

# How Much Variance Does Your Model Explain? A Clarifying Note on the Use of Split-Half Reliability for Computing Noise Ceilings

Sander van Bree<sup>\*1,2,3</sup>, Malin Styrnal<sup>1,2,3</sup>, and Martin N. Hebart<sup>†1,2,3</sup>

<sup>1</sup>*Department of Medicine, Justus Liebig University Giessen, Germany*

<sup>2</sup>*Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*

<sup>3</sup>*Center for Mind, Brain and Behavior (CMBB), Universities of Marburg, Giessen and Darmstadt, Germany*

## Abstract

Noise ceilings estimated from a dataset’s split-half reliability offer a powerful way to quantify how much variance a model can in principle explain given the noise in the dataset, allowing researchers to assess model performance relative to an upper bound. In this work, we caution against a common pitfall in this approach to estimating noise ceilings. Specifically, even though the split-half reliability is expressed as a correlation coefficient, it reflects the maximum *explained variance* of a perfect model, *not* the maximum correlation. This subtle misinterpretation leads to artificially lower noise ceilings and, as a consequence, may inflate how close models appear to be to the noise ceiling. A systematic literature analysis suggests that this overly permissive ceiling is the most prevalent interpretation of noise ceilings estimated through split-half reliability. The purpose of this work is to explain when the mistake happens, why it happens, what its consequences are, and how to avoid it. Toward this end, we offer a general explanation showing how split-half reliabilities relate to the performance of a maximally predictive model, supplemented by simulations, and mathematical derivations. Overall, this clarifying piece is meant to help researchers better understand the statistical underpinnings of noise ceilings and support more consistent reporting across studies.

## Introduction

Empirical studies in psychology, cognitive neuroscience, and computational neuroscience commonly require robust statistical methodologies to interpret experimental data. One prevalent approach involves statistical modeling of observed data patterns, such as modeling neural or behavioral responses from humans or animals. These statistical models typically use hypothesized variables, or ones that originate from a computational model. For example, a psychologist might be interested in modeling response time patterns in a visual attention paradigm (Carasco, 2011), a cognitive neuroscientist might test an encoding model to predict fMRI responses (Kay et al., 2008; Dumoulin and Wandell, 2008; Naselaris et al., 2011), or a cognitive computational neuroscientist may predict behavioral or brain responses using deep neural networks (Yamins and DiCarlo, 2016; Conwell et al., 2024).

---

\*Corresponding author: [sandervanbree@gmail.com](mailto:sandervanbree@gmail.com)

†Corresponding author: [martin.hebart@uni-giessen.de](mailto:martin.hebart@uni-giessen.de)

A fundamental challenge in evaluating the performance of such models lies in the noise inherent in empirical data, which can significantly influence the assessment of model performance. If a model only captures 20% of the variance, this could indicate poor model performance. However, if there is a high degree of noise in the data, 20% may be as good as it gets. To quantify how well a given model performs relative to the noise level of a dataset, it has become increasingly popular to estimate a so-called "noise ceiling", which represents the theoretical maximum performance any model can achieve given the noise and variability inherent to the data (Lage-Castellanos et al., 2019; Nili et al., 2014).

Among the various methods developed for computing noise ceilings, a widely adopted approach to determine this ceiling is based on the split-half reliability of a repeated measurement (Lage-Castellanos et al., 2019; Hsu et al., 2004). This ceiling is particularly common for evaluating individual participant data (Huth et al., 2012; Cadieu et al., 2014), but has also been used when the average of a group of participants is compared to a model (Konkle and Alvarez, 2022). The computation of noise ceilings based on split-half reliability generally involves two steps. First, in the case of individual-participant data, the repeated measurements (e.g. stimulus repetitions) are divided into two halves, and the responses in one half are correlated with those in the other half. Since the split-half reliability relies on two subsets of the data and thus underestimates the reliability of the full dataset, a second step is usually carried out, which involves applying the Spearman-Brown correction (Spearman, 1910; Brown, 1910) to estimate the reliability of the full dataset:

$$r_{YY} = \frac{2 \cdot r_{sh}}{1 + r_{sh}}$$

The procedure is the same when comparing a model to the average response of a participant group, except that the group of participants is split in half instead of the repeated measurements of one individual.

