Experimental Operationalization and Inferential Conclusions in fMRI Investigation

Submitted to the University of Alabama at Birmingham, Undergraduate Neuroscience Program in partial fulfillment of the requirements for an honors degree in Neuroscience.

December 6th, 2019

<u>Written by:</u>

Samuel J Castillo

<u>Mentor:</u>

Dan Mirman, Ph.D.

Department of Psychology

University of Alabama at Birmingham

Table of Contents

Chapter **Page**

Supplement **Page**

Introduction: A Brief Overview of Neuroscientific Methodology

The human brain is arguably the most complex structure in the known universe. The functional mechanics of the three-pound cluster of neural tissue that drives the human experience are beyond our ability to grasp entirely—yet, this very complexity demands the utmost degree of curiosity. Our efforts to explore the brain and nervous system have historically taken the form of philosophy, theology, psychology, biochemistry, and everything in between. Ultimately, our studies have driven us to the foundation of an independent yet highly interdisciplinary field dedicated solely to understanding the structural and functional mechanics of the nervous system: neuroscience.

As a discipline, neuroscience is unique in its multifaceted breadth of methodology—methods in neuroscientific research range from the molecular to the mass population scale. Yet, nearly all approaches aim to explore the components of the nervous system and how they relate to one another to build complex structures such as the human brain. One significant facet of this goal is functional localization, i.e., the association of cognitive processes and functions with a specific area or network of areas within the brain.

In the early years of neuroscience, functional localization studies were possible only on a case-by-case basis. If we were to learn something about a certain region of the brain, we were likely to learn it simply by observing the effects of an injury to that region. Perhaps the most notorious case-study of functional localization is that of Phineas Gage, the mid-nineteenth century railroad worker whose personality suffered a severe shock after a traumatic injury to the frontal cortex (Harlow, 1999). Gage's transformation was fascinating, especially since his cognitive faculties somehow remained largely intact—the only apparent serious change was that of his temper and overall perspective of life (O'Driscoll & Leach, 1998). This posed several issues with the understanding of mind and brain at the time, but it eventually sparked a driven curiosity to discover what areas of the brain are responsible for the variety of cognitive processes we collectively refer to as the human experience.

Over the last half-century or so, modern medicine and research have promoted the value of localizing and describing human neurological function in a manner that basic methods or case studies do not fulfill. We cannot simply wait in curious expectation for cases like Phineas Gage to arrive at our doorstep, especially if we aim to ultimately characterize the entire brain. Thus, investigators have extended our breadth of discovery to higher degrees of complexity via the most recent and lasting method of neuroscientific analysis: neuroimaging.

Neuroimaging describes a class of methods used to generate meaningful images of an individual's brain structure and/or function. These methods range from minimally invasive to noninvasive techniques that allow varying degrees of spatial and temporal resolution. One of the earliest developed methods of image acquisition and analysis is magnetic resonance imaging (MRI). MRI developed as a neuroimaging method in 1977, and it has since become the basis for countless investigations of brain structure and function (Glover, 2011).

MRI: An Overview

As implied by the name, MRI involves the use of strong magnetic fields to induce resonance of protons within the tissue. In this case, resonance is a proton's tendency to shift its orientation when it encounters a certain frequency.

Through a series of radio-frequency pulses, the scanner retrieves information about the position and orientation of different protons responding to the magnetic fields. Each proton will respond to the magnetic fields differently based on the properties of the tissue surrounding it. The scanner obtains a signal that, through complex mathematical processing, reveals the location and type of tissues subjected to the magnetic fields (Huettel, Song, & McCarthy, 2014). For instance, MRI analysis can reveal the spatial divide between gray matter and white matter within the brain, allowing a detailed look at a participant's neural structure. Since the same principles of resonance apply to all tissues, MRI can also localize blood—specifically, hemoglobin proteins within blood. The scanner's capacity to detect blood and blood flow provides the basis for functional magnetic resonance imaging, or fMRI (Glover, 2011).

