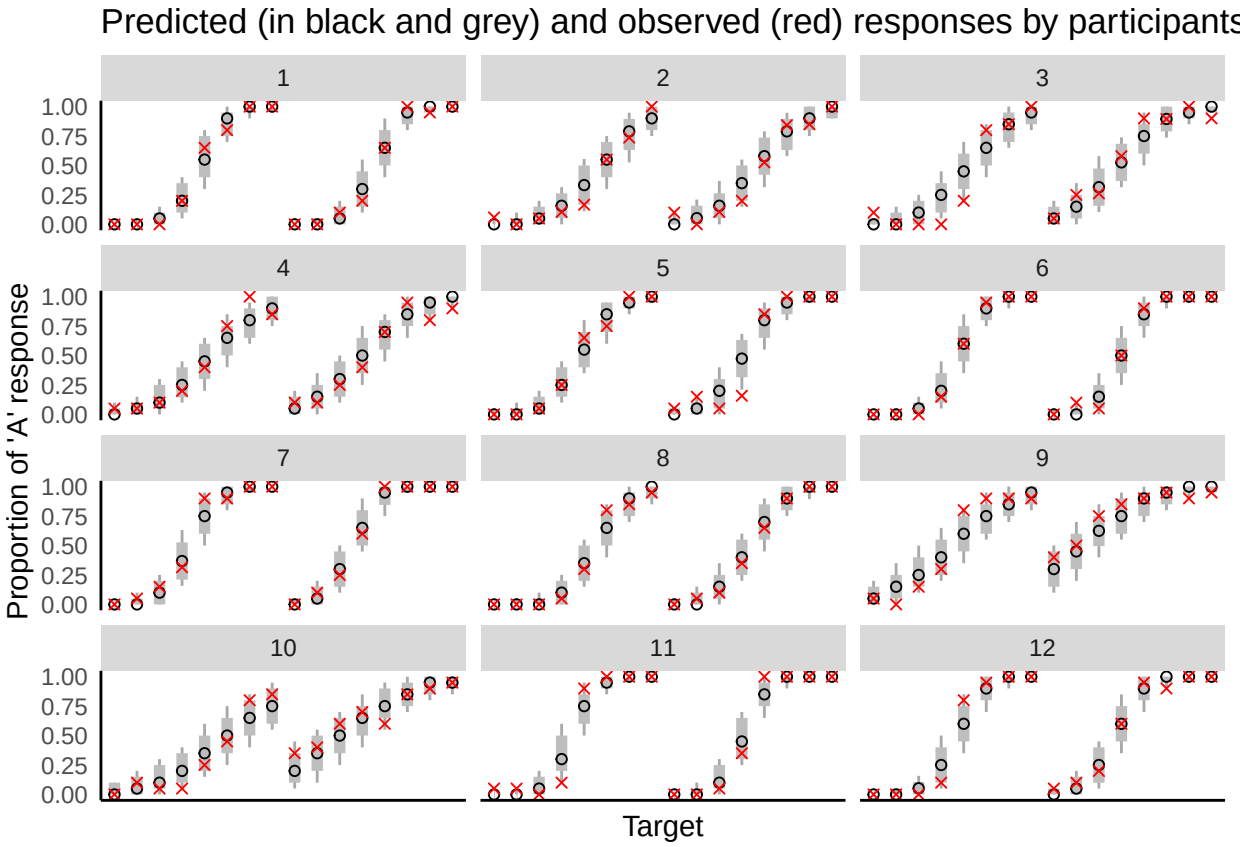

Figure 22. Individual-level proportions of "A" responses in the morph-flanker task for each condition separately based on 1000 posterior predictive samples. The red crosses denote the observed data and the black circle the predictions (the medians). The grey vertical bars denote 80% credible intervals and the vertical lines 95% credible intervals. For each participant, the data on the left shows the observed and predicted responses for each target in the A-distractor condition and the data on the right shows the same for the H-distractor condition.