This approach is well-motivated and straightforward, and it continues to be the prevailing method for noise ceiling estimation. However, as we will discuss in this work, this approach is susceptible to a subtle but important misinterpretation. Specifically, it is often assumed that this reliability ( $r_{SB}$ ), which is expressed as a correlation coefficient, represents the upper bound for the highest correlation coefficient ( $r$ ) achievable by a model. This interpretation is intuitively appealing: if the noise ceiling is defined using correlation coefficients ( $r_{SB}$ ), it seems natural to treat it as the upper bound on a model's correlation coefficient ( $r$ ). Yet, contrary to this intuition, the split-half correlation reflects the upper bound of the explainable *variance* ( $R^2$ ) of the dataset. Consequently, the true upper bound for a model's correlation coefficient is the square root of this reliability:

$$r_{max} = \sqrt{r_{SB}}$$

In simpler terms, reliabilities derived from split-half procedures tell us how much variance in the data can in principle be explained, not how strongly two sets of measurements can correlate (Table 1; Morgan and Schwarzkopf, 2019; Allen et al., 2022).

<b>Performance metric</b>	<b>Noise ceiling</b>
correlation coefficient ( $r$ )	square root of the SB reliability ( $\sqrt{r_{SB}}$ )
explained variance ( $R^2$ )	SB reliability ( $r_{SB}$ )

Table 1: Mapping between performance metrics and noise ceiling metrics.

This subtle misunderstanding can cause models to appear substantially better than they are. For example, expressing model performance as the fraction of variance explained relative to the squared reliability can make models seem to account for more variance than they actually

do. This matters especially when evaluating claims that a model captures substantial portions of explainable variance, a point we return to in the discussion section.

In this work, we (i) show that noise ceilings obtained from split-half reliability represent the maximum explainable variance rather than the maximal achievable model correlation, (ii) explain why this matters, and (iii) approximate the prevalence of such systematic overestimation of model performance relative to noise ceilings. To achieve these ends, we convey the general intuition of why the central claim holds true and perform a literature analysis to estimate how widespread this interpretation of noise ceilings is. In the supplementary materials, we offer in-depth technical information for readers interested in additional details, including mathematical proofs and a simulation.

## The basic intuition

Why does the split-half reliability underestimate the maximum correlation a model can achieve? One way to approach this is through the classic correction for attenuation formula (Spearman, 1904). This formula dictates that the maximum correlation ( $r_{max}$ ) between two variables  $A$  and  $B$  is limited by the square roots of their respective reliabilities ( $r_{AA}$  and  $r_{BB}$ ):

$$r_{max} = \sqrt{r_{AA}} \cdot \sqrt{r_{BB}}$$

This formula is useful because it allows us to analyze how the noise ceiling misapplication might happen. First, the data’s reliability is estimated via a split-half procedure, and usually corrected using the Spearman-Brown formula ( $r_{SB}$ ). The issue arises when this value is treated as the ceiling for correlation. To see why this is so, consider what it is that we are trying to derive: the maximum correlation any model can have with the data:

$$r_{max} = \sqrt{r_{SB}} \cdot \sqrt{r_{model}}$$

When using  $r_{SB}$  directly, this erroneously assumes that the model suffers from the same noise as the data, placing the noise factor on both sides of the formula:

$$r_{max \text{ incorrect}} = \sqrt{r_{SB}} \cdot \sqrt{r_{SB}} = r_{SB}$$

In contrast, the best possible model corresponds to the true data-generating process. Since this process is deterministic (i.e., noise-free), its reliability is perfect ( $r_{model} = 1$ ). When we substitute this into the attenuation formula, the maximum achievable correlation becomes:

$$r_{max} = \sqrt{r_{SB}} \cdot \sqrt{1} = \sqrt{r_{SB}}$$

This demonstrates that the theoretical limit for the correlation coefficient is  $\sqrt{r_{SB}}$ —not  $r_{SB}$ . As a result, the split-half reliability  $r_{SB}$  corresponds to the maximum *explained variance* ( $R_{max}^2$ ), as summarized in Table 1. Intuitively, this is because measurement noise weakens the signal by a fixed factor. Correlating two noisy halves applies this penalty twice (once for each half), whereas correlating a noise-free model with the data applies it only once.

