Getting Blood from a Stone: Improving Neural Inferences without Additional Neural Data

David J. Halpern[1] and Todd M. Gureckis[1]

[1]Department of Psychology, New York University

January 19, 2021

Abstract

In recent years, the cognitive neuroscience literature has come under criticism for containing many low-powered studies, limiting the ability to make reliable statistical inferences. Typically, the suggestion for increasing power is to collect more data with neural signals. However, many studies in cognitive neuroscience use parameters estimated from behavioral data in order to make inferences about neural signals (such as fMRI BOLD signal). In this paper, we explore how cognitive neuroscientists can learn more about their neuroimaging signal by collecting data on *behavior alone*. We demonstrate through simulation that knowing more about the marginal distribution of behavioral parameters can improve inferences about the mapping between cognitive processes and neural data. In realistic settings of the correlation between cognitive and neural parameters, additional behavioral data can lead to the same improvement in the precision of inferences more cheaply and easily than collecting additional data from subjects in a neuroimaging study. This means that when conducting an neuroimaging study, researchers now have two knobs to turn in a design analysis: the number of subjects collected in the scanner and the number of behavioral subjects collected outside the scanner (in the lab or online).

1 Statistical Power and Neuroimaging

One factor which limits progress in human cognitive neuroscience is statistical precision and power (Munafò et al., 2019; Cremers et al., 2017; Yarkoni, 2009). Many reported effects are small relative to the noise in neuroimaging signals, a problem compounded by the small samples sizes used in most studies. Within a hypothesis testing framework, a lack of precise estimates of some measurements can make it difficult to replicate reported effects. For example, even quite strong, known effects such as the relationship between motor movements and fMRI BOLD activation in sensorimotor regions can be unlikely to reach significance in current standard sample sizes of around 25-30 subjects (Poldrack et al., 2017). Having lower precision also means that any estimated effect that is significant will also have a high probability of the estimate being much greater in magnitude or even having the wrong sign compared to the true effect (Gelman and Tuerlinckx, 2000; Cremers et al., 2017; Yarkoni, 2009). However, due to a combination of factors (including lack of sufficient funding for large sample studies), research in human cognitive neuroscience tends to be under-powered in many cases (Poldrack et al., 2017).

Several remedies have been proposed for this situation, the most obvious being to simply collect more neuroimaging data (Poldrack et al., 2017; Munafò et al., 2019). Because the funding available to most labs is relatively small (compared to the cost of a large scale neuroimaging study), this might require a move towards working in larger consortia around critical topics. Another approach might be to leverage the power of publicly available open datasets, which can allow for meta-analyses across several smaller experiments (Poldrack et al., 2017). However, increasing sample sizes is not the only option for dealing with a lack of precision. For example, increasing the signal-to-noise ratio of neuroimaging measures by creating more detailed statistical models of the neuroimaging signal (Lindquist et al., 2009), improving experimental design (Durnez et al., 2017) or even improving the measurement process itself (Feinberg and Yacoub, 2012; Lombardo et al., 2016) are all measures that are likely to help. In addition, it has been suggested that a focus on computational cognitive modeling and behavior can also allow for extracting more signal (Krakauer et al., 2017; Palmeri et al., 2017) by creating better models of the underlying cognitive and neural processes. In that theme, this paper explores a potentially under-studied option for improving the statistical precision of human cognitive neuroscience studies: collecting additional behavioral data without neural recordings.

We begin by noting that many common statistical analyses in cognitive neuroscience involve behavioral data, typically collected from the subjects who also provided neural data. In one common scenario, the interest is in the relationship between individual differences in behavioral traits and more static measures of neural activity, such as performance in a task and functional connectivity (Rosenberg et al., 2016). In another typical analysis, researchers are interested in how within-subject changes in a cognitive state variable over time relate to changes in the neural signal from a particular brain region, such as the prediction error in a reinforcement learning task (O'Doherty et al., 2003).

One interesting, but often overlooked, point is that these behavioral measurements (i.e. survey questionnaires, behavioral responses to stimuli, summaries of task performance or estimates of cognitive model parameters) are themselves noisy estimates of an individual's underlying cognitive processes and the distributions of these patterns that exist in the population. Thus our ability to infer relationships between brain and behavior is limited by *sampling error* and *estimation error*. While we define these types of error in greater detail below, we give a quick summary here. Sampling error is a result of the fact that a small sample may not be representative of the population, so estimates of the relationship based on the sample may be far from the true relationship. A small sample therefore prevents us from being confident about the value of a parameter in a statistical model. Estimation error is a form of measurement error that arises because researchers cannot measure latent cognitive variables directly and instead use models to estimate their values. The relationship between neural recordings and cognitive variables measured with error will necessarily not be as strong as the relationship involving the true values, a notion sometimes known as regression dilution or attenuation. Ignoring these two sources of error can prevent us from learning as much as possible from our neural data.

In the following, we demonstrate that one way to reduce both sources of error and improve statistical estimates of the relationship between the brain and behavior is to collect behavioral data from subjects *without collecting corresponding neuroimaging data*. This idea is somewhat counterintuitive in light of the fact that most common approaches in neuroimaging analyses assume that the only relevant data is from individuals who's neural activity was recorded. However, our interest in this issue stems from the fact that behavioral data (especially data collected over the Internet) is often significantly less resource-intensive to collect than neuroimaging data, in terms of money and the time and effort of both researchers and human subjects. We show how using modern statistical methods that can leverage information from the additional behavioral data can help mitigate the two types of error. To anticipate one consequence of our analysis, we are able to provide a cost-benefit analysis (in terms of statistical power and precision) of collecting data from subjects with only behavior against collecting more expensive additional data from subjects with both behavior and neural recordings. This analysis shows that in some cases, due to its relatively low cost, a more effective use of resources is to collect more behavioral data instead of additional neuroimaging subjects. This result has implications for designing studies in cognitive neuroscience both prospectively and retrospectively (i.e., the re-analysis of archived open-science data sets.)

1.1 Common paradigms for relating brain and behavior in neuroimaging studies

For the purposes of articulating our thesis, it will be useful to describe two common types of analyses used in cognitive neuroscience that use behavioral data to interpret neural signals and lay out some notation for the rest of the paper. While this is by no means an exhaustive list, we claim that many analyses

can be understood as falling into two broad categories.

Regression (Conditional Distribution) In many cognitive neuroscience studies, researchers are interested in how the conditional distribution of neural measures changes as a function of variation in cognitive states induced by a task. For instance, researchers might be interested in regions of the brain that track whether or not it is viewing a face (Kanwisher et al., 1997), what the valence of a stimulus is (Chikazoe et al., 2014) or the prediction error in a reinforcement learning model of decision making (O'Doherty et al., 2003). More formally, a subject i presented with a set of stimuli $\mathbf{s}_i thatoccurontrials(ortimes)$t $\in \{1..T_i\}$. Each stimulus presentation s_{it} is assumed to induce some latent cognitive state c_{it} that is associated with some neural activity n_{it}. Researchers can identify neural signals related to the cognitive states by fitting a regression model where

$$n_{it} = \alpha_i + \beta_{c_i} c_{it} + \epsilon_{it} \tag{1}$$

where ϵ_{it} is uncorrelated noise (for the purposes of this discussion, we ignore features of particular signals like the fMRI BOLD hemodynamic response). However, this is not quite so straightforward because researcher need to derive estimates of the cognitive state for these analyses. The most common assumption is that the relevant states are closely tied with objective features of the stimulus, such as the color of an image or experimental conditions like the reward offered for a choice, so c_{it} can reasonably be replaced by the stimulus feature or experimental condition itself.

In other cases, the cognitive states of interest are tied to subjective features of the stimulus, such as the valence associated with s_{it} or the typicality of s_{it} for a particular category. In some cases, researchers collect behavioral ratings b_{it} which depend on these latent states. Researchers then assume that the ratings come from a cognitive process C such that $b_{it} \sim C(c_{it}, s_{it})$ and obtain the c_{it} from that model. In the simplest case, this can mean simply using the valence ratings given by subjects (Chikazoe et al., 2014) or using the average typicality from a group of subjects tested separately (Wilson-Mendenhall et al., 2015).

More recently, "model-based" neuroimaging analyses have investigated cognitive states that are involved in a particular cognitive process C such that in addition to the current stimulus s_{it}, c_{it} may also depend on θ_i are latent cognitive "trait" variables that do not vary over the relevant time as well as the past states $\mathbf{c}_{i_{1:(t-1)}}$. Researchers can then use the behavioral data to fit the model $b_{it} \sim C(\theta_i, c_{it}, s_{it})$ and obtain the relevant c_{it}. These states may be, for instance, the reward prediction error or predicted value in a reinforcement learning model of sequential decision making (O'Doherty et al., 2003).

Because the latter two types of analyses involve behavioral data, those will be our focus in this paper. In particular, we will focus on the model-based case but nearly all of our analyses will also apply to the second case.

