

Hierarchical-Model Insights For Planning and Interpreting Individual-Difference Studies  
of Cognitive Abilities

Jeffrey N. Rouder<sup>1</sup> & Mahbod Mehrvarz<sup>1</sup>

<sup>1</sup> University of California, Irvine

Author Note

Version 2, February, 2023. *Author Contributions:* JNR wrote the paper, analyzed the Stroop and flanker effect data, and provided the mathematical derivations. MM analyzed the visual illusion data and overall speed measures. Both authors jointly edited the paper. *Open Science Practices:* All data, analyses, and code for drawing the figures and typesetting the table are available at [github.com/specl/ctx-reliability](https://github.com/specl/ctx-reliability). JNR was supported by NSF 2126976.

Correspondence concerning this article should be addressed to Jeffrey N. Rouder, Department of Cognitive Science, University of California, Irvine, CA, 92697. E-mail: [jrouder@uci.edu](mailto:jrouder@uci.edu)

## Abstract

Although individual-difference studies have been invaluable in several domains of psychology, there has been less success in cognitive domains using experimental tasks. The problem is often called one of reliability—individual differences in cognitive tasks, especially cognitive-control tasks, seem too unreliable (e.g., Enkavi, et al., PNAS, 2019). In this paper, we use the language of hierarchical models to define a novel reliability measure—a signal-to-noise ratio—that reflects the nature of tasks alone without recourse to sample sizes. Signal-to-noise reliability may be used to plan appropriately powered studies as well as understand the cause of low correlations across tasks should they occur. Although signal-to-noise reliability is motivated by hierarchical models, it may be estimated from a simple calculation using straightforward summary statistics.

*Keywords:* individual differences, reliability, cognitive control, cognitive abilities, hierarchical models

## Hierarchical-Model Insights For Planning and Interpreting Individual-Difference Studies of Cognitive Abilities

It is popular to study individual differences in cognitive tasks. By understanding how individuals' performance covaries across these tasks, it is perhaps possible to recover an underlying structure. The classic example in cognitive control comes from Miyake et al. (2000), who used latent-variable models to decompose individual differences in cognitive-control tasks into three factors (inhibition, shifting, updating). In the individual-differences approach, participants complete a battery of tasks such as the Stroop task (Stroop, 1935), the flanker task (Eriksen & Eriksen, 1974), and the antisaccade task (Kane, Bleckley, Conway, & Engle, 2001) among many others. On each of these tasks, a task score is computed per individual. The matrix of scores per individual across the tasks serves as input (see Figure 1) to structural-equation modeling where the covariation across tasks is decomposed into latent variables (Bollen, 1989; Skrondal & Rabe-Hesketh, 2004). The relations among these latent variables purportedly reveal the underlying structure of cognitive processes.

The results using this approach in cognition have been less than stellar, and there is substantial disagreement about the factor structure of cognitive control, attention, and working memory (cf., Rey-Mermet, Gade, and Oberauer (2018), Schubert, Hagemann, and Frischkorn (2017)). Perhaps these disagreements reflect a statistical concern—called here the *reliability crisis*—that cognitive tasks may not be sufficiently reliable to perform latent-variable modeling (Enkavi et al., 2019; Hedge, Powell, & Sumner, 2018). If tasks have low reliability, then correlations are attenuated, and it is difficult to extract the underlying latent structure of covariation. There are two signatures to the reliability crisis: First, in the domain of cognitive control, several tasks that purportedly measure the same construct do not correlate well. For example, the correlation between flanker and Stroop effects in large studies is often near .1 and rarely greater than .25 (Enkavi et al., 2019;

Rey-Mermet et al., 2018; Rouder, Haaf, & Kumar, in preparation). These results indicate that even if these tasks are truly correlated, the correlation may be so attenuated by low reliability as to not be recoverable. Second, latent-variable decomposition in cognitive-control domains seem unreplicable. This lack of replicability is showcased by Karr et al. (2018) who showed that latent-variable analysis with simulated data infrequently recovered the generating model when the sample sizes and parameter values used in simulation came from extant studies.

