Supplementary Materials A

Below we provide a brief summary for each of the 30 tasks from which we sourced datasets. More detailed descriptions along with demographics can be found in the original papers.

Description of congruency-based tasks

1-2. Thirty-eight participants performed a Stroop and a Simon task (Pratte et al., 2010). In the Stroop task, they had to name the colour of the word. The main condition was the written word, which could be congruent, incongruent, or neutral. In the Simon task, participants were shown a red or green square, which they had to press a response button with their right- or left-hand respectively.

The main condition of interest is the side of the presented stimulus (left or right of a central fixation), which could be congruent (e.g., red square presented on the right side of the screen) or incongruent (e.g., red square presented on the left side) to their response hand. For both tasks, participants completed 7 blocks of 74 trials, which were presented with a fixed ITI (700 ms).

3-4. Hundred-and-seven participants took part in two sessions (three weeks apart) in a laboratory experiment, in which they completed the Eriksen Flanker, Stroop, Go/No-Go, and Stop-signal (Hedge et al., 2019). We analysed the data series from the Eriksen Flanker and Stroop tasks, using only the first session.

In the Eriksen Flanker task, participants had to discriminate the direction of a central arrow (left- or rightwards). The main condition for our analysis is the congruency of the four surrounding flankers to the target (congruent, incongruent, or neutral). In the Stroop task, participants had to name the colour of a word. The main condition was the congruency of the written word to the colour (congruent, incongruent, or neutral). For both tasks, participants completed 5 blocks of 144 trials, which were presented with a fixed ITI (1000 ms).

5. Fifty-four participants performed a digit distance task, in which they had respond whether the presented digit was higher or lower than 5 (Rouder et al., 2005). The main condition of interest is the difficulty (distance from 5), which could be easy (2 or 8), medium (3 or 7), or hard (4 or 6). Participants performed 6 blocks of 60 trials, presented at a fixed ITI of 400 ms, during which they received auditory feedback on their accuracy.

6-13. 130 young (18-28 years) and 159 older (64-76 years) participants took part in a two-day experiment involving 11 different neurocognitive tasks. We used the data from two Stroop (Colour and Number) and two Flanker (Arrow and Letter) tasks (Rey-Mermet et al., 2018). Data from the two age groups were analysed separately – thus resulting in 8 separate tasks for our analyses. As per the original paper, participants with high scores on dementia and depression screening questionnaires (Mini-Mental State, Beck Depression Inventory II, Geriatric Depression Scale) were excluded (17 participants), and another two participants were excluded due to experimenter error in the testing session.

In the Colour Stroop task, participants had to name the colour of the word while ignoring the written word, while in the Number Stroop task, they had to respond how many digits were shown, while ignoring the value of the digits. In the Arrow Flanker task, participants had to discriminate the direction of the arrow, while ignoring flankers. In the Letter Flanker task, they had to respond whether the central letter was a vowel (‘S’ or ‘H’) or a consonant (‘E’ or ‘U’), while ignoring the surrounding congruent (e.g., central vowel with vowels), neutral (% or #), or incongruent (e.g., central vowel with consonants) flankers.

For each task, the main condition is congruency (congruent, neutral, or incongruent). Participants completed 4 blocks of 74 trials for each task, presented at a fixed ITI of 1000 ms, during which they were presented with feedback on response accuracy. There was a response deadline of 1900 ms after stimulus onset. All participants were presented with the same stimulus sequence, which contained no complete stimulus repetitions.

14. Thirty-one participants performed a priming congruency task, in which they had to discriminate the left- or rightwards direction of an arrow (Desender et al., 2016). Prior to stimulus onset, a smaller priming arrow was shown for 34 ms, which fitted into the contour of the stimulus arrow (and hence was invisible to the participant), followed by a black screen for 34 ms. If a response was made within 3000 ms, participants were subsequently asked to rate the confidence in their response. The main condition is the congruency of the prime (congruent or incongruent). Participants completed 8 blocks of 80 trials, presented at a fixed ITI of 800 ms.

15. Eighty-six participants performed a priming congruency task (Desender et al., 2014). On each trial, they were first shown a priming arrow for 23 ms (pointed left- or rightwards), which was followed by two masks and a blank screen of 23 ms each. Next, they were presented with a stimulus arrow (160 ms) and were asked to discriminate the direction of the arrow (left or right). The main condition is the congruency of the prime (congruent or incongruent). Participants completed 8 blocks of 60 trials. The ITI was self-paced, and there was a response deadline of 3000 ms after stimulus onset.