We can offer a visual intuition for this distinction by considering how noise incrementally affects a model’s correlation with empirical data. Specifically, consider a continuum ranging from a noise-free True Model to versions of that model corrupted by increasing amounts of noise (Figure 1). The correlation between the True Model and the data corresponds to the maximum achievable correlation (green dashed line). Next, we add noise to the True Model up to a point that matches the measurement noise intrinsic to the data. As a consequence, the correlation to the data will drop. This scenario of having a model attenuated by noise to the same degree as the data mirrors the case of two data splits, each of which is also attenuated by independent (measurement) noise. Crucially, this noise-infused correlation results in an inappropriate ceiling

because it underestimates the maximum attainable correlation. Importantly, as shown above, the relationship between the two is a square operation: the invalid ceiling is the square of the valid ceiling. Since correlations range between 0 and 1, this squared value is systematically lower than the true ceiling, effectively setting the bar for model performance too low.

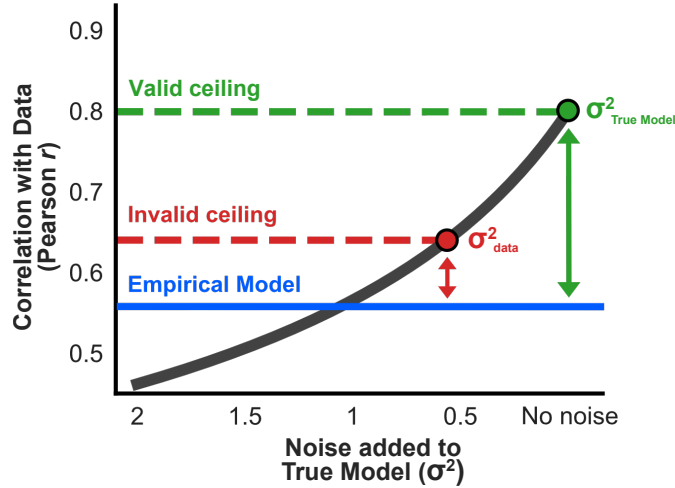


Figure 1: **Conceptual illustration.** An analogy showing why split-half reliabilities underestimate a true model’s achievable correlation. The green dashed line represents the true ceiling any model can achieve. The red line represents a scenario in which the true model is infused with noise equal to that of the data, which more closely matches the scenario of the correlation between two data splits. This ceiling is invalid because it underestimates the correlation a true model would obtain. The blue line shows the performance of a hypothetical empirical model which a researcher might obtain. Note: apparent crossings with the ceiling at high noise are an artifact of this didactic illustration.

To illustrate the practical consequences, consider a concrete scenario aligned with Figure 1. Assume a researcher wants to know how well their neural network captures brain responses in a region of interest. They correlate the model predictions with the observed responses and find  $r = 0.56$ . Next, they estimate a noise ceiling using the Spearman-Brown corrected split-half reliability and obtain  $r_{SB} = 0.64$ . Treating this value as the maximum achievable correlation, the model appears to fall only 0.08 short of the ceiling—indicating near-ceiling performance. However, the correct noise ceiling is  $\sqrt{0.64} = 0.80$ , revealing a gap of 0.24—three times larger than it first appeared. The distortion is perhaps even clearer in explained variance terms: a model with  $R^2 = 0.31$  actually captures 48% of the explainable variance ( $0.31/0.64$ ), yet would appear to capture 76% under the misapplied ceiling of  $r_{SB}^2 = 0.41$  ( $0.31/0.41$ ). Thus, the model initially looked like it reached near-optimal performance, but under the valid ceiling, it becomes clear that there is substantial room for model improvement.

## Prevalence of noise ceiling issue

The misapplication described above would be inconsequential if it rarely occurred in practice. Hence, we asked how common this misapplication is for studies that use noise ceilings. To estimate this, we conducted a literature review (Figure 2). The search was conducted on Google Scholar on October 2nd, 2025, with the following search terms: (“spearman-brown” OR “split-half reliability”) AND (“noise ceiling” OR “explainable variance”), filtering for studies