One proposed solution to the reliability crisis is to use hierarchical models to appropriately partition variability into distinct strata (Haines et al., 2020; Matzke et al., 2017; Rouder & Haaf, 2019). Perhaps by modeling variability and covariability due to trials, conditions, tasks, and people, researchers can improve their recovery of correlations across tasks even in low-reliability environments. Indeed, there is good news on this front—hierarchical models outperform their nonhierarchical competitors and, perhaps more importantly, provide reasonable estimates of uncertainty (Rouder et al., in preparation).

We show here that the *language* of hierarchical models, along with a few quick calculations, can provide a valuable tool in planning and interpreting individual-difference studies. In many ways, our goal is similar to Spearman (1904), who used the language of hierarchical models to develop his famous formula for disattenuating correlations. Our contribution is an update on Spearman’s with new features specific for behavioral experiments.

To see the need for a hierarchical-model language, we need look no further than the concept of *reliability*. Suppose two labs are studying the test-retest reliability of a Stroop task, and everything is the same except that one lab runs 200 trials per person per condition and the other only 20 trials per person per condition. The procedure in the second lab yields a much lower test-retest reliability than the first. Hence, the reliability coefficient is not a property of the task itself. It is not helpful to make statements such as

“The Stroop task has low reliability,” because reliability is critically intertwined with the number of trials per person per condition (henceforth called *trial size*).

Is the reliability crisis merely a crisis of trial size? Is there a measure of reliability that reflects the properties of the task without reference to trial size, and if so, what is it? What is the relationship between trial size, a trial-size invariant measure of reliability and the ability to localize correlations? Hierarchical models provide clear insights into these questions. These insights may be leveraged in planning experiments and interpreting results even if hierarchical models are *not* used in analysis.

### Reliability as Signal-To-Noise Ratios

To answer the above questions, we start with a simple hierarchical model of trial noise in a task. Suppose  $I$  people each run  $L$  trials in congruent and incongruent conditions. Let  $i$  denote people,  $k$  denote conditions, and  $\ell$  denote replicate trials. Response times denoted  $Y_{ik\ell}$  are modeled as  $Y_{ik\ell} = \alpha_i + x_k\theta_i + \epsilon_{ik\ell}$ , where  $x_k$  contrast codes condition and is  $-1/2$  and  $1/2$  for congruent and incongruent conditions respectively. Parameter  $\alpha_i$  is the true overall speed of the  $i$ th participant. Parameter  $\theta_i$  is the true difference between incongruent and congruent conditions—it is the  $i$ th participant’s true Stroop effect and the main target of analysis. The error term is  $\epsilon_{ik\ell} \sim \text{Normal}(0, \sigma^2)$ , with  $\sigma^2$  describing the variability of trial noise. The remaining critical specification is on  $\theta_i$ , the true Stroop effect for the  $i$ th person. Because each  $\theta_i$  describes a latent attribute of a person, it is reasonable to treat it as random effect:  $\theta_i \sim \text{Normal}(\nu, \delta^2)$ , where  $\nu$  and  $\delta^2$  describe the population mean and variance. This last step makes the model hierarchical as variability across trials ( $\sigma^2$ ) and across people ( $\delta^2$ ) are modeled separately.

The usual course is to tabulate a Stroop effect for each person from condition means  $d_i = \bar{Y}_{i2} - \bar{Y}_{i1}$ . It may seem that the sample mean and sample variance of these scores can be used to recover population mean  $\nu$  and variance  $\delta^2$ . The distribution of  $d_i$  is

$d_i \sim \text{Normal}(\nu, \delta^2 + 2\sigma^2/L)$ . From this distribution, it is clear that the sample variance of  $d_i$  does not estimate  $\delta^2$ . Instead, it estimates  $\delta^2 + 2\sigma^2/L$ , a value inflated by trial noise.