16. Thirty participants performed a priming congruency task (Teuchies et al., 2019), in which they had to discriminate the direction of an arrow (left or right), which was shown for 250 ms. Prior to stimulus onset, they were presented with a prime arrow and black screen, both for 17 ms. The prime arrow fitted into the contour of the stimulus arrow and was hence masked. Afterwards, they were asked to rate the confidence in their response.

The main condition is the congruency of the prime (congruent, incongruent, or neutral). Participants completed 6 blocks of 144 trials. The ITI was variable (ranged 1000–5250 ms), and there was a response deadline of 1500 ms after stimulus onset.

Description of perception-based tasks

1. Fifty-eight participants performed an orientation discrimination task, in which they had to respond whether a grating was oriented left- or rightwards (Rouder et al., 2010). The main condition is the difficulty (tilt of the orientation), which could be easy, medium, or difficult (~4°, 2°, and 1.5° respectively). Participants completed 9 blocks of

80 trials, which were presented with a fixed ITI (1000 ms), during which they received auditory feedback on their accuracy.

2. Eighteen participants performed a multisensory perception task, in which they had to discriminate the left- or rightwards motion of moving dots (Kayser & Kayser, 2018). An auditory cue was played simultaneously, which could either be congruent or incongruent to the visual information. The coherence of the visual motion (i.e., percentage of dots moving in one direction) could take on four different values (i.e., difficulty levels), which were participant-depended and based on a prior calibration and were adjusted every 35 trials based on the performance.

The main conditions are the coherence of the visual motion (total number of percentages is participant-depended) and the audio-visual congruency. All participants performed 5 blocks of 240 trials, except one participant who only performed 4 blocks (960 trials). The stimulus was presented for max 1200 ms (or until response), with a variable trial onset (700-1000 ms) and variable ITI (1500-2000 ms).

3. Twenty-three older participants (aged 62-82) performed the same multisensory perception task as in [2] above (Park, Nannt & Kayser, 2020). Likewise, the main conditions are the coherence of the visual motion (total number of percentages is participant-depended) and the audio-visual congruency. All participants performed 3 blocks of 240 trials.

4-5. Sixteen participants performed two separate auditory discrimination tasks (Kayser et al., 2016). On each trial, they heard two different pure tones presented in background noise. In the frequency task, they were asked which of the tones had the highest frequency. In the intensity task, they were asked which tone was the loudest. Note that the first tone was always fixed (i.e., always the same frequency/loudness), while the second one could vary over seven levels. These levels were participant-dependent and based on a prior calibration session, with the lowest level being 0 (Hz or dB) for everyone.

The main condition is the difference in intensity or frequency. Trials were presented at a variable interval (1700-2200 ms) and with a variable duration between the two tones (2400-2562 ms). For the intensity task, all participants performed 4 blocks of 168 (= 672) trials, except one participant who performed 3 blocks (504 trials). For the frequency task, 4 participants performed 3 blocks (504 trials), 1 participant performed 5 blocks (840 trials), and 11 participants performed 4 blocks (672 trials).

6. Thirty participants performed a detection task, featuring auditory, visual, and audio-visual targets (unpublished pilot data). The target was presented at the left or the right side, and participants had to detect it. The main condition is stimulus type. All participants completed 4 blocks, with block size being 239 trials for 7 participants (= 956 trials), and 215 trials for the others (860 trials).

7. Thirty participants performed a noisy detection task (Maniscalco et al., 2017). On each trial, they were shown two patches of noise for 33 ms, and were asked which patch contained a target. The target's contrast was fixed and participant-depended

(based on a 120-trial calibration). Afterwards, participants rated the confidence in their response. The main condition of interest is the location of the target (left or right). Participants performed 10 blocks of 100 trials, presented at a fixed ITI of 1000 ms.

8. Forty-one participants performed a noisy detection task, with an interleaved memory task (Maniscalco et al., 2017). We analyse the data from the detection task only. On each trial, they were shown two patches of noise for 33 ms, and were asked which patch contained a target. The target's contrast was presented at three different difficulties (participant-depended, based on calibration). Afterwards, participants rated the confidence in their response.

The main conditions are the location (left or right) and the contrast of the target. Participants performed 5 blocks of 102 trials, collected over two days. Trials were presented with a fixed ITI of 1000 ms.

9. Fifteen participants performed an orientation discrimination task (Samaha et al., 2016), in which they had respond whether a grating (shown for 33 ms) was titled left- or rightwards at 45°. The main condition is the difficulty (easy or hard), which related to the noise of the grating, and was participant-depended (based on prior calibration). Participants performed 2 blocks of 160 trials, collected over two days. Trials were presented at a variable ITI (300-500 ms).