from 2015 onward. These search criteria were not intended to exhaustively index all papers that compute noise ceilings and likely strongly underestimated the total volume of relevant studies. However, this search served to sample the literature and reveal a representative and recent subset that would yield a sample containing more relevant than irrelevant work, whereas a more exhaustive search would have resulted in a larger fraction of false positives. A total of 244 articles were retrieved. From these, we excluded PhD theses, duplicates, and papers that by their title were clearly irrelevant to the fields of psychology, cognitive neuroscience, or computational neuroscience. This resulted in 164 candidate papers, comprising published work, preprints, and conference proceedings. Because the exact calculation of the noise ceiling and its comparison to model performance was not always apparent, we contacted individual authors from all papers to clarify (i) if a paper computed a noise ceiling based on the regular (uncorrected) split-half reliability or the Spearman-Brown corrected split-half reliability, and if so, (ii) how model performance was compared with the noise ceiling. Some studies used the uncorrected split-half reliability and evaluated models on only half of the dataset. This distinction does not matter for our main assessment, and therefore, we use the term “reliability” to refer to both the Spearman-Brown-corrected and the regular split-half reliability.

In total, we received 55 responses from authors (response rate: 34% out of 164 candidate papers). We supplemented this by manually inspecting 45 randomly chosen additional papers, amounting to a total number of 100 papers on which we report summary statistics. For nine out of the 45 manually inspected papers, we subsequently received responses by authors, the answers of which matched the conclusions from our manual inspection, indicating that the manual inspection was likely done correctly.

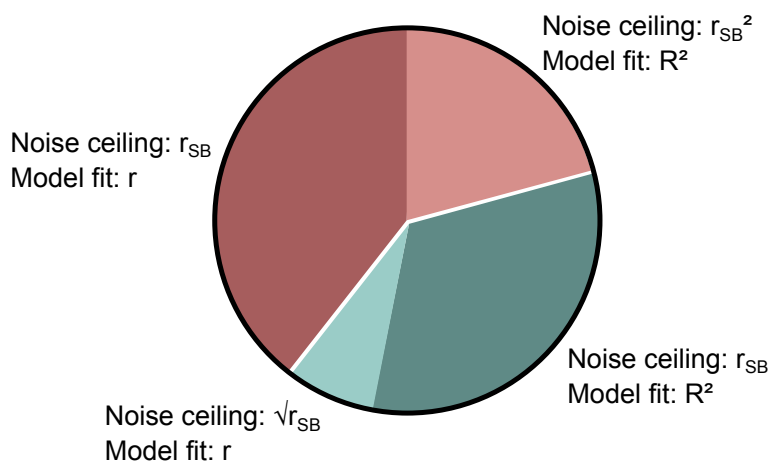


Figure 2: **Prevalence of noise ceiling issue in the literature.** Among relevant papers for which the mapping was determinable, 60% did not show a valid pairing between noise ceiling metric and model performance metric (red). 40% of papers did adopt an adequate mapping in line with Table 1 (green).

Of the full set, 58 papers were relevant to our question because they used a reliability coefficient or Spearman-Brown corrected split-half reliability as a noise ceiling. For five papers we manually checked, we could not be sure if the mistake was made or not. Among the 53 remaining papers, 60 % ( $n = 32$ ) did not correctly match the noise ceiling metric and model performance; of the 32 papers using the incorrect approach, 66 % ( $n = 21$ ) compared a corre-

lation coefficient ( $r$ ) to the reliability ( $r_{SB}$ ), while the remaining 34 % ( $n = 11$ ) compared an explained variance ( $R^2$ ) to the squared reliability ( $r_{SB}^2$ ). In contrast, of the 40 % of papers ( $n = 21$ ) that correctly compared metrics, 19% ( $n = 4$ ) compared a correlation coefficient ( $r$ ) with the square root of the reliability ( $\sqrt{r_{SB}}$ ), and 81% ( $n = 17$ ) compared an explained variance ( $R^2$ ) to the reliability ( $r_{SB}$ ).

In conclusion, this suggests the issue highlighted in this work is relatively widespread, with the majority of papers in our sample using an overly liberal noise ceiling estimate, highlighting the importance of careful reporting and interpretation of noise ceiling estimates.

## Discussion

Split-half reliabilities are widely used to derive an upper bound on how well a model can in principle account for observed data – be it neural, behavioral, or otherwise. In this work, we caution against a pitfall in using reliability-based noise ceilings toward this end: perhaps counterintuitively, the (Spearman-Brown corrected) split-half reliability provides a theoretical maximum of the *explained variance* a model can reach—not a maximum correlation. The practical implication is subtle but crucial: to avoid a systematic overestimation of model performance relative to the noise ceiling, model–data *correlations* should be compared against  $\sqrt{r_{SB}}$ , whereas model *explained variance* should be compared against  $r_{SB}$ . Importantly, this concerns split-half reliability estimates regardless of whether they are adjusted by the Spearman-Brown correction or not. While the correction extends reliability estimates from split halves to the full data, reliability itself, whether it is corrected or not, bounds the explained variance of models predicting the data.