What is the effect of this inflation? Consider a test-retest reliability paradigm and let  $j = 1, 2$  denote the day of data collection. Sample effects for each person and day,  $d_{ij}$  are tabulated and correlated. The test-retest model is  $Y_{ijk\ell} = \alpha_i + x_k\theta_i + \epsilon_{ijk\ell}$ . The distribution of Stroop effects for both days has common variability from  $\delta^2$  and unique variability from  $2\sigma^2/L$ :

$$\begin{bmatrix} d_{i1} \\ d_{i2} \end{bmatrix} \sim N_2 \left( \begin{bmatrix} \nu \\ \nu \end{bmatrix}, \begin{bmatrix} \delta^2 + 2\sigma^2/L & \delta^2 \\ \delta^2 & \delta^2 + 2\sigma^2/L \end{bmatrix} \right).$$

The test-retest coefficient  $r$  estimates  $\delta^2/(\delta^2 + 2\sigma^2/L)$ , that is, the expected value or average of  $r$  across many such experiments is  $E(r) \approx \delta^2/(\delta^2 + 2\sigma^2/L)$ . Reliability is a function of between-participant variability, trial noise, and trial size. As trial size increases, reliability to increases. The variability between-participants and within-trials determines the rate of increase.

What parts of reliability are invariant to trial size? Consider the ratio  $\delta^2/\sigma^2$ . This is a signal-to-noise variance ratio—it is how much more variable people are relative to trial noise. Let  $\gamma^2$  denote this ratio. With it, the reliability coefficient follows:<sup>1</sup>

$$E(r) \approx \frac{\gamma^2}{\gamma^2 + 2/L}. \quad (1)$$

The parameter  $\gamma^2$  serves as a trial-size-invariant measure of the reliability of the task. Tasks with high values of  $\gamma^2$  have variability across people that is greater than trial noise, and localizing individuals' effects may be done with just a small trial size. Tasks with low values of  $\gamma^2$  are difficult. It is hard to localize individuals' effects even with many trials, and recovering latent covariation across such tasks remains intractable in experiments with

---

<sup>1</sup> Eq. (1) is for tasks with contrasts such as the Stroop task. For tasks without contrasts, the appropriate equation is  $E(r) \approx \frac{\gamma^2}{\gamma^2 + 1/L}$ .

reasonable trial size. The parameter  $\gamma = \sqrt{\gamma^2}$  is the signal-to-noise standard-deviation ratio. It is often convenient for communication as standard deviations are sometimes more convenient than variances.

Figure 2 is useful for planning. It shows how signal-to-noise standard-deviation ratio  $\gamma$  and trial size affect the reliability coefficients. Large reliability coefficients can be achieved in a few trials for  $\gamma > 1$ ; and somewhat high values can even be achieved in under  $L = 100$  trials for  $\gamma > .25$ . But tasks with lower signal-to-noise ratios may not be feasible as they require hundreds or thousands replicates per person per condition. The horizontal lines in Fig. 2 show reliability at criterial levels of .7 and .9. The problem then is to know what  $\gamma$  to use in planning experiments, which we address subsequently.

### Calculations of Signal-To-Noise Ratios

It is possible to derive straightforward formula for these ratios without performing any model analysis. For a single task, the sample effect,  $d_i$ , is distributed as  $d_i \sim \text{Normal}(\nu, 2\sigma^2/L + \delta^2)$ . Hence, the usual sample variance has an expectation of  $E[\text{Var}(d)] = 2\sigma^2/L + \delta^2$ . Substituting in  $\gamma^2$  yields,  $E[\text{Var}(d)] = \sigma^2(2/L + \gamma^2)$ . Rearranging yields the following estimator of  $\gamma^2$ :<sup>2</sup>

$$\hat{\gamma}^2 = \frac{\text{Var}(d)}{\hat{\sigma}^2} - \frac{2}{L}. \quad (2)$$

$\text{Var}(d)$  is the sample variance,  $\sum_i (d_i - \bar{d})^2 / (I - 1)$ , and  $\hat{\sigma}^2$  is the MSE given by  $\sum_{ijk} (Y_{ijk} - \bar{Y}_{ij})^2 / (IJ(K - 1))$ .

---

<sup>2</sup> For tasks without contrasts, the analogous formula is  $\hat{\gamma}^2 = \frac{\text{Var}(d)}{\hat{\sigma}^2} - \frac{1}{L}$ .