10-12. Twenty participants performed three tasks each (Samaha et al., 2017). In the contrast perception task, participants were shown two gratings on each trial, and were

asked which grating had the highest contrast. In the orientation discrimination task, participants were also shown two gratings, and had to respond whether their orientations were same or different (25° more tilt). In the visual short-range memory task, participants were shown one target, followed by a 3 s blank screen, and then the second target. Likewise, they had to respond if the orientations were the same or different.

For all tasks, the main condition is the contrast of the stimuli. Participants performed 3 blocks of 100 trials for each task, which were presented at a variable ITI (300-500 ms). Only for the contrast perception task, one participant only completed 2 blocks, and one participant only 1 block – the latter was excluded from analysis on this task. After each trial, participants had to rate the confidence in their response. The contrast was individually determined with a prior calibration and adaptive throughout each task.

13. Twenty-four participants performed a colour judgment task (Desender et al., 2019). On each trial, they were shown 8 coloured dots presented in a circle around a central fixation. Participants had to judge whether the average colour across all dots was bluer or more red. Next, they were asked to rate the confidence in their response (though half of the participants only were asked for a rating in the even blocks).

The main conditions of interest are the mean and the variance of the colour. Half of the participants performed 8 blocks of 60 (= 420) trials, and half performed 8 blocks of 64 (= 512) trials. Trials were presented with a fixed ITI of 1000 ms, and there was a response deadline of 1500 ms.

14. Twenty-one participants performed an emotion discrimination task, in which they had to respond whether the presented face was angry or happy (Johannknecht & Kayser, 2022). The stimulus face was shown for 100 ms. The main condition is the emotion (angry/happy). Participants performed 2 blocks of 200 trials, with trials being presented with a variable trial onset 400-1000 ms) and ITI (1200-1500 ms).

Model fitting

Each dataset was fitted with a GLM. The model for datasets with one experimental manipulation was fitted as:

$$
\begin{gathered}
RT_n = \\
e_1 x1_n + p_1 x1_{n-1} + a_1 RT_{n-1} + a_2 RT_{n-2} + \\
tv_n + tv_n^2 + bw_n + bw_n^2 + tv_n bw_n + tv_n^2 bw_n + tv_n w_n^2 + tv_n^2 w_n^2 + \\
tv_n e_1 x1_n + tv_n^2 e_1 x1_n + bw_n e_1 x1_n + bw_n^2 e_1 x1_n + \\
tv_n bw_n e_1 x1_n + tv_n^2 bw_n e_1 x1_n + tv_n w_n^2 e_1 x1_n + tv_n^2 w_n^2 e_1 x1_n +
\end{gathered}
$$

and the fitted model for datasets with two experimental conditions is:

$$
\begin{gathered}
RT_n = \\
e_1 x1_n + e_2 x2_n + e_1 x1_n e_2 x2_n + p_1 x1_{n-1} + p_2 x2_{n-1} + p_1 x1_{n-1} p_2 x2_{n-1} + \\
a_1 RT_{n-1} + a_2 RT_{n-2} + \\
tv_n + tv_n^2 + bw_n + bw_n^2 + tv_n bw_n + tv_n^2 bw_n + tv_n w_n^2 + tv_n^2 w_n^2 + \\
tv_n e_1 x1_n + tv_n^2 e_1 x1_n + bw_n e_1 x1_n + bw_n^2 e_1 x1_n + \\
tv_n bw_n e_1 x1_n + tv_n^2 bw_n e_1 x1_n + tv_n w_n^2 e_1 x1_n + tv_n^2 w_n^2 e_1 x1_n + \\
tv_n e_2 x2_n + tv_n^2 e_2 x2_n + bw_n e_2 x2_n + bw_n^2 e_2 x2_n + \\
tv_n bw_n e_1 x1_n e_2 x2_n + tv_n^2 bw_n e_1 x1_n e_2 x2_n + tv_n w_n^2 e_1 x1_n e_2 x2_n + tv_n^2 w_n^2 e_1 x1_n e_2 x2_n + \\
tv_n e_1 x1_n e_2 x2_n + tv_n^2 e_1 x1_n e_2 x2_n + bw_n e_1 x1_n e_2 x2_n + bw_n^2 e_1 x1_n e_2 x2_n + \\
tv_n bw_n e_1 x1_n e_2 x2_n + tv_n^2 bw_n e_1 x1_n e_2 x2_n + tv_n w_n^2 e_1 x1_n e_2 x2_n \\
+ tv_n^2 w_n^2 e_1 x1_n e_1 x1_n e_2 x2_n
\end{gathered}
$$

with e_1 and e_2 reflecting the weights for both of the experimental manipulations *e* multiplied respectively by observation *x1* and *x2* on trial *n,* p_1 and p_2 reflecting the weights of the experimental manipulations from the previous trial *n-1*, a_1 and a_2 reflecting the weights for the relation with the RT from trial *n-1* and *n-2*, and *t* reflecting the weight of trial number multiplied by the current blockwise trial number *v*, and *b* reflecting the weight of block number multiplied by the current block number *w.* These models include main effects for all predictors, as well as each possible interaction between current condition, trial number, and block number.