Modern multivariate pattern analysis techniques (Polyn, 2005; Haxby, 2001) invert this regression and try to predict the cognitive state from neural signals. Typically, this means predicting stimulus properties or experimental conditions but in some cases the cognitive state will be inferred from behavioral data (e.g. Mack et al., 2013; Kragel and Polyn, 2016). While the example analyses in the rest of the paper do not directly apply to this setting, similar principles of collecting additional behavioral data may hold there as well.

Correlation (Joint Distribution) In other cases, researchers are interested in the joint distribution of behavioral measures and neural measures (often across subjects). For instance, is there a relationship between individual differences in behavior and individual differences in static measures of the brain? Researchers often use a direct measure of a subject's behavior b_i such as answers to a questionnaire (Treadway et al., 2013) or performance on a behavioral task (such as a sustained attention task Rosenberg et al., 2016). More recently (e.g. Homan et al., 2019), researchers have also used cognitive "trait" variables which can be related to behavioral data through cognitive models (i.e. θ_i above). Static measures can include measures like functional connectivity or average resting state activity that do not depend on time. Typically we only get one n_i per subject and the measure is not always collected simultaneously to the task where b_i is recorded. In this setting, researchers are typically interested in characterizing the joint distribution of neural measures n_i and behavioral (b_i) or latent cognitive (θ_i) trait variables. In particular, if the neural measures and behavioral variables can be approximated by a bivariate normal, researchers want to find neural measures where the correlation ρ is non-zero.

Another common analysis that involves correlation is Representational Similarity Analysis (Kriegeskorte, 2008). In these analyses, researchers correlate the similarities of the neural response to pairs of stimuli with the similarities of the two stimuli according to a model. Most often, these models are created based on stimulus properties but they are sometimes derived from behavioral ratings (Chikazoe et al., 2014; Bruffaerts et al., 2013). In these cases, our analyses about the benefits of additional behavioral data will apply as well.

We argue that these two categories of research design and inferential techniques cover a vast array of studies in cognitive neuroscience and neuroimaging in particular. Each of these analyses can be thought of as attempting to learn about the *conditional* distribution of neural signals given behavioral data (**regression**) or the *joint* distribution of neural and behavioral data (**correlation**). The focus of this paper asks whether we can improve inferences for these two types of analyses by learning more about the *marginal* distribution of behavioral data, that is, collecting more behavioral data without neural data.

2 Two Sources of Error in Neuroimaging Analyses

As the Introduction laid out there are two key sources of error in neuroimaging analyses. In this section, we describe each of the two sources of error in more detail as well as how they can affect the precision of inferences about the relationship between neural signals and behavior.

2.1 Sampling Error

In a **correlation** analysis, we want to know what ρ is in the general population but estimate it by using the sample correlation r. With a small number of subjects, r can be quite far from the true ρ. An approximate confidence interval for ρ can be constructed as $\text{arctanh}(r) \pm \frac{1}{\sqrt{N-3}}$, where N is the number of subjects or, more generally, the number of data points used to estimate r (Fisher, 1915). As this makes clear, the precision of the estimate (and thus finding a ρ that we can say is non-zero with confidence) is constrained by the size of our sample. In many neuroimaging studies this is often relatively small because neuroimaging data is expensive and time consuming to collect. In Figure 1, we show an example of where the smaller sample size can lead to noisy estimates and larger error bars. [1] Sampling error can also be an issue in **regression** analyses as well. However, because in that case we are only interested in the conditional distribution of the neural signals, we cannot reduce sampling error by collecting more behavioral data so do not discuss it here.

2.2 Estimation Error

In the **model-based regression** analysis described above, we want to find a set of neural signals that have a linear relationship with a subject's latent cognitive states $\mathbf{c}_i$. [2] The true model we are interested in is

$$n_{it} = \alpha_i + \beta_{c_i} c_{it} + \epsilon_{it} \tag{2}$$

where in particular we are interested in which neural signals have a significant β_{c_i} parameter (typically, at the group level). In practice, because we don't have direct access to c_{it} we typically replace it with estimates $\hat{c}_{it}$ by fitting the model $C(\theta, \mathbf{c}_{it}, \mathbf{s}_{it})$ to the behavioral data b_{it}. It is typically assumed that the model doesn't perfectly account for behavior and that there is some variability in responses beyond what can be predicted by the model. For instance, when fitting a reinforcement learning model, a decision noise parameter is often used to deal with variability beyond what can be accounted for by the latent representations

[1] All figures and simulations in this paper are created using custom code in `python 3.7.7` using the packages `pandas` (McKinney, 2010), `numpy` (van der Walt et al., 2011), `matplotlib` (Hunter, 2007) and `seaborn` (Waskom et al., 2020)

[2] In the following we will focus on the case where we are interested in a single latent cognitive variable.

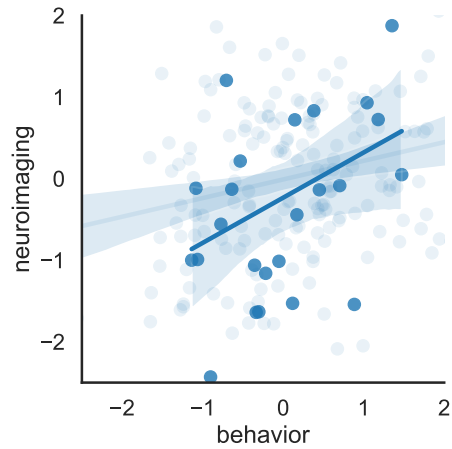

Figure 1: 25 samples from a multivariate Gaussian distribution are overlaid over an additional 100 samples from the same distribution. Regression lines are plotted showing how estimates from smaller samples can add significant variance to the estimated correlation. Error bars show that we cannot rule out a zero correlation in the smaller sample.

in the model. Because we don't have infinite data, this typically means that several sets of parameters could account for the data even if the model is well identified. In other words, the parameters $\hat{\mathbf{c}}_{it}$ are estimated with some uncertainty. This may seem obvious to many readers but standard techniques that are typically used in cognitive neuroscience do not easily allow this uncertainty to be taken into account in the above regression. Instead of the model above, researchers often simply plug in the estimated $\hat{\mathbf{c}}_i$ into the regression, i.e.

$$n_{it} = \alpha_i + \beta_{\hat{c}_i}\hat{c}_{it} + \epsilon_{it} \tag{3}$$

In this model, the value of the slope estimate using the estimated latent variables, $\beta_{\hat{c}_i}$, will not necessarily be the same as when using the true parameter β_{c_i}. In fact, it can be shown that on average, the least squares estimate of $\hat{\beta}_{\hat{c}_i}$ will be

$$E(\hat{\beta}_{\hat{c}_i}) = \beta_{c_i}\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i}) \tag{4}$$

where $\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})$ is the correlation between the true $\mathbf{c_i}$ and the estimates $\hat{\mathbf{c}}_\mathbf{i}$ (Wilson and Niv, 2015). This is perhaps not always relevant since the neuroimaging signal can often only be known up to a multiplicative constant. However, Wilson and Niv (2015) also showed that, as long as the estimates are uncorrelated with the regression error term ϵ_{it}, the expected t-statistic for the $\hat{\beta}_{\hat{c}_i}$ is

$$\hat{t}(\beta_{c_i}, \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i}), T_i, \sigma_\epsilon^2) = \beta_{c_i}\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})\sqrt{\frac{T_i - 2}{\sigma_\epsilon^2 + \beta_{c_i}^2(1 - \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2)}} \tag{5}$$

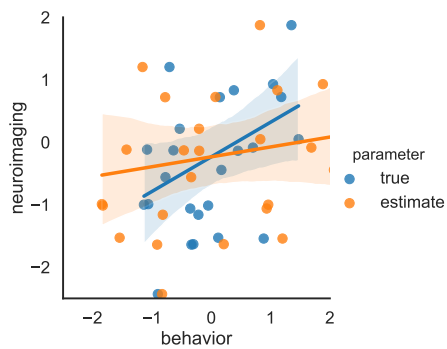

Figure 2: Here we show a case where the latent behavioral measure is estimated with error. The correlation with the estimates is lower than it would be with the true latent behavioral measure.

where T_i is the number of trials from subject i used to estimate the regression and σ_ϵ^2 is the variance of the residuals ϵ_{it}, due to noise in the neuroimaging signal. We can see that this t-statistic gets closer to 0 as $\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})$ gets further from 1. A smaller t-statistic means that the standard error (or any multiple of it such as a confidence interval) is more likely to overlap with 0. Thus having a lower correlation between the estimates and the true values, or, in other words, greater estimation error, can make it harder to find relevant signals in the typical regression setting.

Estimation error can also be relevant for **correlation** analyses looking at correlations between latent cognitive trait variables and neural signals. If we estimate the trait variable θ_i from behavioral data, Katahira (2016) showed that the correlation between estimates $\hat{\theta}$ and structural neural signals $\mathbf{n}$, can be decomposed as

$$\rho(\mathbf{n}, \hat{\theta}) = \rho(\mathbf{n}, \theta)\rho(\theta, \hat{\theta}) \tag{6}$$

Therefore, the correlation is upper bounded by the true correlation between the neural signals and latent trait variables $\rho(\mathbf{n}, \theta)$ and will decrease as the correlation between the estimates and the true values $\rho(\theta, \hat{\theta})$ decreases (or as estimation error increases). Because the size of confidence intervals for a correlation coefficient only depend on the total number of subjects, the intervals for a correlation with a noisier estimate will be more likely to overlap with zero. This means that again, greater estimation error will make it harder to find relevant signals.