### The Consequences of Low and High Signal-To-Noise Ratios

To show the consequences of low and high signal-to-noise ratios, we analyze data from a cognitive-control battery and a visual-illusions battery. There are two analyses: 1. a conventional analysis in which the analysis starts from person-by-task sample effects, and observed correlations among these are the targets (see Fig. 1); and 2. a hierarchical-model analysis where trial noise is explicitly modeled.

The cognitive-control tasks come from Rey-Mermet et al. (2018), who had young and elderly participants perform a large battery. We highlight data from a number-Stroop task and a letter-flanker task. Figure 3A shows results for the Stroop task. Plotted are observed effects  $d_i$  and model estimates of  $\theta_i$ . Here, the two estimators differ, and the model estimators are far more compact or regularized than the corresponding observed effects. The large degree of regularization means that the apparent individual differences in observed effects are due to trial noise and are not replicable. Fig. 3A shows the model estimate of  $\gamma$  ( $\hat{\gamma} = 0.12$ ), and the number of trials ( $L = 93$ ). Figure 3B shows the same for the flanker task; the signal-to-noise ratio is even lower than that for the Stroop task.

Figure 3C shows the correlation among tasks. The observed correlation and associated 95% CI is shown as a large dot and horizontal line near the top of the distribution. The correlation value is attenuated, and the relatively narrow CI reflects the large number of participants without consideration of trial noise. Comparison to the model estimates show that this high degree of confidence is misplaced. The posterior distribution of correlation from the hierarchical model is plotted along with 95% credible intervals. The uncertainty from low signal-to-noise ratios in the tasks is reflected in the large degree of uncertainty in correlation. Of note, the observed correlation, 0.045 is attenuated by a factor of 3.06 compared to the hierarchical estimate of 0.139. In summary, low signal-to-noise ratios may result in much uncertainty when trial noise is considered and much overconfidence in heavily-attenuated values when trial noise is ignored. This



summary holds for reasonably-sized sample, and the situation is not desirable.

The bottom row of Figure 3 shows a more sanguine case. The data are from a pilot study on visual illusions gathered by the authors and Michael S. Pratte. The paradigm for the illusions is shown in Figure 4. For the Mueller-Lyar paradigm, participants adjusted a center arrow so that it bisects the horizontal line. Participants’ tendency is to set the center arrow too far to the left, and we coded that as a positive bias. For the Poggendorf paradigm, participants adjusted the vertical offset of the right segment so that it lined up with the extension of the left segment through the occluded region. Participants’ tendency is to set this segment too far down, and we coded this as a positive bias. A total of 100 individuals from Prolific ran 15 trials in each illusion; of these 100 individuals, 7 were discarded for producing uninterpretable data.

The resulting biases in perception are shown in Figure 3D-E. As can be seen, illusion tasks yield quite high signal-to-noise ratios. These high-ratio agree well with Cretenoud, Grzeczowski, Kunchulia, and Herzog (2021), who studied individual differences in Mueller-Lyar, Ebbinghaus, and Ponzo illusions. With high signal-to-noise ratios, there is little regularization and sample mean estimates match hierarchical estimates even with the limited number of trials. Moreover, with high signal-to-noise ratios, observed and model correlations match in both value and uncertainty (Fig. 3F). In this case, because trial noise is small relative to individual variation, the uncertainty in correlation reflects the moderate number of people rather than the limited number of trials.

### Signal-To-Noise Ratios In A Few Tasks and Measures

The critical quantity for planning experiments and understanding the ability to localize correlations is the signal-to-noise ratio. What are the values for a range of tasks? Table 1 provides some guidance.

The first row is for weight. We used the U.S. Army’s 2012/2014 survey of 6068

soldiers' anthropometric data (Army, 2014). The mean and standard deviation of weight are 175.4 lbs and 34.4 lbs, respectively. How variable are weight measurements? Let's assume that a repeat measurements might have an standard deviation of 3 lbs. Although 3 lbs is likely too high, it is a small amount relative to the variation across participants, and  $\gamma = 11.48$ . Weight is a best-case scenario—the range of human weights compared to the reliability of scales is indeed quite large.