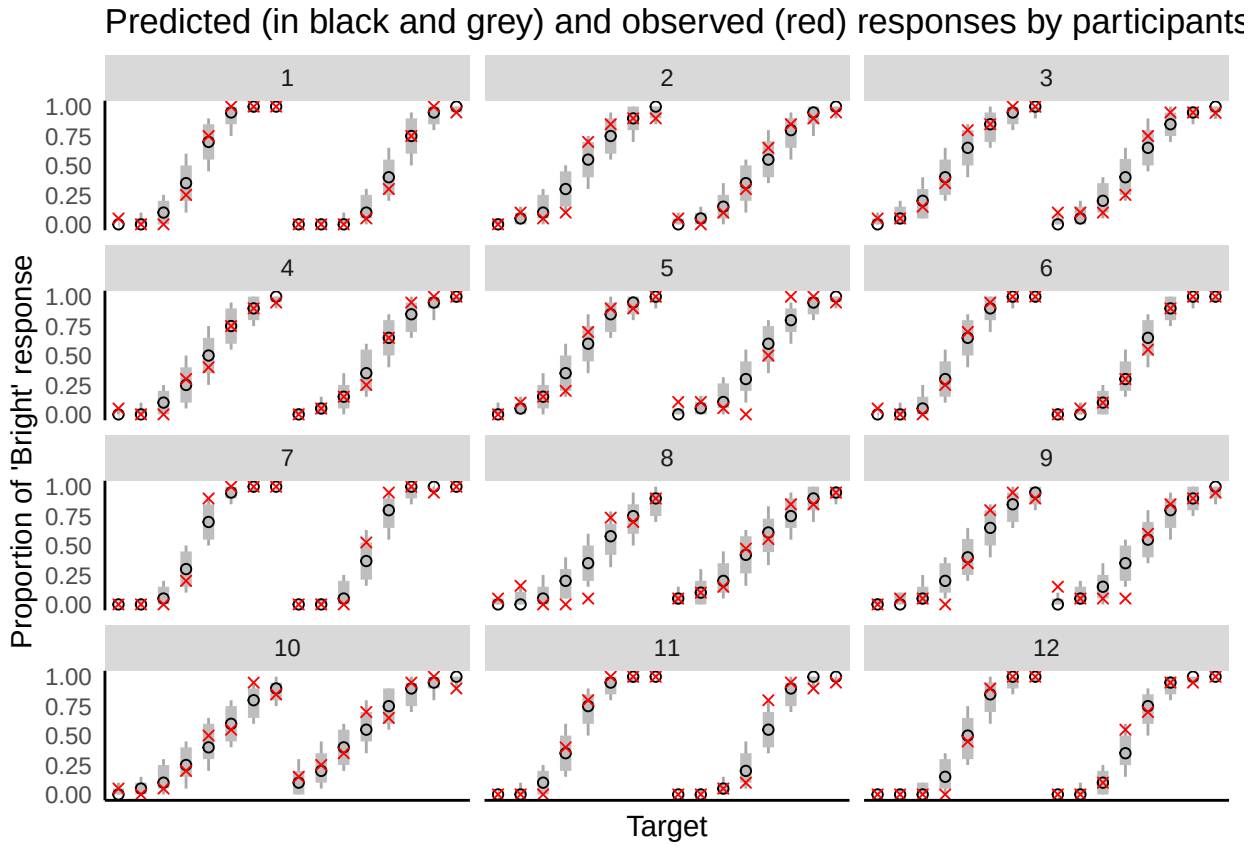

Figure 23. Individual-level proportions of "Bright" responses in the brightness-illusion task for each condition separately based on 1000 posterior predictive samples. The red crosses denote the observed data and the black circle the predictions (the medians). The grey vertical bars denote 80% credible intervals and the vertical lines 95% credible intervals. For each participant, the data on the left shows the observed and predicted responses in the bright-distractor condition for each target and the data on the right shows the same for the dark-distractor condition.