3 The Benefits of Additional Behavioral Data

Having laid out why these two errors can deteriorate our inferences, we now demonstrate how collecting additional behavioral data can help. In addition, for each case, we lay out a tradeoff between collecting cheaper behavioral data

and more expensive imaging data in terms of how much each improves our inferences. These analyses show in which cases it may actually be more valuable to collect behavioral data than neuroimaging data, even if the main object of interest is the relationship between brain and behavior. To preface, section 3.1 demonstrates how little known estimators are able to make use of additional behavioral data to reduce sampling error for correlation analyses. We then follow with simulations to understand where this estimator improves over standard ones. Section 3.2 then documents how the use of hierarchical modeling can allow for combining datasets and improving estimation for common regression analyses as well as correlation analyses. Further analytical work shows the benefits of decreasing estimation error for making group-level inferences about the relationship between cognitive processes and neural signals.

3.1 How to Decrease Sampling Error: Behavioral Data as Neuroimaging Data with a Missing Variable

Typically we are interested in generalizing beyond our sample and characterizing the joint distribution of neural signals and behavioral variables in the population. However, as mentioned, most neuroimaging studies are limited by getting a relatively small sample. One somewhat counter-intuitive notion from the field of statistics focused on missing data is that we can often get better estimates of a parameter in the population (in terms of a lower mean squared error) by including data points that do not have every variable recorded in our analysis (Little and Rubin, 2019). This is because we are often assuming that variables have some joint distribution. If the correlation between two variables is not zero then we can know something about the value of what a missing variable must be from knowing the value of the other variable. If we have collected neural and behavioral data from N_{nb} subjects and only behavioral data from an additional N_b subjects, Anderson (1957) showed that the maximum likelihood estimates of the mean and standard deviation of $\mathbf{n}$, as well as the correlation between $\mathbf{n}$ and $\mathbf{b}$, ρ, are not the same when we include all of the data points as they would be if we only include the data with all variables recorded (the complete data). We can see this by rewriting the likelihood for a bivariate normal. If

$$(\mathbf{n}, \mathbf{b}) \sim N(\mu, \Sigma) \tag{7}$$

where $\mu = (\mu_{\mathbf{n}}, \mu_{\mathbf{b}})$ and $\Sigma = \begin{pmatrix} \sigma_n^2 & \rho\sigma_n\sigma_b \\ \rho\sigma_n\sigma_b & \sigma_b^2 \end{pmatrix}$ then, we can factor the likelihood as

$$P(\mathbf{n}, \mathbf{b}) = P(\mathbf{n}|\mathbf{b})P(\mathbf{b}) = N(\mathbf{n}; \alpha_{nb} + \beta_{nb}\mathbf{b}, \sigma_{nb})N(\mathbf{b}; \mu_b, \sigma_b) \tag{8}$$

The advantage of writing the joint distribution as we did above is that the first term is just the regression of $\mathbf{n}$ on $\mathbf{b}$ which is only defined for the data points that have both variables recorded. Therefore, we can get the maximum likelihood likelihood estimate for α_{nb}, β_{nb} and σ_{nb} using standard least squares regression estimates with the N_{nb} data points. Maximizing the likelihood of the second

term only depends on the **b** data so we can use standard estimates of the mean and variance using all of the data. We can write the correlation as

$$\rho = \frac{\mathrm{Cov}(\mathbf{b}, \mathbf{n})}{\sigma_n \sigma_b} \tag{9}$$

Then plugging in the maximum likelihood estimates $\hat{\beta}_{nb}$, $\hat{\sigma}_{nb}$ and $\hat{\sigma}_b$, we get the maximum likelihood estimate for ρ using all of the data:

$$\hat{\rho} = \frac{\beta_{nb}\sigma_b^2}{\sigma_b\sqrt{\sigma_{nb}^2 + \beta_{nb}^2\sigma_b^2}} = \frac{\beta_{nb}\sigma_b}{\sqrt{\sigma_{nb}^2 + \beta_{nb}^2\sigma_b^2}} \tag{10}$$

Because the maximum likelihood estimate of σ_b using all $N_{nb} + N_b$ data points will not in general be the same as only using the N_{nb} data points, the maximum likelihood estimate of ρ using all the data will not be the same as using only the complete data.

We can now try to illuminate how this new estimate is different from the correlation estimate based on only the N_{nb} complete cases. Little and Rubin (2019) derive the missing data estimator $\hat{\rho}$ in a different way, showing that

$$\hat{\rho} = r\left(\frac{\hat{\sigma}_b}{\tilde{\sigma}_b}\right)^{\frac{1}{2}}\left(\frac{\tilde{\sigma}_n}{\hat{\sigma}_n}\right)^{\frac{1}{2}} \tag{11}$$

where r, $\tilde{\sigma}_b$, $\tilde{\sigma}_n$ are the estimates of the correlation (i.e. Pearson's r), the variance of the behavioral variable and the variance of the neural variable among the N_{nb} complete cases. $\hat{\sigma}_n$ is the variance of n after adjusting for the $\hat{\sigma}_b$ estimate based on all of the data. Little and Rubin (2019) derive the formula for $\hat{\sigma}_n$ as

$$\hat{\sigma}_n = \tilde{\sigma}_n + \hat{\beta}^2(\hat{\sigma}_b - \tilde{\sigma}_b) \tag{12}$$

We can instead write this in terms of r and rearrange terms, i.e.

$$\hat{\sigma}_n = \tilde{\sigma}_n + \frac{r^2}{\tilde{\sigma}_n}(\hat{\sigma}_b - \tilde{\sigma}_b) = \tilde{\sigma}_n\left(1 + r^2\left(\frac{\hat{\sigma}_b}{\tilde{\sigma}_b} - 1\right)\right) \tag{13}$$

Plugging this back in, we get

$$\hat{\rho} = r\left(\frac{\hat{\sigma}_b}{\tilde{\sigma}_b}\right)^{\frac{1}{2}}\left(\frac{\tilde{\sigma}_n}{\tilde{\sigma}_n(1 + r^2(\frac{\hat{\sigma}_b}{\tilde{\sigma}_b} - 1))}\right)^{\frac{1}{2}} \tag{14}$$

$$= r\left(\frac{\hat{\sigma}_b}{\tilde{\sigma}_b}\right)^{\frac{1}{2}}\left(\frac{1}{\left(1 + r^2\left(\frac{\hat{\sigma}_b}{\tilde{\sigma}_b} - 1\right)\right)}\right)^{\frac{1}{2}} \tag{15}$$

To our knowledge, we are the first to show that the Anderson (1957) estimate, $\hat{\rho}$, only depends on the correlation estimate in the complete cases, r and the ratio of the variance of the behavioral variable in the complete cases and the full data set, $\frac{\hat{\sigma}_b}{\tilde{\sigma}_b}$. This form makes clear how the new information on the

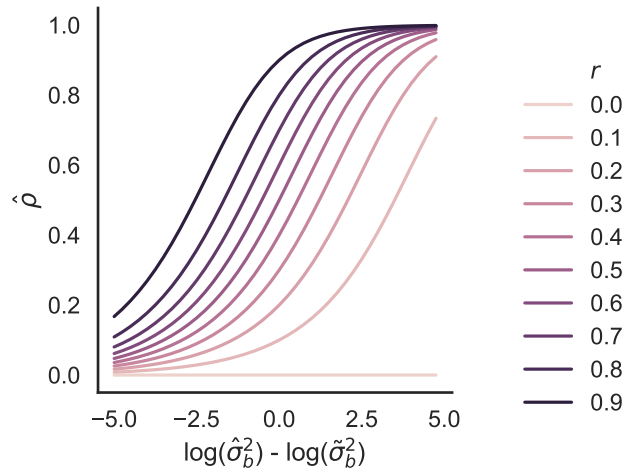

Figure 3: Anderson (1957) estimate ($\hat{\rho}$) of the correlation using all of the data as a function of Pearson's r using only N_{nb} complete cases and the difference in the log of the variance of the behavioral data in the N_{nb} complete cases and all N_b cases. In general $\hat{\rho}$ is modified up or down by the relationship between the variance of the behavioral data on the full data set and that for the complete cases.

marginal distribution of the behavioral data contributes to the estimated measure of association. In figure 3, we plot how $\hat{\rho}$ changes as a function of these two variables. In general, if the variance increases when including the additional data, the correlation estimate increases and visa versa.

3.1.1 When does it work?

Garren (1998) showed analytically that asymptotically, the maximum likelihood estimate $\hat{\rho}$ using all the data will have lower error than estimates for r using only the complete data. This result suggests that there can be cases where, if one variable is more expensive to collect than the other one (for instance neuroimaging data and behavioral data), the same quality of estimates of the correlation can can be achieved for lower cost by collecting some data with the more expensive variable missing (Hocking and Smith, 1972). However, the result is an asymptotic result, meaning that it is only guaranteed to apply in the limit of infinitely large data sets. It is therefore still unclear when collecting more behavioral data will be better than collecting more neuroimaging data in practice. In the following, we conduct a simple simulation to explore this tradeoff.