The next rows are for the Rey-Mermet et al. Stroop task, and Row 2 shows the signal-to-noise ratio for the Stroop effect. The columns **Eq** and **Model** show the estimate of  $\gamma$  from Eq. 2 and the hierarchical model, respectively. These values match well in all cases. Following that are the number of trials needed to attain a criterial level of reliability. For the Stroop effect, the signal-to-noise is low, and the needed trials sizes are large. The following row, Row 3, is for the average or overall speed rather than the contrast. Here, the signal-to-noise ratio is great, that is, the variability in participants in overall speed is large relative to trial noise. Hence, only 10s of trials per person are needed to localize individual differences in overall speed.

There is one new task, lexical distance, which is an implementation of the distance-from-five effect (Moyer & Landauer, 1967). Participants classified digits as either less-than or greater-than five, and did so more quickly if the digit was far from five (digits 2 and 8) than close to five (digits 4 and 6). The contrast row is for the contrast between near and far digits; the speed row is for the overall speed. As with cognitive-control tasks, the signal-to-noise is much greater for localizing individual overall speed effects than for localizing individual distance-from-five effects.

### **The Reliability Crisis Revisited**

The reliability of a task can be profitably assessed without recourse to trial size when reliability is a signal-to-noise ratio. With it, researchers can communicate clearly about

reliability without recourse to trial size. Signal-to-noise ratios succinctly capture how well individual differences may be localized and how well the structure of covariation of individual differences across tasks may be recovered. Moreover, the ratio may be estimated accurately from common sample variances without modeling. The simple formulas provided here may be used for planning an experiment or for interpreting whether small correlations reflect a true lack of correlation or attenuation from low reliability. The advantage of hierarchical-model analysis is the resultant measures of uncertainty on correlations across tasks.

The cause of the reliability crisis is that researchers tend to use too few trials in tasks with too low signal-to-noise ratios. One obvious solution is to use more trials. For example, the correlation between Stroop and flanker can be well localized with  $L = 500$  trials per person per condition. The problem with this obvious solution is that increasing trial size is often unrealistic or inconvenient. Some of the drawbacks to great numbers of trials are that fewer tasks may be run in a battery, effects may attenuate with practice, and people may fatigue or even withdraw.

There are other proposed solutions to the reliability crisis that are not as draconian as implementing excessive trial sizes. These include avoiding difference scores (**Drahiem.etal.2021?**), using so-called gamified tasks (Deveau, Jaeggi, Zordan, Phung, & Seitz, 2015; Kucina et al., 2022), and diffusion modeling (Haines et al., 2020; Lerche et al., 2020; von Krause, Lerche, Schubert, & Voss, 2020; Weigard, Clark, & Sripada, 2021). Understanding the concept of signal-to-noise helps to evaluate these proposals (e.g., Kucina et al., 2022).

Our preference is to avoid tasks with low signal-to-noise ratios because the recovery of covariation is so precarious. High signal instruments include overall speed and biases in visual illusions. We advocate a prioritization of high signal instruments even at the expense of more traditional albeit low-signal instruments (see **Drahiem.etal.2019?** for a related

view). Even without these low signal instruments, there is a rich cognitive world that may be explored fruitfully with individual differences. For example, what is the factor structure of overall speed across a wide range of tasks? Is there a single susceptibility-to-visual-illusions factor? We hope the development facilitates this fruitful exploration.

## References

- Army. (2014). *2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics* (No. ADA611869). Army Natick Soldier Research and Engineering Center. Retrieved from <https://apps.dtic.mil/sti/citations/ADA611869>
- Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.
- Cretenoud, A. F., Grzeczowski, L., Kunchulia, M., & Herzog, M. H. (2021). Individual differences in the perception of visual illusions are stable across eyes, time, and measurement methods. *Journal of Vision*, 21(5), 26–26.
- Deveau, J., Jaeggi, S. M., Zordan, V., Phung, C., & Seitz, A. R. (2015). How to build better memory training games. *Frontiers in Systems Neuroscience*, 8, 243.
- Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, 116(12), 5472–5477.
- 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.
- Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., ... Turner, B. M. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox.
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavioral Research Methods*.
- Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, 130(2), 169–183. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=psych&AN=2001-17501-002&loginpage=Login.asp&site=ehost-live&scope=site>
- Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera,