Based on an analysis of previous work, we find that well over half of the sampled literature uses an overly liberal mapping that overestimates model performance relative to the noise ceiling—namely, model correlation coefficient ( $r$ ) to split-half reliability ( $r_{SB}$ ), or model explained variance ( $R^2$ ) to the squared split-half reliability ( $r_{SB}^2$ ). This literature overview was intended to offer a straightforward attempt at estimating the prevalence of this issue, but there are potential challenges to the validity of this numerical estimate. For one, some authors may have reported making the mistake even though they did not. However, our manual assessments were consistent with the authors’ responses, making this interpretation unlikely. Alternatively, response bias may have led authors who computed the noise ceiling correctly to preferentially respond to our email. At the same time, we find a similar proportion of incorrect noise ceilings in the set of papers we manually reviewed compared to the papers that authors responded to, indicating that response bias may have played a smaller role.

In some cases, we could not determine whether the mistake was made, highlighting the importance of clear reporting on how noise ceilings are computed and compared against model performance. We recommend that authors state explicitly what quantity is used as the noise ceiling and what metric it is compared to, for example by writing: “the reliability ( $r_{SB}$ ) served as the noise ceiling for the model’s explained variance ( $R^2$ )”. Such explicit reporting makes it easier for readers and reviewers to detect mismatches and to verify that model performance is evaluated against an appropriate upper bound.

What are the consequences of the statistical misapplication outlined in this paper? In this work, we would like to distinguish local and global effects. Locally, individual studies may draw overly strong conclusions about their hypothesis of interest. For example, a researcher might reasonably conclude on the basis that a model is close to ceiling that the theoretical framework underpinning that model is likely right, drawing resources and concentration away from alternative hypotheses that may have seen more focus if model performance was correctly computed relative to its upper bound. The impact of the noise ceiling misapplication is more pronounced in the case of low-to-intermediate ceilings, highlighting the need for reliable data to draw strong conclusions about a model’s distance to the noise ceiling.

Globally, such misunderstandings might propagate because researchers draw upon previous successes and failures when choosing modeling approaches, constructing benchmarks, and interpreting patterns across studies. Potentially, if a large enough number of studies within a research program overestimate how close a hypothesis instantiated in a model is to the noise ceiling, a field may prematurely deem a research question answered. We do not believe this is necessarily the case, but in light of the apparent prevalence of the issue, it is something to guard against. As we have explained, this can be done by carefully tracing the metrics used for noise ceiling and model performance. Besides this, simulating a perfect model can offer deeper insight into what behavior is plausible and implausible from an empirical model. In the Supplementary Material, we offer one simple version of such a simulation in its most general form. For a mathematical intuition, we also offer proofs for further reading.

This contribution specifically concerns noise ceilings estimated from the reliability across data splits. This approach towards noise ceilings is both common and powerful when implemented correctly, and this work should not be viewed as a critique of noise ceilings based on split-half correlations, but merely as a cautionary tale of one potential misapplication. Nevertheless, we refer readers to alternative approaches that may be appropriate depending on the nature of a research domain. For example, analytical noise ceiling may be derived in closed form (Allen et al., 2022), or a maximum achievable group-average model performance may be obtained from the across-subject consistency of models (Nili et al., 2014). In this latter method, a model is cross-validated using a population of individuals, where each individual is tested separately. These noise ceiling approaches are not subject to the misapplication identified in this work, which specifically concerns noise ceilings based on reliability estimates from split-half correlations.

To conclude, this paper offers a cautionary note with a constructive outlook. If noise ceiling estimates are made more consistent across future work, it will allow psychologists, neuroscientists and computational modelers to more accurately evaluate how models are performing relative to each other and in relation to a limit set by noise in the dataset.

## Acknowledgments

We thank Elina Schweppe for her assistance with the literature search. We are also grateful to Chris Baker, Mick Bonner, and Robin Ince for their helpful comments, and to Kohitij Kār, Sam Schwarzkopf, and Kendrick Kay for useful discussions.