Following the description above, suppose we have collected a behavioral variable ($\mathbf{b}$) and a neuroimaging variable ($\mathbf{n}$) from N_{nb} subjects. In addition, we collected just behavioral data from N_b subjects. For simplicity, we assume that

we have a single behavioral measure and a single neuroimaging measure from each subject.

In order to specify the generative model with missing data, we can create a matrix of data $\mathbf{BN}$ from our neuroimaging subjects, an N_{nb}-by-2 matrix with $\mathbf{n}$ as the first column and $\mathbf{b}_{1:N_{nb}}$ as the second. The data matrix is generated for $i \in [1, N_{nb}]$ according to

$$(n_i, b_i) \sim N(\mu_{nb}, \mathbf{\Sigma}_{nb}) \tag{16}$$

where $\mathbf{\Sigma}_{nb_{1,1}} = \sigma_n^2$, $\mathbf{\Sigma}_{nb_{2,2}} = \sigma_b^2$ and $\mathbf{\Sigma}_{nb_{1,2}} = \mathbf{\Sigma}_{nb_{2,1}} = \rho\sigma_n^2\sigma_b^2$. The additional behavioral subjects then have distribution

$$\mathbf{b}_{(N_{nb}+1):N_b} \sim N(\mu_b, \sigma_b^2) \tag{17}$$

For the purposes of our simulation, we set $\mu_{nb} = 0$ and $\sigma_n^2 = \sigma_b^2 = 1$.

Having generated data according to the above model, we now compare two ways of estimating the correlation, standard Pearson's r using only data with both variables and the Anderson (1957) estimate of the correlation using all of the data.

We compare estimates with for four values of N_{nb} (25, 50, 100 and 200) and ten values of ρ from 0 to .9. To investigate the benefit of the additional behavioral data for methods that can make use of the information, for each value of N_{nb}, we test five values of N_b (0, 25, 50, 100, 200). Finally, to compare the benefit of the additional behavioral data to collecting an additional neuroimaging subject, we compare all estimates to Pearson's r with one additional nueroimaging subject. For each set of (N_{nb}, N_b, ρ), we generate 10,000 datasets.

Of note here is that collecting neuroimaging data can be quite expensive. For instance in 2019, collecting fMRI data at New York University's Center for Brain Imaging costs around $450 per hour of subject time. In addition to the price, collecting neuroimaging data can have many other non-monetary costs such as requiring the subject to come into the lab, requiring the time of trained researchers or technicians to run the machine and only being able to run one subject at a time. In contrast, many behavioral studies can be easily run in parallel online without significant monitoring by researchers for a cost of around $11 per hour of subject time (Hara et al., 2018; Crump et al., 2013). Thus, if 40 behavioral subjects can improve inferences more than a single fMRI subject, this might be a worthwhile tradeoff in many experimental designs.

Figure 4 plots how additional behavioral data can influence the mean squared error of the correlation estimate using Anderson's estimator. Comparing the two methods of correlation estimation, we find that the effect of the additional behavioral data can be nonlinear in the size of the underlying neuroimaging dataset and the true value of ρ. Having too few neuroimaging data points to estimate a weak correlation (e.g. only 25 subjects for a correlation of .1) makes the behavioral data less useful and it can be better to collect more neuroimaging data. This finding resembles similar simulation results from Garren (1998). At low amounts of neuroimaging data and lower true values of ρ, the Anderson estimate of the correlation using all of the behavioral data can actually perform

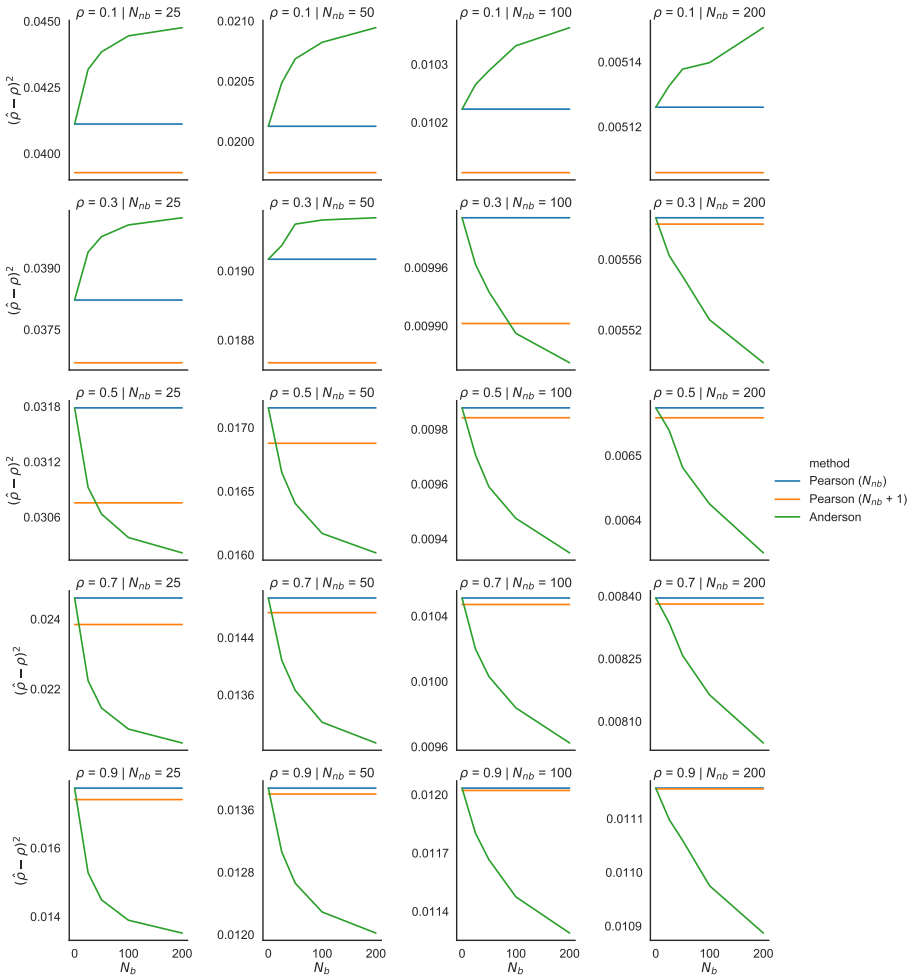

Figure 4: Mean squared estimation error, averaged over 10,000 simulated datasets, for the estimate of ρ from the two estimators as a function of the true correlation, the number of neuroimaging subjects N_{nb} and number of additional behavioral subjects N_b. We also compare to estimates from a neuroimaging dataset with $N_{nb} + 1$ subjects. The orange line is generally below the blue line showing the improved estimate that comes from one additional neuroimaging subject. The curved green line shows the Anderson (1957) $\hat{\rho}$ changes in the estimate as the number of addition behavior only subject are added to the analysis.

worse than the standard Pearson estimate. However, as the true ρ and number of neuroimaging subjects increase, the value of the additional data increases such that collecting additional behavioral data can be better than collecting a smaller amount of additional neuroimaging data. In some cases, collecting just 25 behavioral subjects can improve precision more than another neuroimaging subject. At the cost levels described above, this can mean that behavioral data can be a better investment than equivalently priced neuroimaging data. Therefore, a design analysis that takes costs into account has the potential to recommend collecting behavioral data rather than neuroimaging data.

However, as shown in Figure 4, that is not true across the whole parameter range and, perhaps in particular, it isn't true for low sample sizes and low true correlation values, which is exactly where we might hope to gain in terms power. Why is is that the Anderson estimator performs worse here? To answer this question, we can decompose the mean squared error into the bias and the variance, i.e.

$$MSE(\hat{\rho}, \rho) = E((\hat{\rho} - \rho)^2) = Bias(\hat{\rho}, \rho)^2 + Var(\hat{\rho}) \tag{18}$$

where Bias is defined as

$$Bias(\hat{\rho}, \rho) = E(\hat{\rho} - \rho) \tag{19}$$

As with mean squared error, we can also look at how various methods perform in terms of bias, i.e. is the estimate of ρ equal to ρ on average? It is well known that Pearson's r is biased downward such that, on average, correlation estimates will be lower than their true value (Olkin and Pratt, 1958). In Figure 5, we can see that the Anderson estimator of ρ always decreases the bias in estimating ρ, relative to an N_b of 0 where it is equivalent to the Pearson r. This means that the increase in mean squared error in the Anderson estimate for lower values of ρ and N_{nb} is due to the behavioral data adding variance to the estimate. It may be that the adjustment shown in Figure 3 is only likely to be in the right direction when there is enough data to estimate ρ and σ_b precisely. We hope that future statistical work will address this question and potentially develop new estimators that can control this variance better in the low sample size, low correlation regime. For now, the choice of whether to use the Anderson (1957) estimator will depend on the research question.