- M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*, *144*(11), 1147.
- Kucina, T., Wells, L., Lewis, I., de Salas, K., Kohl, A., Palmer, M., ... Heathcote, A. (2022). A solution to the reliability paradox for decision-conflict tasks.
- Lerche, V., von Krause, M., Voss, A., Frischkorn, G. T., Schubert, A.-L., & Hagemann, D. (2020). Diffusion modeling and intelligence: Drift rates show both domain-general and domain-specific relations with intelligence. *Journal of Experimental Psychology: General*, *149*, 2207–2249. doi:10.1037/xge0000774
- 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).
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.
- Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520.
- 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>
- Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, *26*(2), 452–467. Retrieved from <https://doi.org/10.3758/s13423-018-1558-y>
- Rouder, J. N., Haaf, J. M., & Kumar, A. (in preparationin preparation). *Why Most Studies of Individual Differences With Cognitive Tasks Are Bound To Fail*.
- Schubert, A.-L., Hagemann, D., & Frischkorn, G. T. (2017). Is general intelligence little

- more than the speed of higher-order processing? *Journal of Experimental Psychology: General*, 146(10), 1498.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: CRC Press.
- Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *American Journal of Psychology*, 15, 72–101. Retrieved from <https://www.jstor.org/stable/pdf/1412159.pdf>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.
- von Krause, M., Lerche, V., Schubert, A.-L., & Voss, A. (2020). Do Non-Decision Times Mediate the Association between Age and Intelligence across Different Content and Process Domains? *Journal of Intelligence*, 8(3, 3), 33. doi:10.3390/jintelligence8030033
- Weigard, A., Clark, D. A., & Sripada, C. (2021). Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. *Cognition*, 215, 104818. doi:10.1016/j.cognition.2021.104818

Table 1

Tasks	Design	Signal-To-Noise SD Estimates		Needed Trial Size	
	Contrast	Sample	Model	$r = 0.7$	$r = 0.9$
Body Measure					
1. Weight	N	11.48	-	1	1
Stroop					
2. Effect	Y	0.12	0.12	158	607
3. Speed	N	0.96	0.96	3	10
Flanker					
4. Effect	Y	0.07	0.08	353	1359
5. Speed	N	0.70	0.7	5	19
Lexical Distance					
6. Effect	Y	0.11	0.11	193	742
7. Speed	N	0.60	0.62	7	24
Illusions					
8. Mueller-Lyar	N	0.72	0.71	5	19
9. Poggendorf	N	1.46	1.41	2	5

*Note.* The standard deviation of repeated weight measurements was assumed at 3 lbs.



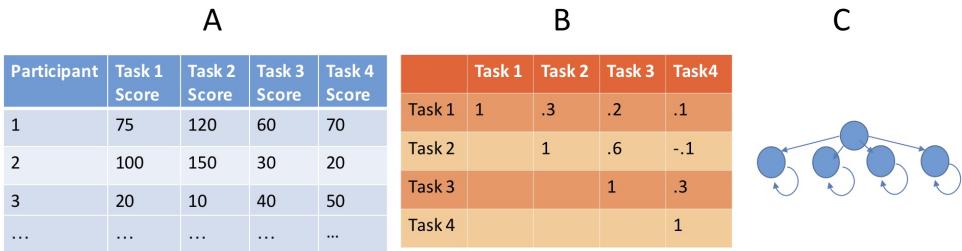


Figure 1. Usual analysis: The raw data are used to tabulated into individual scores (A). The covariation among these individual scores may be computed (B). These covariances are decomposed with structural equation models (C).