Supplementary Materials B

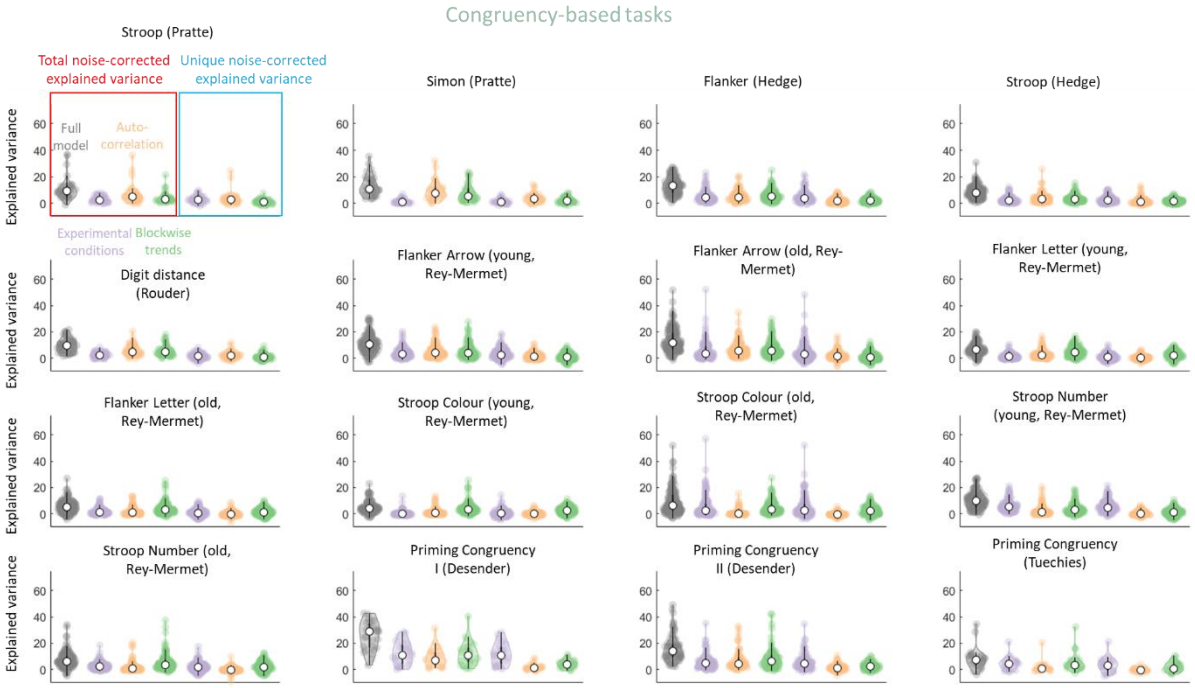

Figure B1. Percentage of noise-corrected total explained variance and unique explained variance in congruency-based tasks. Each panel shows one task, with the task name and first author name on top. In each panel, each dot represents one dataset, the white circle represents the median of the distribution, and the width of the dot cloud indicates distribution density – by the three sources combined (full model or R^2_{abc}; grey) as well as separately (purple, orange, and green for experimental condition R^2_c, autocorrelational R^2_a, and blockwise trends R^2_b sources respectively) as estimated by the dataset level GLMs. The first four violins show the total explained variance (red square) and the last three show the unique explained variance (light blue square). Negative values indicate instances in which the predictors explained less variance than random noise.