One somewhat counter-intuitive consequence of the above discussion is that Anderson's estimator can help not only in studies of individual differences across subjects but also in studies using representational similarity analyses where the matrix of item-item similarities based on a model is correlated with the similarities based on the multivariate neural signal (Kriegeskorte, 2008; Edelman et al., 1998). If researchers simply increase the number of similarities computed for the model, the above analyses suggest that there may be cases where this increases the power to detect a correlation with neural measures. In practice, since representational similarity matrices are often created from behavioral ratings (e.g Chikazoe et al., 2014; Bruffaerts et al., 2013), this is yet another way in which collecting additional behavioral data can increase power. While it is common in this literature to use Spearman's rank correlation to assess similarities, we

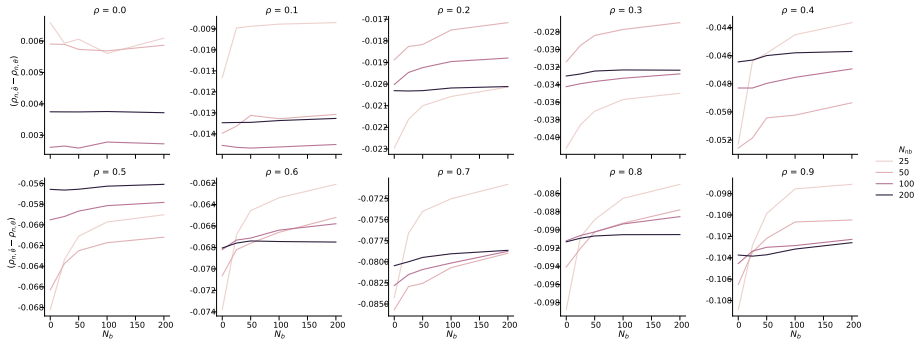

Figure 5: Bias in the Anderson (1957) estimate of ρ, over 10,000 simulated datasets, as a function of as a function of the true correlation, the number of neuroimaging subjects N_{nb} and number of additional behavioral subjects N_b.

can put this within the linear correlation framework by first ranking the two variables and then computing the correlation. For Pearson's r, this is exactly equivalent to computing the Spearman correlation.

3.2 How to Decrease Estimation Error: Improving Within-Subect Model Estimates by Collecting Data from Additional Subjects

We now investigate how collecting additional behavioral data can also help reduce estimation error, making it applicable to the regression analyses described above. In our previous description of estimation error, we were agnostic as to how exactly parameters were estimated from behavioral data. The most common method of estimating cognitive model parameters for use in a **model-based regression** is to use maximum likelihood estimation (Myung, 2003). The maximum likelihood estimate of parameters θ_i for subject i is the set of parameters in a cognitive model C that maximizes the probability of that subject's behavioral data $\mathbf{b}_i$ i.e.

$$\hat{\theta}_i = \text{argmax}_{\theta_i} P(\mathbf{b}_{it}|C, \theta_i, \mathbf{c}_{it}) \tag{20}$$

One problem is that many neuroimaging experiments are relatively short so not many trials can be collected per subject. In complex cognitive models, when the number of data points is small, many sets of parameters can be consistent with the data and, therefore, two datasets generated from the cognitive model with the same θ_i could result in very different maximum likelihood estimates $\hat{\theta}_i$, i.e. $\rho(\hat{\theta}_i, \theta_i)$ can be low on average. As demonstrated in the above estimation error section, this will result in lower power for estimating correlations and regression slopes when attempting to relate such variables to neural recordings.

Two methods have been used in the literature to increase $\rho(\hat{\theta}_i, \theta_i)$ and $\rho(\hat{\mathbf{c}_{it}}, \mathbf{c}_{it})$: sharing information across subjects or constraining the possible pa-

rameter values. One way to use information from other subjects data is to simply assume that all subjects have the same θ values which means that they can be estimated using all of the behavioral data collected in the experiment (Daw, 2011). This increases the data used to fit the parameters by a factor of n. The $\mathbf{c}_{it}$ will still vary by subject because of the dependence on the stimuli in the experiment. If the variability in true parameters θ_i across subjects is small relative to the estimation error, this can be an effective, heuristic way to increase $\rho(\hat{\mathbf{c}_{it}}, \mathbf{c}_{it})$. However, this is not an option in a **correlation** analysis when we are often interested the variation across subjects itself. In addition, many cognitive processes do have large individual differences so this may actually decrease $\rho(\hat{\mathbf{c}_{it}}, \mathbf{c}_{it})$.

Another option is to constrain parameter values by adding a prior, $P(\theta_i|\phi)$. If ϕ is chosen correctly, using a prior can decrease the likelihood of implausible parameter values such as a learning rate of 0 (Daw, 2011). We can then find the parameters that maximize the posterior, i.e.

$$\hat{\theta}_i = \text{argmax}_{\theta_i} P(\mathbf{b}_{it}|C, \theta_i, \mathbf{c}_{it}) P(\theta_i|\phi) \tag{21}$$

This is known as the maximum a posteriori (or MAP) estimation. By using the prior, we can force θ_i to be closer to more plausible values. However, it can be difficult to know in general if your priors are constraining the parameters to be in the correct space. If your priors are wrong, this could also decrease $\rho(\hat{\theta}_i, \theta_i)$.

One principled way to incorporate information from other subjects and use priors is hierarchical modeling (Gelman, 2006). This is an increasingly popular method in cognitive science and cognitive neuroscience (Ahn et al., 2011; Rouder and Lu, 2005) that assumes that there is a potentially nonuniform distribution $P(\theta_i|\phi)$ of parameters in the population. Functionally, this constrains the possible values of θ_i as in the prior above but because this is describing the population distribution, we can now estimate ϕ from all of the subjects data rather than setting ϕ by hand. The resulting estimate of θ_i will be a weighted average of the estimate from the population with the estimate that just uses an individual's data. The weights are determined by the strength of the individual data. For a subject whose data provides less constraint on the parameters of the cognitive model (i.e. if they had fewer trials or made more inconsistent choices), their estimates will be closer to the prediction from the prior. If a subject made very consistent choices, we might already have a lot of information about the parameters of their cognitive processes and we do not need to use as much information from the population. Thus, hierarchical modeling provides a way to decide from the data how much we should pool information from the population in creating each individual estimate. Katahira (2016) demonstrates that estimates from a hierarchical model strictly dominate both individual maximum likelihood estimates and population maximum likelihood estimates in terms of bias and variance.

While there exist frequentist methods for estimating hierarchical models (such as restricted maximum likelihood as implemented in popular packages for fitting hierarchical linear models like `lmer` (Bates et al., 2015)), for arbitrary

cognitive models, fitting in this way often requires a complicated derivation or numerical integration which can be challenging in high dimensions. With modern Bayesian inference tools like `Stan` (Carpenter et al., 2017), it is usually much more straightforward to place a prior on ϕ and use Bayesian inference to compute a joint posterior over ϕ and θ. The posterior for the hierarchical model is then

$$P(\theta, \phi|\mathbf{b}) \propto P(\mathbf{b}|C, \theta, \mathbf{c})P(\theta_i|\phi)P(\phi) \tag{22}$$

In addition to Stan, many R or python packages have made hierarchical Bayesian versions of specific cognitive models particularly easy to fit (e.g. (`hddm` (Wiecki et al., 2013) for drift diffusion models or `hBayesDM` (Ahn et al., 2017)) for many types of decision making models).

One feature of hierarchical modeling is that individual hierarchical estimates can often be improved simply by sampling more data from the population. That is, the cognitive model parameter estimates for subjects whose data you already have, i.e. subjects who contributed neuroimaging data, can be improved by adding data from new subjects. This is because the error of the estimates of ϕ (or the standard deviation of the prior) will, for most models, converge towards the correct estimates with increasing the number of subjects.

To demonstrate this, we conduct a simulation with a very simple hierarchical model to show that additional behavioral data from new subjects can improve inference for latent cognitive parameters θ_i for subjects $i \in [1, (N_{nb} + N_b)]$, N_{nb} of whom were collected in an original set neuroimaging experiment and N_b who were collected in a separate behavioral experiment. We assume each subject's latent cognitive parameter θ_i is drawn from a Gaussian population distribution with mean $\mu_\theta = 0$ and variance $\sigma_\theta^2 = 1$. Each subject then provides behavioral data b_{it} and to keep things simple, we assume that the cognitive state variables c_{it} are essentially equal to b_{it} such that

$$b_{it} \sim N(\theta_i, \sigma_{b_i}) \tag{23}$$

where $\sigma_{\mathbf{b}i}$ is drawn from a uniform distribution between .5 and 1 for each subject. Further simplifying, we summarise b_{it} by its mean $\bar{b}_i$ and assume that there are enough trials t that we can treat σ_{b_i} as known.