The "functional" qualifier of fMRI refers to its ability to localize neurological functions according to blood flow. The underlying premise is that more active neurons invoke a higher metabolic rate than less active neurons, resulting in a greater need for oxygen. Oxygen is transported via hemoglobin proteins within the blood; the proteins arrive at tissues as oxyhemoglobin and exit tissues as deoxyhemoglobin (having released oxygen for metabolic use). fMRI involves the localization of deoxyhemoglobin over time, indicating which regions of the brain required greater oxygen input and thus were most active during a certain period. The signal obtained in fMRI, then, is dependent upon the levels of oxygenated versus deoxygenated blood. For this reason, investigators have coined this method of analysis the blood-oxygen level dependent signal, or BOLD signal (Ogawa, Lee, Kay, & Tank, 1990).

Since the BOLD signal is an indicator of tissue metabolism rather than synaptic communication, fMRI is a temporally indirect method of observing neural activity. Contrary to instantaneous changes in electrical activity recorded by other imaging methods, changes in blood oxygenation (i.e., the hemodynamic response) typically peak approximately 6 seconds after the onset of activity. The total duration of the hemodynamic response varies according to vascular diameter and capillary networks, both of which vary according to the physiological characteristics of the tissue. So, different areas of the brain demonstrate somewhat variable hemodynamic responses. In general, however, the response can be modeled by the hemodynamic response function, an example of which is shown in **Figure 1** (Lindquist, Meng Loh, Atlas, & Wager, 2009).

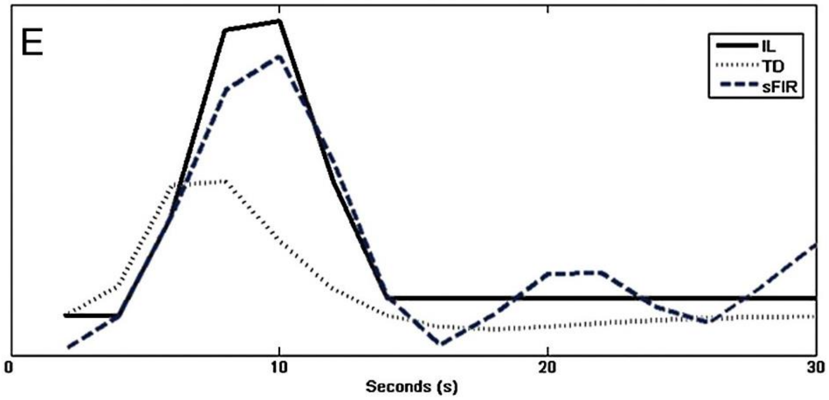

Figure 1. Potential Models of the Hemodynamic Response Function. *Demonstrates three models used to characterize the hemodynamic response: inverse logit (IL), temporal derivative (TD), and semiparametric smooth finite impulse response (sFIR).*

Thus, fMRI provides a relatively low temporal resolution compared to waveform-based counterparts such as electro- and magnetoencephalography (EEG and MEG, respectively). The mechanisms of EEG and MEG render these methods capable of capturing the direct effects of neuronal signaling, placing their temporal resolution on the order of milliseconds (compared to fMRI's order of seconds). However, these waveform-based methods demonstrate improved temporal resolution at the expense of spatial resolution, which typically spans a few centimeters—much too large for precise functional localization. fMRI, on the other hand, captures data with a spatial resolution as low as 3-4 mm, small enough to determine precise regions and subregions of neural activity. Ideally, then, functional neuroimaging is best accomplished via compound approaches involving both fMRI and EEG/MEG methods, allowing investigators to capitalize on the resolution strengths of each method. Unfortunately, logistic limitations (funding, operating space, lack of expertise, etc.) prevent most labs from conducting multifaceted imaging studies. Given these limitations that leave room for only one method, fMRI remains one of the foremost approaches for modern neuroimagers.

Functional Localization Via fMRI: Validity at Stake

Given modern developments in the technical and analytical aspects of neuroimaging, fMRI has become highly cut out to explore the depths of functional localization. We've reached a stage at which clean and publishable results are relatively easy to obtain by even amateur researchers, allowing the field of neuroimaging to advance much further than we anticipated in such a short time. Throughout this time of great development, though, I believe our standards of experimentation have failed to follow suit. The problem is not the limitations of fMRI methodology, as every investigative technique bears its own limitations. Rather, we as researchers have lost some sense of responsibility to implement tools like fMRI in a manner that respects these limitations. Our boundaries of interpretive inference have become convoluted due to a collective misunderstanding of fMRI as an operationalization.