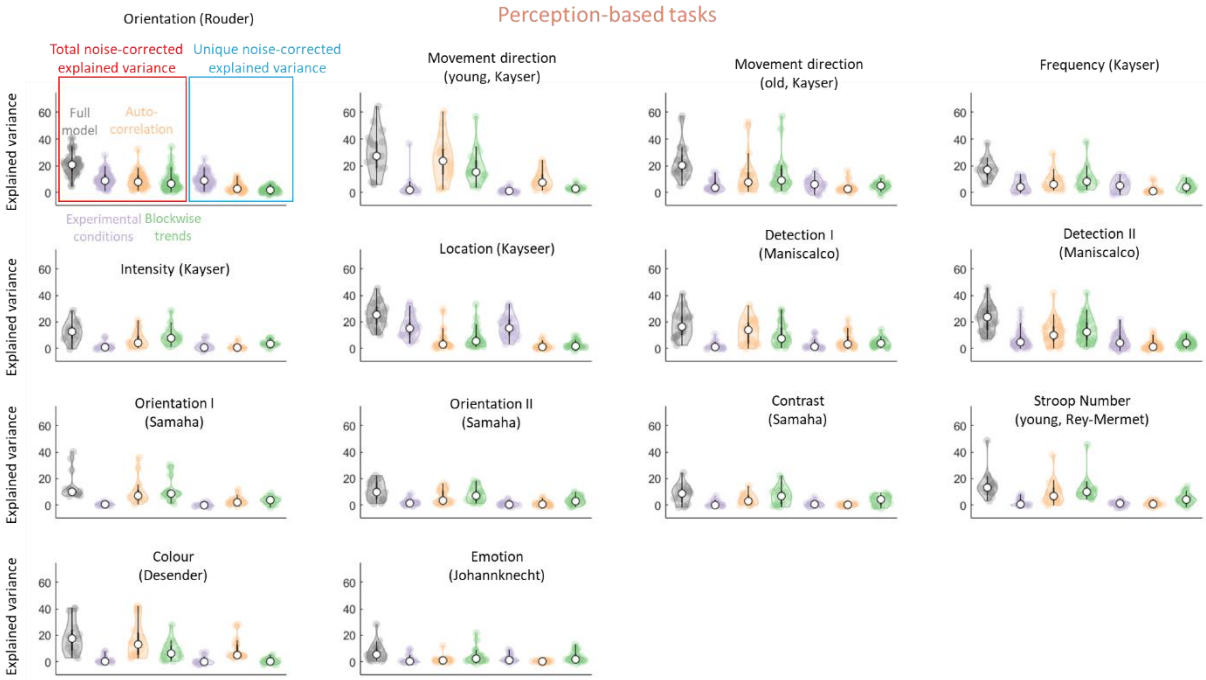

Figure B2. Percentage of noise-corrected total explained variance and unique

explained variance in perception-based tasks. Conventions are the same as in Figure B1.

Supplementary Materials C

Here we report the results of various control analyses that were conducted to check the robustness of our result patterns. We check for the effect of different analysis choices, additional systematic sources of variance not implemented in the current models, impact of multicollinearity, and impact of experimental condition and blockwise trends on the autocorrelation. Congruency- and perception-based tasks were combined for these control analyses, based on the consistent patterns we had observed across both task types.

Because of the number of observations per distribution, inferential tests like paired t-tests can likely detect very small but meaningless differences – and thus do not inform about the robustness of the result patterns. Therefore, we compare the median values of explained variance between our original and control analysis. Any differences < 1% were considered as not noteworthy.

Effect of analysis choices

Here, we ask how our results are affected by preprocessing choices. In our main analyses, the RT series in each dataset were cleaned and transformed prior to analysis. RT outliers were removed with one common criterion for outlier exclusion across all participants: a lower cut-off of 150 ms and an upper cut-off of 3 SD above

the median RT for outlier exclusion. Next, RT series were inverse-transformed, and then the GLMs were run on to predict to RT series and obtain the measures of explained variance (as shown in Figure 5). The first results row of Table C1 shows the explained variance by our original analyses by taking the median across all datasets.

Here we instead use logarithmic (natural log) and square root to transform the RT series prior to analysis. Likewise, GLMs were run on to obtain the measures of explained variance, and the median explained variance was calculated for each source. GLMs were also run on the raw RT series (i.e., no transformation at all). This allows for comparison of the median explained variance across four different transformation choices.

We also checked for the effect of outlier exclusion. The RT exclusion method was taken from the original paper if available (reported in Table 2 and 3). Note that some papers only analysed accuracy and therefore did not include an RT exclusion method. For these papers we kept the < 150 ms and > median RT plus 3 SD(RT) cut-off. Again, analyses were run using all four transformation methods (Table C1; bottom).

For each transformation separately, the total and unique explained variance was extracted for each dataset. We obtained qualitatively very similar results across all analysis choices (see Table B1 for the median explained variance across all participants and tasks). GLMs with raw RT series explained the least amount of variance, but the differences between inverse, log, and square root transformations in total explained variance were all lower than 1%.

Table C1. Median percentage of noise-corrected explained variance across all participants and all tasks for the full model (R^2_{abc}) and the total plus unique explained variance for each of three sources separately under (2 outlier exclusion x 4 data transformations =) 8 different analysis choices. The numerically highest value for each R^2 is shown in bold.