For the hierarchical model, we need to put a prior on σ_{θ_i}. While in practice it is often better to use a weakly or strongly informative prior, for generality, in these simulations, we assume an improper uniform prior, i.e. $P(\sigma_1) \propto 1$. Using the derivation in (Gelman et al., 2013), we construct a grid approximation to the posterior for θ given $\mathbf{b}$. We now investigate how the correlation between the true subject parameters θ_i and their posterior mean estimate $\hat{\theta}_i$ for the N_{nb} neuroimaging subjects changes as we add a number of new behavior-only subjects, N_b, to the experiment. Note that we are only computing the correlation within the original N_{nb} subjects. As we demonstrated in the section on estimation error, this correlation is what limits the power to detect true correlations or regression slopes between model estimates and neural signals.

Figure 6 shows that across a range of N_{nb} values, the addition of N_b new behavioral subjects can increase the correlation between model estimates for

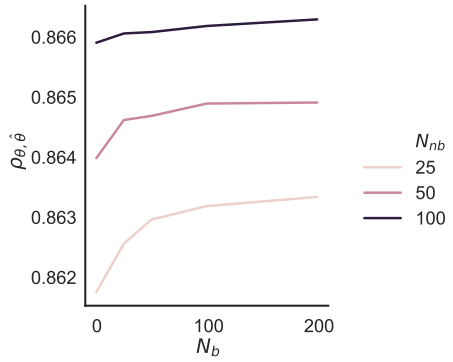

Figure 6: Correlation between the estimated $\hat{\theta}_{[1:N_{nb}]}$ and $\theta_{[1:N_{nb}]}$ from the hierarchical model, averaged over 10,000 simulated datasets, as a function of the number of neuroimaging subjects N_{nb} and number of additional behavioral subjects N_b.

the N_{nb} neuroimaging subjects $\hat{\theta}_{[1:N_{nb}]}$ and true parameters $\theta_{[1:N_{nb}]}$.

This implies that one way to increase $\rho(\hat{\theta}_i, \theta_i)$ (and therefore also increase $\rho(\hat{\mathbf{c}_{it}}, c_{it})$) for subjects in a neuroimaging experiment is to simply collect behavioral data from more subjects. However, the exact rate at which more subjects improve the model estimates will in general depend on the distribution of uncertainty in the individual θ_i as well as the particular form of the model you use.

Beyond it's use in improving hierarchical estimates of cognitive model parameters, collecting additional behavioral data can improve neuroimaging regression inferences in several other ways. While not computationally feasible yet for most neuroimaging analyses, joint modeling (Turner et al., 2013) is a recently proposed technique in which the cognitive model and the neuroimaging model are jointly fit, hierarchically, to both the neuroimgaging and behavioral data. This can allow for even more efficient use of information by allowing both sources of data to inform the cognitive model parameters. In addition, it allows the uncertainty in the cognitive model parameters to be propagated into the estimates of the regression or correlation, resulting in more powerful estimates without the biases mentioned above. Turner et al. (2016) demonstrated how joint modeling can be used to improve multimodal inferences even when we don't have data for every mode for every row in the dataset (i.e. some subjects' data were collected with EEG and others with fMRI). This can be generalized to the case we discuss where we only have rows in the dataset with only one type of data (e.g. behavior).

In addition, to improving estimates within a particular model, collecting additional behavioral data can be useful in many other ways for improving model estimates. For instance, collecting additional behavioral data in experimental conditions or on novel stimuli not in the neuroimaging experiment could assist

in model selection. Choosing the correct model could potentially have a much larger impact on the correlation between model estimates and true parameters.

Finally, we can view reducing estimation error as relevant not only for model-based regression but also for regression approaches where the stimulus features are subjective or latent, such as the valence of a stimulus. The covariates in this case are often created from ratings based either on subjects in the experiment or outside behavioral subjects. Similar to their use with cognitive models, hierarchical models with additional behavioral data can reduce error in these stimulus ratings as well. Even just aggregating the average rating however, will be improved through greater amounts of data per item. Indeed, going back to at least Kent and Rosanoff (1910), it is already quite common in cases like this to use large "norming" studies to understand stimuli, particularly for word stimuli.

3.2.1 When does it work?

Having demonstrated that it is possible to reduce estimation error in a neuroimaging dataset by collecting more behavioral data, we might now wonder when this is likely to help. As we mentioned above, the amount that hierarchical modeling can help will in general depend on the model and the experimental design. In addition, collecting additional behavioral data can improve estimates in other ways such as through model selection. In order to say something general about how improvements in estimation improve statistical power to find neuroimaging effects, we will simply assume that it is possible to increase the correlation between model estimates and true parameters using behavioral data by a certain amount and investigate how that changes estimates about the relationship with neural data. To begin, we will build on the derivation from Wilson and Niv (2015) where they showed how the least squares estimate of the regression slope is impacted by misestimation (i.e., equations 4 and 5). Wilson and Niv (2015) only solved this for the case where the analyst is trying to infer a single regression slope from the data (i.e. a non-hierarchical regression model). But in model-based neuroimaging analyses, it is rarely assumed that all subjects have the same effect size. Therefore, the most common analysis method in model-based analysis is to use a two stage approach where a first-stage regression is fit to each subject (or each run) and then the subject regressions are aggregated at the group-level by approximating a hierarchical model, testing whether the average effect is significantly different from zero (Beckmann et al., 2003; Woolrich et al., 2004). This means that the benefit of collecting an additional neuroimaging subject cannot be captured by the Wilson and Niv (2015) equations alone. Following Friston et al. (2002) and Beckmann et al. (2003), we assume a Gaussian hierarchical model of population effects:

$$\beta_{c_i} = \beta_G + \eta_i \tag{24}$$

$$n_{it} = \beta_{c_i} c_{it} + \epsilon_{it} \tag{25}$$

where the noise distributions are both Guassian:

$$\eta_i \sim N(0, \sigma_G^2) \tag{26}$$

$$\epsilon_{it} \sim N(0, \sigma_\epsilon^2) \tag{27}$$

Using the same notation as above, for subject i on trial t, c_{it} is that subject's cognitive state at that time and n_{it} is the associated neural signal. β_{c_i} represents the relationship between the cognitive state and the neural signal for a particular subject and β represents that average relationship in the population. σ_G^2 represents the variance of the magnitude of the relationship in the population and σ_ϵ^2 represents the additional variance in the neural signal that is not due to the cognitive state (either due to noise in the measurement or the signal being related to other unrelated cognitive processes as well.)

From Wilson and Niv (2015), we know the distribution of estimates in the first level regression, i.e.

$$\hat{\beta}_{c_i} \sim N\left(\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})\beta_{c_i}, \frac{\sigma_\epsilon^2 + \beta_{c_i}^2(1 - \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2)}{T-2}\right) \tag{28}$$

Given the population distributions above and using standard formulas for linear combinations of Gaussian random variables, we can get

$$\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})\beta_{c_i} \sim N\left(\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})\beta_{c_i}, \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2\sigma_G^2\right) \tag{29}$$

We can now write the distribution of first level estimates as

$$\hat{\beta}_{c_i} \sim N\left(\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})\beta_G, \frac{\sigma_\epsilon^2 + \beta_{c_i}^2(1 - \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2)}{T-2} + \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2\sigma_G^2\right) \tag{30}$$

The parameter we are interested in is β_G, the average effect in the population. While there are more efficient but computationally expensive estimators (Beckmann et al., 2003; Woolrich et al., 2004), the most common way to estimate β_G is the OLS method, i.e.

$$\hat{\beta}_G = \frac{\sum_{i=1}^{N_{nb}} \hat{\beta}_{c_i}}{N_{nb}} \tag{31}$$

$\hat{\beta}_G$ is then just the mean of independent Gaussian random variables and will therefore have a Gaussian distribution with mean $\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})\beta_G$ and variance

$$\sigma_{\hat{\beta}_G}^2 = \left(\frac{1}{N_{nb}}\right)^2 \sum_{i=1}^{N_{nb}} \left(\frac{\sigma_\epsilon^2 + \beta_{c_i}^2(1 - \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2)}{T-2} + \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2\sigma_G^2\right) \tag{32}$$

$$= \left(\frac{1}{N_{nb}}\right)^2 \left(N_{nb}\left(\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2\sigma_G^2 + \frac{\sigma_\epsilon^2}{T-2}\right) + \frac{(1 - \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})^2)}{T-2}\sigma_G^2 \sum_{i=1}^{N_{nb}} \frac{\beta_{c_i}^2}{\sigma_G^2}\right) \tag{33}$$

The first term is a constant. In the second term, because β_c is a Gaussian random variable, the normalized squared sum has a non-central chi-squared distribution,

$$\sum_{i=1}^{N_{nb}} \frac{\beta_{c_i}^2}{\sigma_G^2} \sim \chi^2(N_{nb}, \frac{\beta_G^2}{\sigma_G^2}) \tag{34}$$

and a non-central chi-squared multiplied by a constant will have a generalized chi-squared distribution. Plugging in the mean of that distribution, the expectation of the variance is now:

$$E(\sigma_{\hat{\beta}_G}^2) = \left(\frac{1}{N_{nb}}\right)^2 \left(N_{nb}\left(\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2\sigma_G^2 + \frac{\sigma_\epsilon^2}{T-2}\right) + \frac{(1-\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2)}{T-2}\sigma_G^2\left(N_{nb} + N_{nb}\frac{\beta_G^2}{\sigma_G^2}\right)\right) \tag{35}$$

$$= \frac{1}{N_{nb}(T-2)}(\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2\sigma_G^2(T-2) + \sigma_\epsilon^2 + (1-\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2)(\sigma_G^2 + \beta_G^2)) \tag{36}$$