## Funding

MNH was supported by the ERC Starting Grant COREDIM (ERC-2021-STG-101039712), a LOEWE Start Professorship of the Hessian Ministry of Higher Education, Research, Science and the Arts, and the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG), under the Collaborative Research Center “Cardinal Mechanisms of Perception” (222641018–SFB/TRR 135 TP C11) and Germany’s Excellence Strategy (EXC 3066/1 “The Adaptive Mind”, Project No. 533717223).

## Competing interests

The authors declare that they have no competing interests.

## References

- Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., and Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.*, 25(1):116–126.
- Brown, W. (1910). SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES <sup>1</sup>. *Br. J. Psychol.*, 3(3):296–322.
- Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.*, 10(12):e1003963.
- Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Res.*, 51(13):1484–1525.
- Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.*, 15(1):9383.
- Dumoulin, S. O. and Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–660.
- Hsu, A., Borst, A., and Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network*, 15(2):91–109.
- Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.
- Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.
- Konkle, T. and Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.*, 13(1):491.
- Lage-Castellanos, A., Valente, G., Formisano, E., and De Martino, F. (2019). Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Comput. Biol.*, 15(3):e1006397.
- Morgan, C. and Schwarzkopf, D. S. (2019). Comparison of human population receptive field estimates between scanners and the effect of temporal filtering. *F1000Res.*, 8:1681.
- Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410.
- Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Comput. Biol.*, 10(4):e1003553.
- Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.*, 15(1):72.
- Spearman, C. (1910). Correlation calculated from faulty data. *Br. J. Psychol.*, 3(3):271–295.
- Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365.

# Supplementary Materials

## Simulation

These simulations serve to demonstrate, using artificial data, how Spearman-Brown corrected split-half reliability metrics map onto model performance metrics. The results confirm two assumptions about the Spearman-Brown corrected split-half reliability:

(i) First, less centrally, it offers an unbiased estimation of the data’s reliability (under the simulated assumptions, which include additive noise and tau equivalence). That is, the Spearman-Brown correction derives the reliability of the dataset in its entirety, across all measurements, by extrapolating from the reliability of split halves.

(ii) Second, and key to our main claims, the corrected split-half reliability reflects the explained variance of a true model ( $R^2$ ), rather than its correlation coefficient ( $r$ ).

The simulation was designed as follows. First, we drew  $n = 1000$  data points from a latent distribution (i.i.d. standard normal distribution). For each data point, we generated  $m = 10$  noisy repetitions by adding independent Gaussian noise. We anchored the signal variance at  $\text{Var}(\text{signal}) = 1$  and defined

$$\text{SNR} = \frac{\text{Var}(\text{signal})}{\text{Var}(\text{noise})} = \frac{1}{\text{Var}(\text{noise})},$$

with SNR values logarithmically spaced from 0.01 to 2.56. Each simulation is performed  $k = 200$  times, randomizing the noise in each iteration.

To evaluate claim (i), we generated two independent datasets, A and B, drawn from the same latent distribution. We first calculated the Spearman-Brown corrected split-half estimate solely within Dataset A by splitting the  $m$  repetitions for each stimulus into a first and second half. Then, we averaged the  $m/2$  repetitions within each half and computed Pearson’s  $r$  between the two halves across the  $n$  stimuli. After computing the split-half correlation, we applied the Spearman-Brown correction ( $r_{\text{SB}} = 2r/(1+r)$ ) to the split-half reliability to estimate the reliability intrinsic to the full data. To evaluate if the Spearman-Brown correction achieves this, we established a benchmark for the ‘true’ reliability of the full data. Specifically, we computed the Pearson correlation between the averaged responses of Dataset A and the averaged responses of Dataset B. Since A and B represent independent realizations of the same  $m$ -trial experiment, their correlation ( $r_{A,B}$ ) reflects the actual test-retest reliability of the full dataset. Then, we compared the Spearman-Brown estimate to this benchmark. We found that the corrected split-half estimate consistently tracked the correlation between the two datasets across SNR conditions (Figure S1). This confirms that the Spearman-Brown correction successfully extrapolates from half the data to estimate the reliability of the full dataset.