		Full model	Condition		Autocorrelation		Blockwise	
			total	unique	total	unique	total	unique
	inverse	**9.51**	2.50	2.18	**2.97**	**.007**	**4.50**	1.85
<150ms and	raw	8.22	2.64	**2.41**	1.95	.004	3.95	1.91
>3SD+median	log	8.98	2.66	2.35	2.55	.005	4.32	1.89
	square root	8.62	**2.70**	2.40	2.30	.005	4.18	**1.92**
	inverse	**9.72**	2.53	2.14	**3.22**	**.009**	**4.63**	1.97
Original paper	raw	8.28	2.38	**2.33**	2.07	.005	4.04	2.03
	log	9.32	**2.55**	2.32	2.66	.007	4.42	2.01
	square root	8.80	2.46	2.28	2.43	.006	4.24	**2.05**

Additional sources of variance

Our analyses were focused on the experimental conditions of interest. However, the analysed tasks contain additional exogenous sources, such as a variable ITI or a redundant visual feature of the stimulus (e.g., the actual direction of the target arrow

in the Flanker task). On the group level, the effects of arrow direction are often not statistically different from zero, but for each individual participant, the explained variance will numerically be higher than zero. As a control analysis, we therefore ran single-subject GLMs to examine if the amount of explained variance substantially increases when taking additional exogenous sources into account.

GLMs were computed for all datasets that included additional exogenous sources that were uncorrelated with the experimental conditions, if made available at their archival source (11 datasets in total across congruency and perception-based tasks). For example, in the Simon task from Pratte et al. (2010), the colour and location of the stimulus varied across the trials independently from the experimental condition. Both of these factors were added to the new GLM. In this particular example, the additional factors explained .007% of the total variance. It should be noted that the interpretation of these sources can be difficult. For example, in the flanker task, participant respond to right arrows with their right hand and to left arrows with their left hand. This means that any additional explained variance could be due to the visual stimulus property, the motor response bias (i.e., being faster with dominant hand), or a combination of both. We therefore do not make strong claims about what causes this increase in explained variance, but only note that it is relatively small.

Table C2 shows an overview of the additional exogenous factors added to the GLM, as well as the original explained variance by the full model (Figure 3 and 5) and the extended model's explained variance. Overall, increases in explained variance were small, with a median increase of 1.16%.

Table C2. Overview of analysed datasets in GLMs featuring additional exogenous sources. Shown are the 11 included task, with the exogenous sources that have been added to the new GLM (and the exogenous sources that could not be included due to non-availability in the datasets, where applicable), with their original and new amount of noise-corrected explained variance (R^2). Increases in explained variance are minimum for most of the datasets.

	Added predictors (levels)	**Not available in dataset**	**Original R^2**	**New R^2**
Congruency-based tasks				
2. Simon task	Colour of stimulus (2); Location of stimulus (2)		10.71	11.44
5. Digit distance	Distance from 5		9.61	9.56
14. Priming congruency	Direction of stimulus (2)		28.93	29.90
15. Priming congruency	Direction of stimulus (2)	Variable ITI	14.02	15.31
16. Priming congruency	Direction of stimulus (2)	Variable ITI	7.32	9.15
Perception-based tasks				
2. Random dots direction discrimination (young)	ITI (R); Pre-trial fixation time (R);		27.39	28.39

	Direction of visual stimulus (2); Direction of auditory stimulus (2)			
3. Random dots direction discrimination (older)	ITI (R); Pre-trial fixation time (R); Direction of visual stimulus (2); Direction of auditory stimulus (2)		20.26	20.82
10. Grating orientation discrimination	Stimulus orientation (2)	Variable ITI	10.08	12.13
11. Grating orientation discrimination	Stimulus orientation (2)	Variable ITI	9.84	13.33
12. Grating contrast discrimination	Stimulus orientation (2)	Variable ITI	8.81	8.72
13. Short term memory of grating orientation	Stimulus orientation (2)	Variable ITI	13.30	16.48

Multicollinearity

As our models feature a large number of predictors, we tested if these predictors met the assumption of absence of multicollinearity. For each participant, the tolerance value and Variance Inflation Factor were calculated for each predictor. There is reason to assume severe multicollinearity for 8 of the 31 datasets, defined as a tolerance > .10 and a VIF > 10, for the experimental predictors for almost all participants. These were tasks that either featured two experimental conditions and/or a stimulus that changes throughout the task (e.g., stimulus contrast adjusting to current performance),

and were all perception-based tasks. We removed these 8 tasks from the dataset and assessed whether the overall result pattern would be affected; such a change of results would imply that our analysis and conclusions are contaminated by predictor multicollinearity. However, the amount of explained variance for the different variability sources was similar with and without the 8 identified tasks: the total explained variance dropped by less than 1%, and by .001-.39% for each separate source (see Table C3), suggesting that multicollinearity did not impact our conclusions.