As a response to the shortfalls of our field, the aim of this paper is to explore the possibilities and boundaries of fMRI operationalization as it relates to interpretative inference of functional localization paradigms. To supplement these ideas, I will provide brief overviews of the methodological background of fMRI research. Ultimately, I'll provide recommendations of how to realistically move forward according to recent studies and ideologies of neuroimaging.

Inference: The Fundamental Basis of Data Interpretation

The field of neuroimaging generally relies on two main forms of inference to draw functional conclusions from observed data: forward inference and reverse inference (Poldrack, 2011). The less common approach among neuroimagers is reverse inference. Yet, its breadth of discovery and application has invoked growing pursuit over the last several years. To draw a reverse inferential conclusion, one asks the following: "What is the probability of the presence of a certain cognitive state given a certain pattern of neural activity?" Essentially, reverse inference allows investigators to potentially determine someone's mental state simply by observing neural activity via some imaging method. The applications of this form of inference are quite vast, particularly for clinical psychologists who aim to understand the potential impairments in patients' cognitive states by analyzing their neural activity relative to healthy controls.

Reverse Inference: P (M|A)

Equation 1. Reverse Inference Probability Statement. *"M" indicates a specified mental state; "A" indicates observed neural activity.*

As expected, reverse inferential conclusions are difficult to establish due to the multi-modal capacities of many brain regions (Sprooten et al., 2017; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). For instance, participants subjected to a fearful stimulus generally exhibit activity in the amygdala, but activity in the amygdala may indicate a mental state of several different basic emotions (Geng et al., 2018; Goodwin & Norbury, 2016). So, it would be erroneous to infer that activation in the amygdala necessarily indicates that one must be experiencing fear. Despite this inherent limitation of reverse inference, its potential applications within cognitive neuroscience render it a tool worth exploring. Certain authors have informally dubbed reverse inference as modern day "mind-reading" due to its premises; however, the process of probabilistically determining someone's cognitive state via neural activation is more formally known as decoding (Poldrack, 2011; Yarkoni et al., 2011). Effective decoding requires strong inferential foundations, and it has become one of the mainstream foci of modern neuroimaging research.

Logically speaking, though, the validity of reverse inference relies on the conditional probability that a particular neural region is truly associated with a selective function—Poldrack pointed this out in his own overview of inference (Poldrack, 2011). In other words, we cannot infer the presence of a certain cognitive state in light of neural activity if we do not first gauge the probability of the activity pattern appearing in conjunction with that cognitive state. Obtaining this knowledge via fMRI analysis requires the experimental induction of a specific cognitive state followed by observation and analysis of the resulting neural activity, i.e., forward inference.

Forward inference is the more common inferential approach of fMRI analysis, as it is seemingly less vulnerable to obstacles of generalization than reverse inference. To summarize the definition of forward inference presented in the preceding paragraph, a forward inferential conclusion is drawn as such: “What is the probability of observing a certain pattern of neural activity given the presence of a certain cognitive state?”

$$\textit{Forward Inference: } P\ (A|M)$$

Equation 2. Forward Inference Probability Statement. *”M” indicates a specified mental state; “A” indicates observed neural activity.*

A common example of forward inference is the n-back task discussed in detail later in this paper (**Figure 4**). The paradigm is designed to invoke working memory, so any neural activity contrast deemed significant would be attributed to the cognitive processes involved in working memory. If these results are replicated several times over using this paradigm (and they have been), then it is established within the scientific community that the forward inferential conclusions regarding the neural correlates of working memory using the n-back task are valid. All the necessary components of scientific inquiry appear present: the formation and experimental analysis of a hypothesis, coherent and generalizable results, and replicable data to confirm the validity of the model. The integrity of fMRI analysis has rested on this forward inferential foundation for decades. However, this general scheme of inference-validation isn’t as foolproof as we would like for it to be.