Combining this with the expected estimate, we can get an t-statistic at the expected variance

$$\hat{t}(\rho(\mathbf{c_i}, \hat{\mathbf{c_i}}), \beta_G, N_{nb}, T, \sigma_G^2, \sigma_\epsilon^2) = \rho(\mathbf{c_i}, \hat{\mathbf{c_i}})\beta_G\sqrt{\frac{N_{nb}(T-2)}{\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2\sigma_G^2(T-2) + \sigma_\epsilon^2 + (1-\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2)(\sigma_G^2 + \beta_G^2)}} \tag{37}$$

Rather than parameterizing this in terms of the raw effect size, β_G, it can be useful for understanding to standardize the variables by the total standard deviation, $\sqrt{\sigma_\epsilon^2 + \sigma_G^2}$, similar to the derivation in Wilson and Niv (2015). For clarity, we define two new variables. Following Wilson and Niv (2015), we call the standardized effect size, $\frac{\beta_G}{\sqrt{\sigma_\epsilon^2+\sigma_G^2}}$, the contrast-to-noise ratio (CNR). In multilevel modeling, an important statistic is the intra-class correlation coefficient or ICC which is the proportion of the total variance that is explained by the subject level variance, i.e. $\frac{\sigma_G^2}{\sigma_G^2+\sigma_\epsilon^2}$ (Chen et al., 2018; Shrout and Fleiss, 1979). If we multiply the right hand side of the above equation by $\frac{\sqrt{\sigma_\epsilon^2+\sigma_G^2}}{\sqrt{\sigma_\epsilon^2+\sigma_G^2}}$, we can now write the equation in terms of these two variables, i.e.,

$$\hat{t}(\rho(\mathbf{c_i}, \hat{\mathbf{c_i}}), \mathrm{CNR}, \mathrm{ICC}, N_{nb}, T) = \rho(\mathbf{c_i}, \hat{\mathbf{c_i}})\mathrm{CNR}\sqrt{\frac{N_{nb}(T-2)}{\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2\mathrm{ICC}(T-2) + (1-\mathrm{ICC}) + (1-\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})^2)(\mathrm{ICC} + \mathrm{CNR}^2)}} \tag{38}$$

Having derived the expected t-statistic, we can now compute the power under that t-statistic using standard formulas. This allows us to ask the question "how much would additional behavioral data have to improve our latent variable

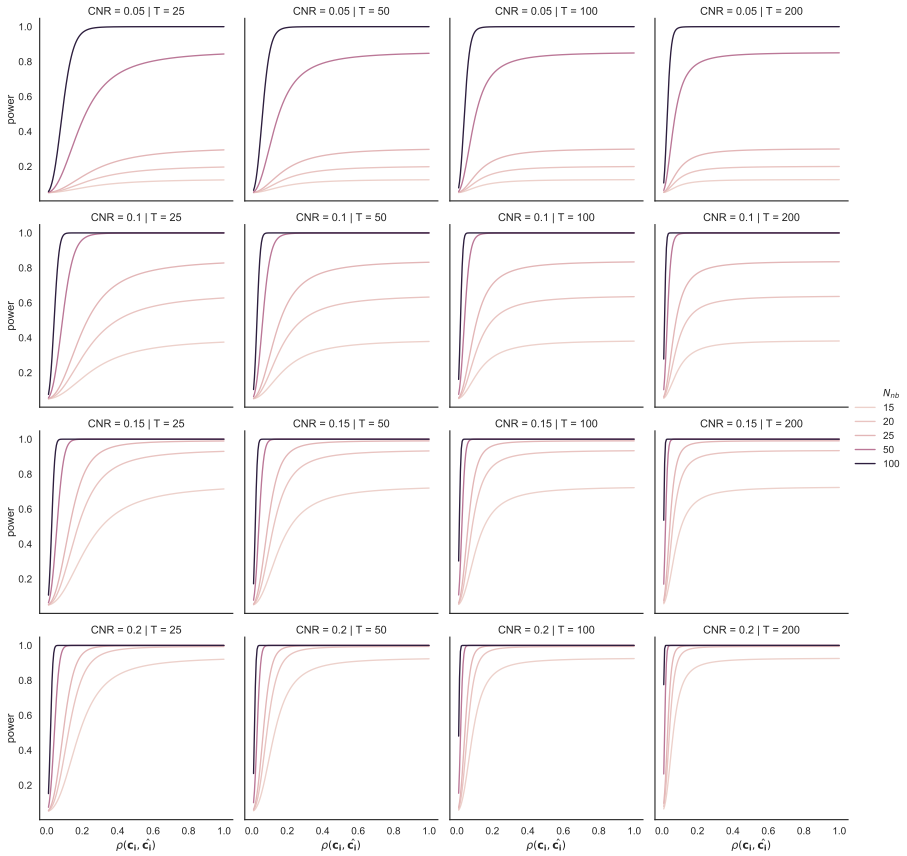

Figure 7: Power as a function of $\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})$ for a range of CNR, T and N_{nb} values. ICC is set to $\frac{2}{3}$ corresponding to a between subjects variance that is twice as large as the neuroimaging noise. In general, power increases as the number of subjects increases, as the number of trials increases and as accuracy of the model estimates (the inverse of the estimation error) increases.

estimates in order to improve power as much as another neuroimaging subject?" That is, what is the increase in $\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})$ that is equivalent to increasing N_{nb} by one? More formally, if $P(t, N)$ is the power of the one-sample t-test with a t-stat of t and N subjects, we can write

$$P(\hat{t}(\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i}) + \delta_\rho, \text{CNR}, \text{ICC}, N_{nb}, T), N_{nb}) = \\ P(\hat{t}(\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i}), \text{CNR}, \text{ICC}, N_{nb} + 1, T), N_{nb} + 1) \quad (39)$$

and ask what value of δ_ρ makes this equation true?

Since δ_ρ must be between 0 and $1 - \rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})$, this is a straightforward constrained root finding problem that we can solve exactly using numerical methods. We can solve this for several values of the other parameters (CNR, ICC, T and N_{nb}) using Brent's method (Brent, 1972) as implemented in the `scipy` package (SciPy 1.0 Contributors et al., 2020) in `python 3.7.7`. We plot the power as a function of CNR, T and N_{nb} values in Figure 7.

Which settings of the parameters in equation 39 are reasonable for neuroimaging experiments? Wilson and Niv (2015) derive estimates for CNR from two different studies with a very wide range of .4 to 11. Their CNR definition was only normalized by σ_ϵ and not the total standard deviation including the group variance as we have done here. Therefore, to translate those values to values in equation, we need to take into account the ICC. In for this simulation, we set the ICC to $\frac{2}{3}$, corresponding to a between subjects variance that is twice as large as the neuroimaging noise, which seems to be quite large. With that value of ICC, Wilson and Niv (2015)'s CNR of .4 corresponds to a CNR of $\frac{.4}{\sqrt{3}}$ which is approximately .231. We can see in Figure 7 that a value of .2 for the CNR leads to somewhat implausible power estimates for most cognitive neuroscience task-based studies (Button et al., 2013; Turner et al., 2018). Therefore, we will assume that the CNR is closer to .1 in the following simulation. However, with the above derivation, it is easy to compute this for other values of the generative model and task parameters. Also of note in these plots is that for a particular value of N_{nb} the power function will asymptote as a function of $\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})$. This means that given a good enough model estimate, it is impossible to improve power more by improving the estimate than by collecting more subjects.

Figure 8 shows the solution to equation 39 in terms of δ_ρ as a function of the correlation between estimate and true parameters with one more neuroimaging subject. Overall, as noted above, there is a range of parameter settings where improving model fit (potentially by collecting more behavioral subjects) cannot help improve the power more than collecting another neuroimaging subject. Wilson and Niv (2015) showed that, for a reinforcement learning model, it is fairly easy to obtain a high $\rho(\mathbf{c_i}, \hat{\mathbf{c}}_\mathbf{i})$, even if the latent subject-level parameters are severely misestimated. This suggests that in that setting, model-fitting "isn't necessary", or at least improving the model fit may be unlikely to significantly affect power. However, this may not be the case for other cognitive models, although it seems unlikely that a model that is fit to the data will have a correlation as low as .2 or .3.

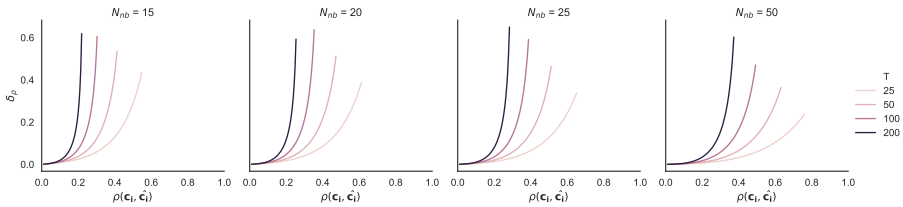

Figure 8: Solution to equation 39 as a function of parameters $\rho(\mathbf{c_i}, \hat{\mathbf{c_i}})$ for a range of T and N_{nb}. CNR is set to .1 and ICC is set to $\frac{2}{3}$. Lines end when it is no longer possible to achieve better performance by increasing the correlation.