Table C3. Median percentage of noise-corrected explained variance across all participants for the full model (R^2_{abc}) and the total plus unique explained variance, after excluding the datasets with indications of serious multicollinearity.

	Full model	Condition		Autocorrelation		Blockwise	
		total	unique	total	unique	total	unique
All datasets	9.51	2.50	2.18	2.97	.007	4.50	1.85
W/o multicollinearity datasets	8.56	2.34	2.06	2.58	.006	4.18	1.67

Correcting the autocorrelation for the other predictors

In our main analyses, we first estimate our three sources separately and combined, and then calculate the communalities and unique explained variance. Alternatively, one may employ a sequential approach. Here, RT was first regressed on the

experimental conditions and blockwise trends. We then obtain the previous RT on trial *n* and *n-1* from the residuals, and run the full model with these residuals, experimental conditions and blockwise trends. Again, the amount of explained variance for was similar, differing only .2% in explained variance (see Table C4).

Table C4. Median percentage of noise-corrected explained variance across all participants for the full model (R^2_{abc}) and the total plus unique explained variance, when first regressing the experimental conditions and blockwise trends from the previous RT series before calculating the models.

	Full model	Condition		Autocorrelation		Blockwise	
		total	unique	total	unique	total	unique
Standard previous RT	9.51	2.50	2.18	2.97	.007	4.50	1.85
Residual previous RT	9.31	2.56	2.23	.007	.006	4.49	3.94

Supplementary Materials D

Linear Mixed Models (LMM) explain variance as parts common to all instances of a grouping variable (i.e., fixed effects) and the variability of these common parts across the grouping variable (i.e., random effects). The more random effects are included in

the model, the more data and computation time are required for model fitting. Moreover, correlations between predictors further increase the required amount of data for fitting. For our purposes, we ideally would want to assign all model terms to both the fixed and the random effects, effectively allowing the model to identify variability of all factors and interactions of our model that is common to all participants, but also to determine the variability between them for all fixed model terms. This model did not run successfully due to being too unstable. We therefore chose the common strategy of entering all model terms in the fixed effects while making the random effects increasingly complex.

With this approach, the fixed factor structure is identical to that of our GLMs. The difference between our GLM approach and an LMM approach is that we fit a separate GLM for each dataset and inspect the resulting amounts of explained variance, whereas the LMM fits all participants at once and assigns variability to the fixed and random model parameters. Thus, the information provided between the GLM and LMM approaches is different, because there is no equivalent of the random factor structure in our GLM approach. Thus, the LMM assigns some part of the variance to differences between participants and might, thus, explain more variance overall than the GLM. However, this additional explained variance does not address the question we posed in our paper, namely how well single trial RT data can be predicted; instead, it quantifies variability between participants.

We explored two different LMM approaches. First, we conducted LMMs on the RT series across all datasets separately for congruency- and perception-based tasks. Next, we conducted LMMs on the RT series across all datasets separately for each task.

All datasets

LMMs were run in R (R Core Team, 2013) with the lmer function in the lme4 package (Bates et al., 2015) for each of the three sources separately as well as combined (full model), mimicking the GLM analyses presented throughout the paper. For each source, three models were run with increasingly complex random terms, including: 1) only the intercept, 2) the intercept plus slope A, and 3) the intercept plus slope A plus slope B. As the more complex models took several hours to days to runs and did not converge reliably, we did not add any more terms. The random effects were allowed to vary over unique participant number (note, rather than by dataset) and task. This strategy takes into account when a participant contributed data to several experiments for estimating model parameters. First, for experimental condition, slope A was the experimental condition of the current trial, and slope B was the condition of the previous trial – giving us the following three random effects:

- *(1 | participant) + (1 | task)*
- *(1 + condition | participant) + (1 + condition | task)*
- *(1 + condition + previous condition | participant) + (1 + condition + previous condition | task)*

In Figure D1, these are indicated by the red triangle, blue plus, and green star respectively. Secondly, for the autocorrelation, we included RT-1 as slope A and RT-2 as slope B. Thirdly, for the blockwise trends model, we included trial number as slope A and block number as slope B. The most complex models (with slope A and B) frequently did not converge and once did not successfully run at all (blockwise trends model) – indicating that even the large amounts of data are

insufficient to fit them. For the full model, we included condition and previous condition as slope A and B, because these are the simplest terms. Because the most complex model again had severe convergence problems, we did not include any of the other model terms.