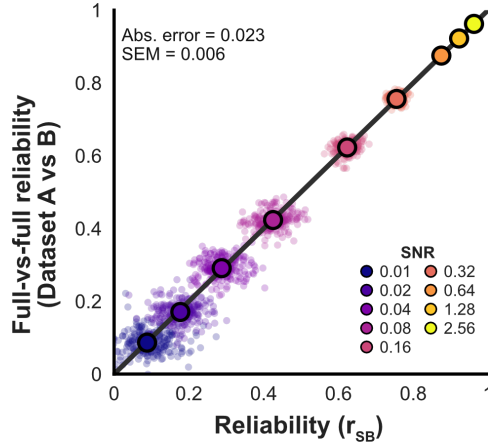


Figure S1: **Spearman-Brown correction captures full reliability.** Across noise levels, the Spearman-Brown corrected split-half reliability approximates the reliability of the full data, as operationalized by the correlation between two datasets drawn from the same underlying distribution. Individual points reflect simulation iterations ( $k = 200$ ) for a given SNR condition. Absolute error = mean absolute difference between  $r_{SB}$  and the full-data reliability across SNRs and iterations.

Second, to evaluate claim (ii), we devised a model that perfectly captures the true underlying signal used to generate the ground-truth data. We then illustrate the relation between the performance of this *oracle model* and the Spearman-Brown corrected reliability, which we have shown matches the data’s full reliability. Consistent with the central message of this paper, the model’s explained variance ( $R^2$ ) matches the corrected reliability ( $r_{SB}$ ), while its correlation aligns with the square root of that reliability (model  $r \approx \sqrt{r_{SB}}$ ; Figure S2). This confirms the mapping outlined in Table 1, underscoring that the reliability represents the proportion of observed variance that is true signal variance.

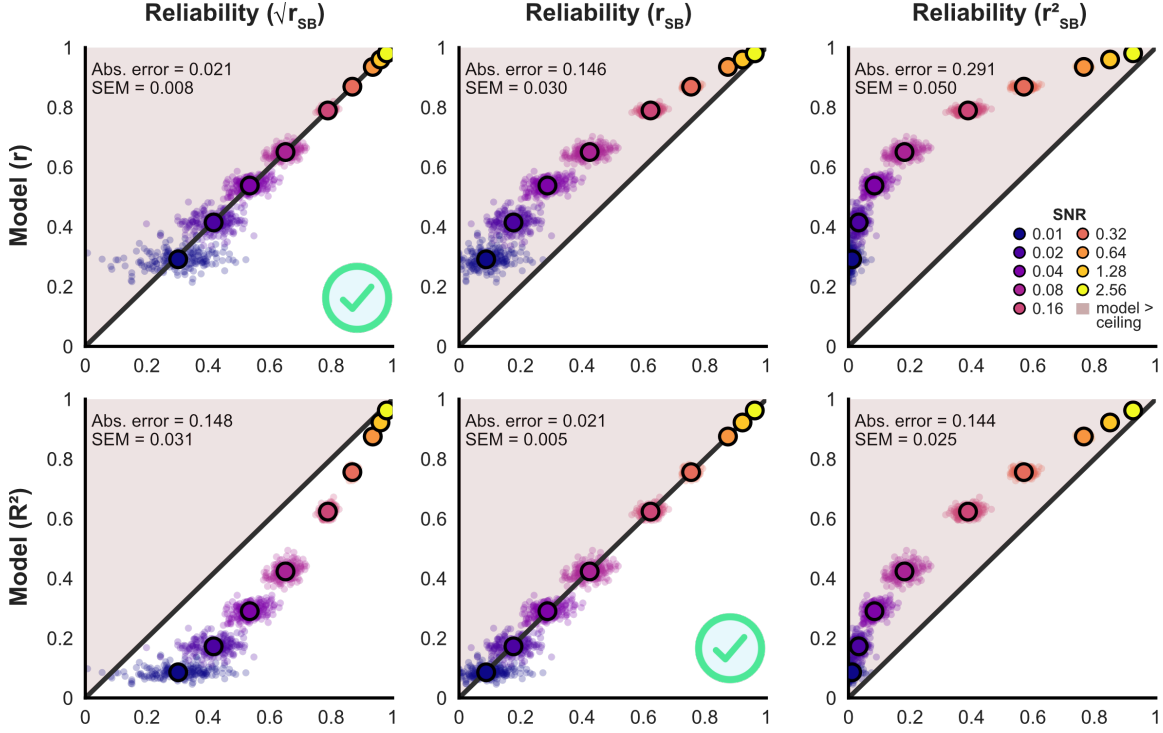


Figure S2: **Mapping model performance to noise ceiling.** Oracle model performance (model  $r$ ) tracks the square root of the Spearman-Brown corrected split-half reliability, and its explained variance ( $R^2$ ) approximates the reliability ( $r_{SB}$ ). Using model correlation in combination with reliability ( $r_{SB}$ ) leads to an overestimation of model performance relative to ceiling. The shaded red area reflects instances of a true model exceeding the noise ceiling.