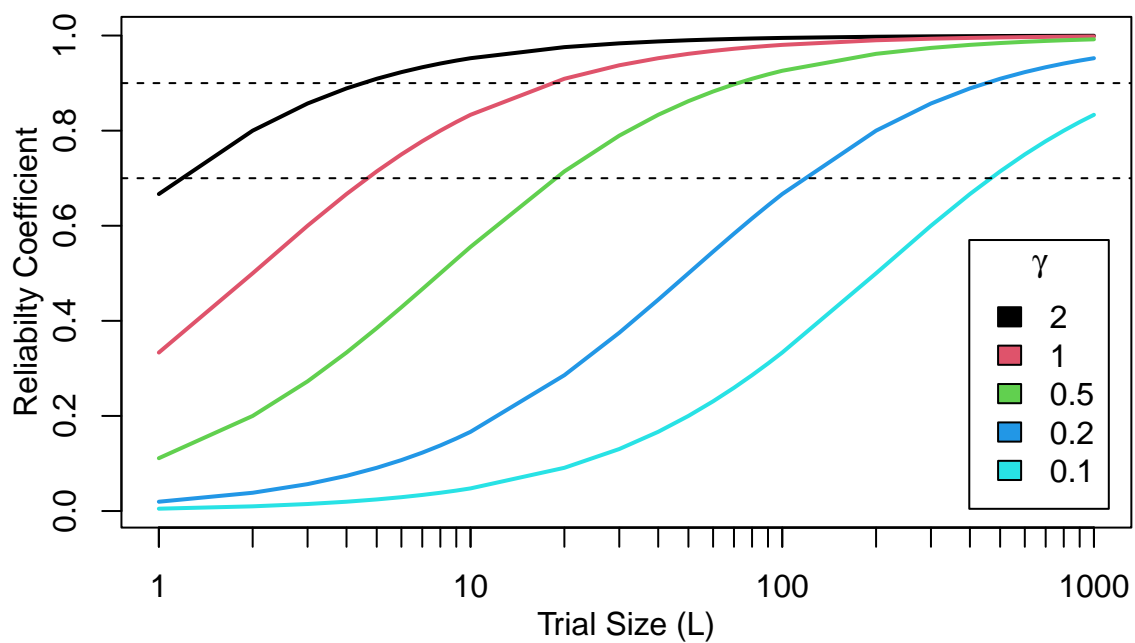


Figure 2. Reliability coefficients as a function of trial size for various signal-to-noise-standard-deviation ratios  $\gamma$ . Horizontal lines at .7 and .9 can be used to plan the trial size for tasks with contrasts across conditions such as the Stroop task.