Consider the central question of forward inference as presented in **Equation 2**. For the inferential conclusion to be valid, the neural activity observed in the experiment is ideally a representation of the activity required to maintain the neuropsychological function induced by the experimental paradigm. In order for this to occur, the theoretical perspectives of the verbal hypothesis must be entirely captured by the experimental operationalization. If the operationalization captures more or less than the verbal hypothesis, the reliability of the inference is at stake (Yarkoni, 2019).

Operationalization: The Fundamental Basis of Inference

Operationalization can most generally be defined as one's means of determining something (J. Sullivan, 2007, 2015; Yarkoni, 2019). In the context of fMRI investigation, experimental paradigm design operationalizes the verbal hypothesis (J. Sullivan, 2015), and statistical analysis operationalizes the paradigm (Poline & Brett, 2012; Yarkoni, 2019). Logically speaking, then, the results of any fMRI functional localization study are not necessarily the neural correlates of the psychological process in question—they are the neural correlates of operationalizations of those processes.

Statistical Analysis: An Operationalization of Experimental Paradigm

To understand the operationalization of paradigm via statistical analysis, we must first understand the nature of imaging data. fMRI data analysis is based on the structural map of the brain as it relates to the time-course of the BOLD signal. As the scanner collects data from a participant, the three-dimensional anatomical image of the brain is reconstructed into hundreds of thousands of tiny "voxels" (i.e., volume-pixels). Voxels typically span 1—8 mm^3 in volume, thus encompassing precise neural regions (Glover, 2011). The BOLD signal is recorded per individual voxel until the entire brain volume is accounted for. This process is repeated for each epoch throughout the duration of the scan, often every two seconds. The final compilation of raw data, then, typically includes hundreds of thousands of spatially and temporally specific BOLD signal recordings over the course of several minutes.

Ultimately, the goal is to determine which signals are representative of neural activity and which are not. The unrepresentative signals are generally referred to as "noise," and they encompass anything from shortcomings of the hardware to head motion within the scanner. Some of these factors can be controlled for via experimental design, but each fMRI scan is still bound to generate a plethora of noise. Unless the noise appears as signal outside of the cortex, there is no perfect way of differentiating between true and false positives. This complication is the primary roadblock to most fMRI analyses, particularly when the only method of determining which signals are valid often relies on previous fMRI data and experience, posing an issue of circularity. The most effective means of sifting between true signal and noise is to subject each BOLD signal in each voxel to rigorous statistical analysis, both at the individual and group levels.

General Linear Model

The statistical method most commonly applied in fMRI analysis is the general linear model (GLM) (Poline & Brett, 2012). The GLM is a relatively simple means of fitting a model of general regressors to the signal obtained from the scanner to mathematically determine which signals are likely to be true or false. It is typically represented by the following function:

$$Y = X\text{ß} + \varepsilon$$

Equation 3. Simplified General Linear Model.

In the representation above, *Y* indicates the total BOLD signal collected by the scanner. It is composed of two major components: *Xß* and *ε. X* entails the various factors that each contribute differently to the neural activation indicated by the BOLD signal, and the conditions that influence variable contribution are accounted for by *ß* (according to mathematical convention, *ß* is the weighted slope of a variable's contribution to the function). After considering the expected contributions to total BOLD signal, there remains an unexpected residual – this error is captured by *ε.*

Equation 3 above demonstrates the basic form of the GLM, but the function is realistically represented by a series of variable matrices to account for each component. For instance, investigators need to analyze the BOLD signal throughout the duration of the scan, so the GLM must account for variables at each time point. Additionally, the model must consider the variety of factors that contribute to observed changes in BOLD signal. An example of the GLM variable matrices is shown in **Equation 4** (Poline & Brett, 2012).

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix}$$

Equation 4. Variable Matrix of General Linear Model. *In this representation, matrices account for several time points y_1 thru y_9. The design matrix includes three conditions of three time points each, the effects of which are modeled by the three associated beta weights.*

Regressors included in the GLM typically account for the expected neural response to the experimental stimulus, head motion within the scanner, scanner drift (i.e., a tendency for baseline signal to increase throughout the scan), and masks that block all signal outside of the cortex (Huettel et al., 2014). These are all modeled by *Xß.* Any signal *Y* that isn't accounted for by the expected components is considered noise, *ε*.