While in this paper we have assumed that data collected in the scanner and out of the scanner (or with and without an EEG cap) are exchangeable, that is unlikely to be exactly true in practice. This does not make the described methods impossible to use but does potentially require more complex models that allow for differences between the data collected under different modalities. That being said, there is relatively little work documenting the magnitude of these purported differences. If these differences were found to be large, this would likely have implications for learning about cognition through neuroimaging in general, not just with the statistical methods described here. We strongly encourage both future empirical work documenting these differences and future statistical work on how to aggregate inferences across data collection modalities.

In future work, we hope to explore this in the context of other cognitive models. Assuming that correlations between are reasonably high and that a smaller δ_ρ is easier to achieve, Figure 8 shows that improving the model fit with additional behavioral data is most useful when there is already a reasonably sized N_{nb} and T is smaller. While even 50 subjects is already significantly above average for a neuroimaging study, most use numbers of trials that are much higher than 50. One thing that this suggests is that using extra behavioral data may be particularly useful for neuroimaging studies with small numbers of trials. These are not common because of the low power they achieve but methods like those mentioned in the previous section may make them more feasible. This opens the door to more neuroimaging studies that are difficult to run with high numbers of trials such as long-term memory studies where subjects are unlikely to be able to remember a large number of items over a long delay.

4 Discussion

In this paper, we have presented two ways in which collecting additional behavioral data outside of a neuroimaging experiment has the potential to improve the precision of an estimate of the relationship between brain measures and cognition. We also attempted to quantify a tradeoff between collecting each type of data, showing how this could potentially affect the design of future studies

or reanalyses of past work. In both cases, behavioral data had the potential to help but not in all cases. Unfortunately, it was particularly unhelpful when both true effect sizes and the number of subjects were small, which is exactly the case when more power is needed. However, our analyses pointed to possible future statistical work, in particular the potential to improve the Anderson (1957) correlation estimate by appropriately reducing the variance. In addition, our analyses suggest new regimes where behavioral data could be helpful; for instance, the potential for improving model fits with behavioral data could allow for feasibly running neuroimaging studies with small numbers of trials. In many cases, such as in long term memory studies, it is not feasible to run large numbers of trials because participants cannot reasonably remember hundreds of items. Both of these directions will require more work to be put into practice but we hope this paper will inspire collaborations between cognitive neuroscientists and statisticians to work more on ways of fusing data sources in order to learn about brain and behavior correlations.

Finally, while we have mostly discussed collecting additional behavioral data in the context of designing a new study, it could also be a consideration in a study that resuses an old dataset. It could be that a model is not identifiable with a particular dataset but it could become identifiable by collecting behavioral data in more conditions. With the advent of more open neuroimaging data sets, and particularly with the current difficulty in collecting new data, these methods point to a way to learn new things from old datasets simply by collecting more cheap behavioral data.

References

Ahn, W.-Y., Haines, N., and Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, 1:24–57.

Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R., and Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics*, 4(2):95–110.

Anderson, T. W. (1957). Maximum Likelihood Estimates for a Multivariate Normal Distribution when Some Observations are Missing. *Journal of the American Statistical Association*, 52(278):200–203.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, 67(1).

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, 20(2):1052–1063.

Brent, R. P. (1972). *Algorithms for minimization without derivatives*. Prentice-Hall series in automatic computation. Prentice-Hall, Englewood Cliffs, N.J.

Bruffaerts, R., Dupont, P., Peeters, R., De Deyne, S., Storms, G., and Vandenberghe, R. (2013). Similarity of fMRI Activity Patterns in Left Perirhinal Cortex Reflects Semantic Similarity between Words. *Journal of Neuroscience*, 33(47):18597–18607.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). *Stan* : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1).

Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., Leibenluft, E., Brotman, M. A., and Cox, R. W. (2018). Intraclass correlation: Improved modeling approaches and applications for neuroimaging. *Human Brain Mapping*, 39(3):1187–1206.

Chikazoe, J., Lee, D. H., Kriegeskorte, N., and Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, 17(8):1114–1122.

Cremers, H. R., Wager, T. D., and Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLOS ONE*, 12(11):e0184923.

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, 8(3):e57410.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models: (Tutorial Review). In *Decision Making, Affect, and Learning*. Oxford University Press, Oxford.

Durnez, J., Blair, R., and Poldrack, R. A. (2017). Neurodesign: Optimal Experimental Designs for Task fMRI. preprint, Neuroscience.

Edelman, S., Grill-Spector, K., Kushnir, T., and Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, 26(4):309–321.

Feinberg, D. A. and Yacoub, E. (2012). The rapid development of high speed, resolution and precision in fMRI. *NeuroImage*, 62(2):720–725.

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10(4):507.

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). Classical and Bayesian Inference in Neuroimaging: Theory. *NeuroImage*, 16(2):465–483.

Garren, S. T. (1998). Maximum likelihood estimation of the correlation coefficient in a bivariate normal model with missing data. *Statistics & Probability Letters*, 38(3):281–288.

Gelman, A. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics*, 48(3):432–435.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 0 edition.

Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14, Montreal QC, Canada. ACM Press.

Haxby, J. V. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539):2425–2430.

Hocking, R. R. and Smith, W. B. (1972). Optimum Incomplete Multinormal Samples. *Technometrics*, 14(2):299–307.

Homan, P., Levy, I., Feltham, E., Gordon, C., Hu, J., Li, J., Pietrzak, R. H., Southwick, S., Krystal, J. H., Harpaz-Rotem, I., and Schiller, D. (2019). Neural computations of threat in the aftermath of combat trauma. *Nature Neuroscience*, 22(3):470–476.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11):4302–4311.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, 73:37–58.

Kent, G. H. and Rosanoff, A. J. (1910). Part II. Association in insane subjects. In *A study of association in insanity.*, pages 16–72. American Journal of Insanity.

Kragel, J. E. and Polyn, S. M. (2016). Decoding Episodic Retrieval Processes: Frontoparietal and Medial Temporal Lobe Contributions to Free Recall. *Journal of Cognitive Neuroscience*, 28(1):125–139.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, 93(3):480–490.

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.

Lindquist, M. A., Meng Loh, J., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mismodeling. *NeuroImage*, 45(1):S187–S198.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical analysis with missing data*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, third edition edition.

Lombardo, M. V., Auyeung, B., Holt, R. J., Waldman, J., Ruigrok, A. N., Mooney, N., Bullmore, E. T., Baron-Cohen, S., and Kundu, P. (2016). Improving effect size estimation and statistical power with multi-echo fMRI and its impact on understanding the neural systems supporting mentalizing. *NeuroImage*, 142:55–66.

Mack, M., Preston, A., and Love, B. (2013). Decoding the Brain's Algorithm for Categorization from Its Neural Implementation. *Current Biology*, 23(20):2023–2027.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. pages 56–61, Austin, Texas.

Munafò, M. R., Cremers, H. R., Wager, T. D., and Yarkoni, T. (2019). Power and design considerations in imaging research. In *Casting Light on the Dark Side of Brain Imaging*, pages 73–78. Elsevier.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, 38(2):329–337.

Olkin, I. and Pratt, J. W. (1958). Unbiased Estimation of Certain Correlation Coefficients. *The Annals of Mathematical Statistics*, 29(1):201–211.

Palmeri, T. J., Love, B. C., and Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76:59–64.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126.

Polyn, S. M. (2005). Category-Specific Cortical Activity Precedes Retrieval During Memory Search. *Science*, 310(5756):1963–1966.

Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., and Chun, M. M. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*, 19(1):165–171.

Rouder, J. N. and Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12(4):573–604.

SciPy 1.0 Contributors, Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, , Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Treadway, M. T., Buckholtz, J. W., and Zald, D. H. (2013). Perceived stress predicts altered reward and loss feedback processing in medial prefrontal cortex. *Frontiers in Human Neuroscience*, 7.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., and Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206.

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., and Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, 128:96–115.

Turner, B. O., Paul, E. J., Miller, M. B., and Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, 1(1):62.

van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30.

Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., Ruiter, J. D., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Swain, C., Miles, A., Brunner, T., O'Kane, D., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., and Brian (2020). mwaskom/seaborn: v0.10.1 (April 2020).

Wiecki, T. V., Sofer, I., and Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, 7.

Wilson, R. C. and Niv, Y. (2015). Is Model Fitting Necessary for Model-Based fMRI? *PLOS Computational Biology*, 11(6):e1004237.

Wilson-Mendenhall, C. D., Barrett, L. F., and Barsalou, L. W. (2015). Variety in emotional life: within-category typicality of emotional experiences is associated with neural activity in large-scale brain networks. *Social Cognitive and Affective Neuroscience*, 10(1):62–71.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4):1732–1747.

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, 4(3):294–298.