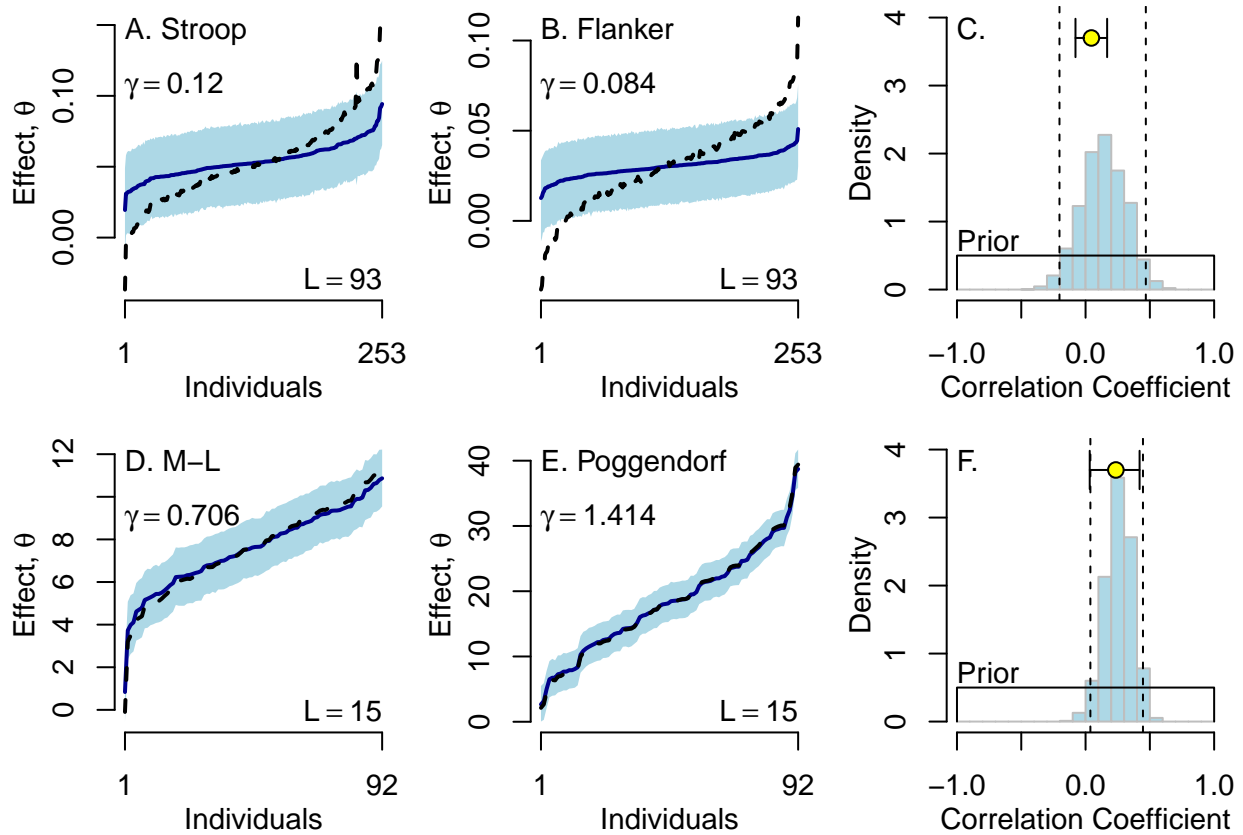
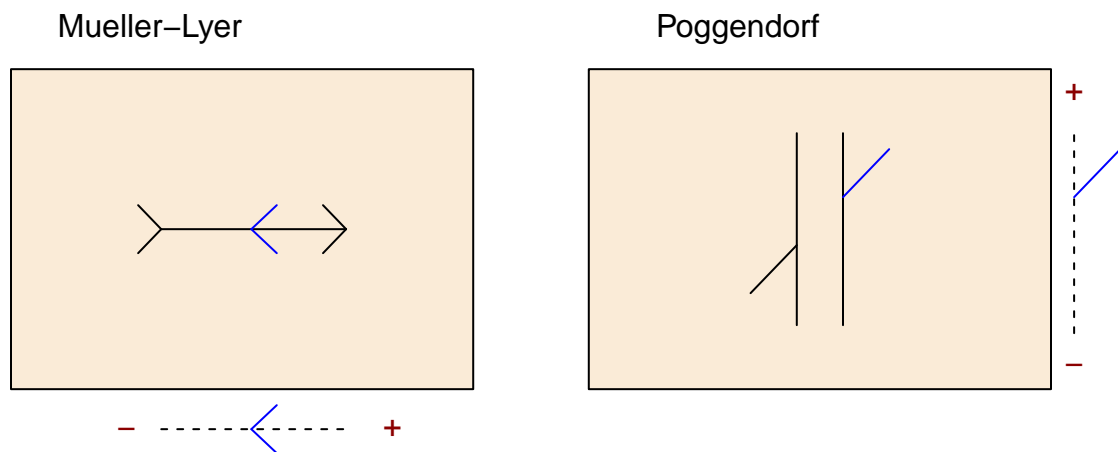


Figure 3. A-B. Observed effects ( $d_i$ , dashed line) and model-based estimates ( $\theta_i$ , solid line) for Rey Mermet et al.'s (2018) Stroop and flanker tasks. The shaded area shows the 95% credible interval for model-based effects. The signal-to-noise ratio  $\gamma$  is low indicating much trial noise. C. Posterior distribution of model-based correlation ( $\rho$ ) between Stroop and flanker effects. The dashed lines denote the 95% credible interval. The point and segments above the distribution show the observed correlation coefficient and associated 95% CI. D-E. Analogous plots for the Mueller-Lyar and Poggendorf illusions F. Analogous plot for the correlation between Mueller-Lyar and Poggendorf effects.



*Figure 4.* Paradigms for assessing visual illusions. Left: For the Mueller-Lyer paradigm, participants adjusted a center arrow so that it bisects the horizontal line. Right: For the Poggendorf illusion, participants adjusted the vertical offset of the right segment so that it lined up with the extension of the left segment through the occluded region.