Once the GLM is fitted to the obtained signal such that residuals are minimized (i.e., unexplained error is as low as possible), investigators can estimate the degree to which the experimental stimuli invoked a hemodynamic response in each voxel throughout the duration of the scan. This is commonly referred to as the estimation of *ß*-weights according to the variables of the model. *ß*-weights can be analyzed and reported at the individual level for a single voxel or entire region-of-interest (ROI), but these values don't necessarily reveal much in and of themselves. Rather, the most common approach is to generate statistical contrasts (t-tests, f-tests, ANOVA, etc.) between the *ß*-weights of different conditions (e.g., experimental stimulus condition – control condition)—in this manner, each potential hemodynamic response can be quantitatively described according to its statistical significance.

It's important to note that the GLM is not the only statistical technique used to analyze fMRI data, and the example provided above is merely a simplified representation of an actual model. Other techniques involve data-driven rather than hypothesis-driven approaches, such as multi-voxel pattern analysis, independent components analysis, and several others.

Regardless of the statistical approach, the goal is always to generate some quantitative contrast between features that addresses the nature of the paradigm. In most GLM applications, statistical contrasts determine the likelihood of an observed result in the event of a true null hypothesis. For instance, a null hypothesis would assume the difference between *ß*-weights of the experimental stimulus condition and control condition to be zero, implying that there is no difference in neural activity due to the stimulus. Then, any apparent difference between conditions is tested for statistical significance—the signals that surpass the significance threshold are interpreted as unlikely in the event of a true null hypothesis, supporting the alternate hypothesis that a real difference in hemodynamic response exists between the experimental and control conditions. This difference is quantified in each voxel at every time point in all subjects of a single fMRI study.

The statistical contrasts based on the *ß*-weights of the GLM are ultimately averaged across subjects and plotted onto a structural image of a normalized brain, an example of which is shown below in **Figure 2** (Castillo, 2018).

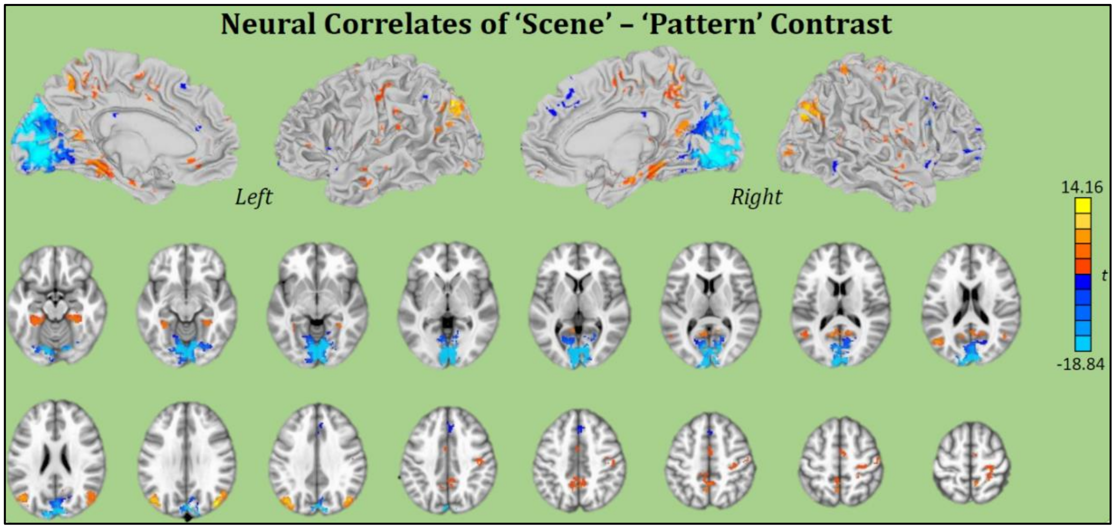

***Figure 2. Neural Correlates of 'Scene' - 'Pattern' Contrast**. Representation of a simple statistical map showing a t-test sampled between two conditions: scene viewing and pattern viewing. Upper images show lateral and medial perspectives; lower images show a montage of axial slices. Warm colors depict scene-favored viewing, and cool colors depict pattern-favored viewing.*