## Proof from test theory

Suppose we try to measure the same data generating quantity  $T$  twice, yielding two parallel measurements  $Y$  and  $Y'$ . Parallel means they target the same  $T$  and, importantly, have the same error variance. Thus, we can write

$$Y = T + e, \quad Y' = T + e' \quad (1)$$

with three important assumptions: First,  $e$  and  $e'$  are independent of each other, which is standard in repeated measurements so that random noise sources are unshared; second, the errors are independent of  $T$ , since by definition, measurement error cannot be related to signal; and third,  $\text{Var}(e) = \text{Var}(e')$ , since the measurements are assumed to be equally precise. Note that equal variance does not imply the errors are identical; they merely have the same degree of variability.

Our aim here is to determine how well measurement  $Y$  can predict another measurement  $Y'$ . Specifically, we are interested in the correlation between  $Y$  and  $Y'$ , defined as:

$$r_{YY'} = \frac{\text{Cov}(Y, Y')}{\text{SD}(Y) \cdot \text{SD}(Y')} \quad (2)$$

Let's first focus on the numerator. Given additivity of covariances, we can write

$$\text{Cov}(Y, Y') = \text{Cov}(T + e, T + e') = \text{Cov}(T, T) + \text{Cov}(T, e') + \text{Cov}(e, T) + \text{Cov}(e, e')$$

Given the independence assumptions, the terms  $\text{Cov}(T, e')$ ,  $\text{Cov}(e, T)$ , and  $\text{Cov}(e, e')$  must equal zero. Thus, we have:

$$\text{Cov}(Y, Y') = \text{Var}(T) \tag{3}$$

Next, for the denominator, the variance of each test  $Y$  or  $Y'$  is the sum of the variance of the true score and the variance of the measurement error:

$$\text{Var}(Y) = \text{Var}(T + e) = \text{Var}(T) + \text{Var}(e), \quad \text{Var}(Y') = \text{Var}(T + e') = \text{Var}(T) + \text{Var}(e')$$

Recall that  $\text{Var}(e) = \text{Var}(e')$ . Therefore, the denominator simplifies to:

$$\text{SD}(Y) \cdot \text{SD}(Y') = \sqrt{\text{Var}(T) + \text{Var}(e)} \cdot \sqrt{\text{Var}(T) + \text{Var}(e)} = \text{Var}(T) + \text{Var}(e)$$

Putting numerator and denominator together, the correlation simplifies to:

$$r_{YY'} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(e)} = \frac{\text{Var}(T)}{\text{Var}(Y)} \tag{4}$$

Notice that this highlights that reliability can alternatively be defined as the correlation of parallel measurements or the fraction of observed variance attributable to the true score  $T$ .

If this proof is not fully convincing yet, we can now introduce  $X$  as the perfect predictor of  $T$ , which is the best possible model. Then,

$$\begin{aligned} r_{XY} &= \frac{\text{Cov}(X, Y)}{\text{SD}(X) \cdot \text{SD}(Y)} = \frac{\text{Var}(T)}{\sqrt{\text{Var}(T)} \cdot \sqrt{\text{Var}(T) + \text{Var}(e)}} \\ &= \frac{\sqrt{\text{Var}(T)}}{\sqrt{\text{Var}(T) + \text{Var}(e)}} = \sqrt{\frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(e)}} = \sqrt{r_{YY}} \end{aligned} \tag{5}$$

Hence, for this perfect predictor,

$$r_{XY}^2 = r_{YY} \tag{6}$$

Therefore, the fact that one of the square roots in the denominator cancels part of the numerator turns the entire expression for the reliability into a square root. This proof from classical test theory thus confirms the earlier intuition that the maximum proportion of variance that can be explained is given by the reliability of the data and explains its source.