Estimates of explained variance was extracted for each model using the MuMIn package (Johnson, 2014; Nakawaga et al., 2017; Nakagawa & Schielzeth, 2013). For the congruency-based tasks, datasets from all 16 tasks were run in one model. For the perception-based tasks, this was not possible, because 7 included one experimental condition while the other 7 included two (see Table 2). We therefore ran LMMs separately for one- and two-condition tasks, and then calculated the mean across both. As with the GLM, we performed a noise correction procedure by estimating the explained variance on 100 shuffled RT series and subtracting the mean from the original estimates – though in contrast to the GLM, these values were near zero, as the LMM is more protective against overfitting. These corrected estimates were plotted in Figure D1A, alongside the median explained variance across datasets by the GLMs (as reported in Table 3). The total explained variance across the four models was similar, with all symbols nearly overlapping. In contrast, the experimental conditions explained more variance in the GLM, while both temporal dependencies explained more variance in the LMM.

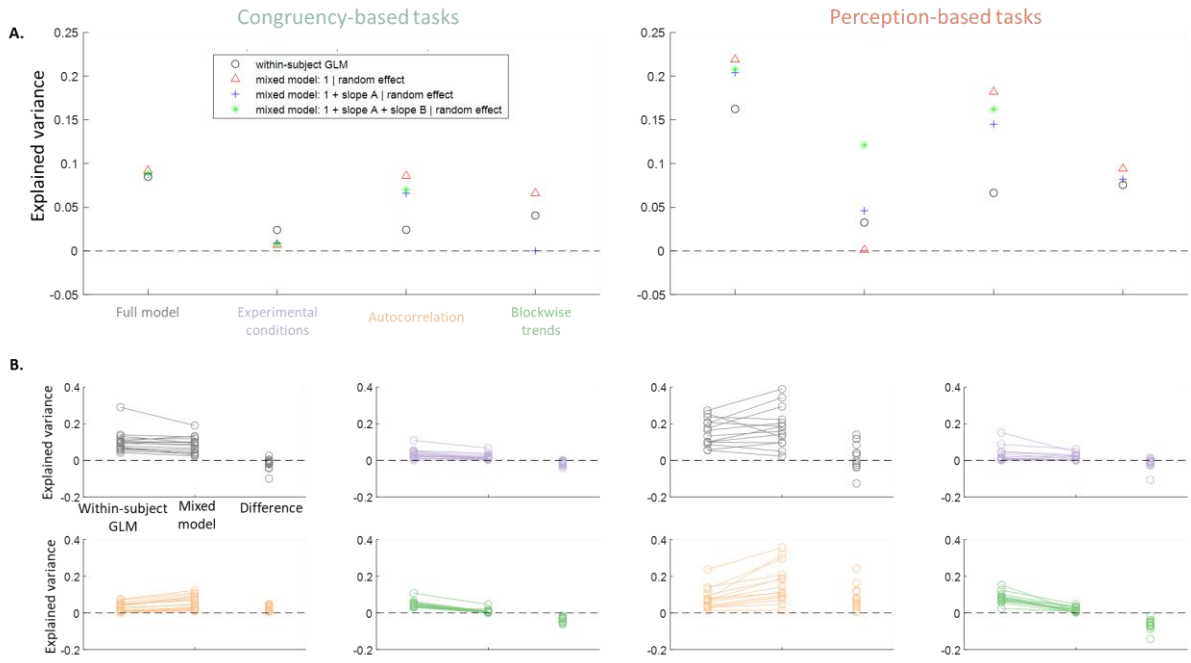

Figure D1. Comparison of the within-subject GLM analyses and across-subject LMM analyses. The percentage of noise-corrected total explained variance is shown for the full model (grey), experimental conditions model (purple), autocorrelation model (orange), and blockwise trends (green) model. ***A.*** *Median explained variance by the GLM (black circle) and explained variance by three LMMs with increasing complexity (red triangle, blue plus, and green star respectively), separately for congruency- (left) and perception-based tasks.* ***B.*** *Median explained variance by the GLM and explained variance by the LMM, with each line representing one task. The right circles indicate the difference between the GLM and LMM, with difference values above zero indicating that the GLM explained more variance.*

Datasets per task

Next, we ran LMMs separately for each task. Again, we chose to run each model with increasing complexity, including intercept, slope A, and slope B. Here, the random

model terms can only vary over participants, rather than over participants and tasks. For example, for the experimental condition model, we fit the random terms:

- *(1 | participant)*
- *(1 + condition | participant)*
- *(1 + condition + previous condition | participant)*

Again, the more complex models did not converge reliably or did not fit successfully. Furthermore, adding more random model terms did not seem to impact the explained variance, with differences between the levels of complexity often < .001. Here, we therefore only include the models with a random intercept, which converged well. Again, all estimates of explained variance were corrected for noise. Figure D1B shows median explained variance per task by the GLM and the explained variance by the LMM alongside their difference, for the full model and each source separately.