Contrary to popular belief, depictions such as **Figure 2** (which are present throughout the entirety of fMRI literature) do not represent the sole locations of neural activity for a certain cognitive process. More appropriately, they are maps of statistical contrast. When the contrast is based on the GLM described above, these maps reveal areas where the BOLD signal is deemed to be significantly greater during a condition designed to invoke a specific cognitive process than during a control condition in which the process is not necessary. However, as indicated earlier, several other forms of statistical analysis can be applied to fMRI data to address an array of different hypotheses and contrasts. The method of choice depends largely on the question at hand, and further reviews of the costs and benefits of various methods can be found in (Huettel et al., 2014; Mahmoudi, Takerkart, Regragui, Boussaoud, & Brovelli, 2012). Additionally, every type of analysis is subject to some form of error.

Type I and Type II Errors

When accounting for the massive volume of data generated by nearly all fMRI statistical analyses, investigators must consider the potential presence of two threats to the validity of their results: Type I and Type II errors.

A Type I error is also known as a false positive, i.e., the incorrect validation of an observed BOLD signal that isn't representative of a true hemodynamic response. A type II error, or false negative, describes the incorrect rejection of a BOLD signal that *is* representative of a true hemodynamic response. Naturally, these two error types are at odds with each other in terms of correction: the mitigation of one inevitably leads to an increase of the other. Thus, investigators are tasked to find the most appropriate balance of Type I and Type II error control in their data.

This balance is modulated by statistical thresholding. Earlier, I mentioned that signals recorded above a specific threshold are deemed significant—but, what determines the threshold? To be quite honest, thresholds throughout the history of psychological science have been somewhat arbitrary, especially the magic "$p < 0.05$" value set by Fisher in 1926 (Fisher, 1926; Lieberman & Cunningham, 2009). It's difficult to precisely determine the most acceptable amount of potential error—we only know that errors will happen, and we can try our best to keep the error count low while still capturing true data. The stricter the statistical threshold, the less Type I errors will occur because only the strongest signals will pass the test; but, any true signals that happen to be weak will be neglected. The looser the threshold, the less Type II errors will occur because true weak signals will still pass the test, but other weak signals that aren't true will also pass.

The decision that faces investigators, then, is one of weighted importance: which is more deleterious to neuroimaging analyses, false positives or false negatives? This remains a debatable topic, as both errors impact studies differently. According to Bennett (and most other neuroimagers), Type I errors pose the biggest threat due to their effect on generality (Craig M Bennett, Wolford, & Miller, 2009). If a study unknowingly reports falsely positive results, e.g., activation in the inferior frontal gyrus (IFG) when there is none, then attempts to replicate this study will result in one of two events: 1) the replicators will yield activity in the IFG as well, corroborating a false finding and laying a foundation of untruth, or 2) the replicators will not yield IFG activity, and their attempt would be considered an unfit use of valuable funds, scanning time, etc. Attempted replication of a false positive is a “catch-22” that leads to replication crises amidst neuroimagers. So, the prevailing attitude is that these dangerous errors should be staunchly corrected for via principled multiple comparison corrections such as False-Discovery Rate (FDR) or Family-wise Error Rate (FWER) control, examples of which are shown in **Figure 3** (Craig M Bennett et al., 2009). For those not convinced of the prevalence of Type I errors, Bennett made a spectacle of statistical analysis by demonstrating his ability to yield a recorded BOLD signal in a dead Atlantic salmon simply by reporting uncorrected results (C M Bennett, Miller, & Wolford, 2009).

What’s the debate, then? According to some investigators, the more looming threat remains the Type II error. The primary detriment of a false negative is that, unlike a false positive, it goes entirely unseen. Over time, Lieberman points out, we can expect Type I errors to wash out of the larger body of data because they’re not logically replicable. Type II errors, however, will never be washed out over the course of replications because they already are—*that’s the problem* (Lieberman & Cunningham, 2009). There’s no way to address missing truth, as we would never know where to look. Lieberman goes on to explain that the mainstream focus on Type I error correction not only drastically increases unfixable Type II errors, but also biases heavily in favor of large, obvious effects rather than complex cognitive and affective processes. Studies that focus on more subtle neuropsychological phenomena are subject to great strain when faced with unreasonable statistical thresholds (Lieberman & Cunningham, 2009). Those in agreeance with Lieberman claim that neuroimagers should lighten thresholds overall and rely on the collective body of data to draw inferential conclusions rather than single studies.

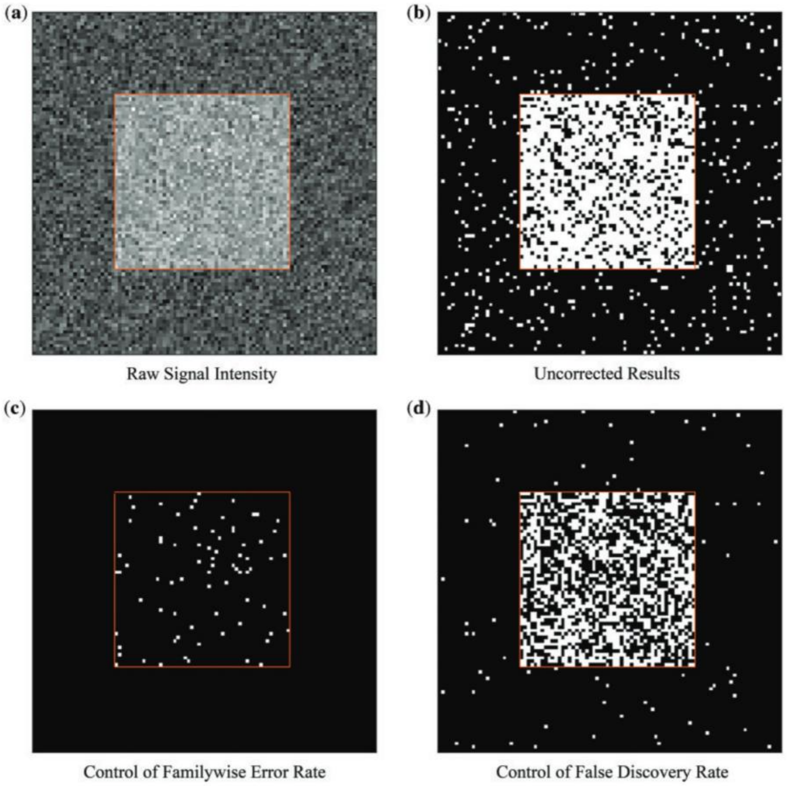

***Figure 3. Demonstration of Correction Methods for the Multiple Testing Problem**. **(a)** A raw image of the simulated data used in this example. A field of Gaussian random noise was added to a 100x100 image with a 50x50 square section of signal in the center. **(b)** Thresholded image of the simulated data using a pixelwise statistical test. The threshold for this test was P < 0.05. Power is high at 0.80, but a number of false positives can be observed. **(c)** Thresholded image of the simulated data using a Bonferroni FWER correction. The probability of a familywise error was set to 0.05. There are no false positives across the entire set of tests, but power is reduced to 0.16. **(d)** Thresholded image of the simulated data while controlling the false discovery rate. The FDR for this example was set to 0.05. Out of the results, 4.9% are known to be false positives but power is increased to 0.54.*

And so, regardless of how neuroimagers decide to threshold their functional localization analyses, errors will occur. While this hindrance is certainly not unique to fMRI analysis, it is only one of many statistical obstacles in our comprehensive exploration of the functional mechanics of the human brain. Another equally prevalent yet relatively unaddressed issue is that of unmeasured factors.

Unmeasured Factors

Generally, the scientific community accepts the fact that any quantification of a qualitative characteristic (like a cognitive process) is bound to miss some facet of the characteristic itself—in many cases, this isn't a problem so long as it's acknowledged. Still, the breadth of statistical analyses can fall too short to account for the questions we ask. The issue is broader and deeper than the control of Type I and Type II errors I discussed earlier—we face the regular danger of imposing shallow models onto our data that fail to capture our research questions, giving rise to conclusions that are statistically significant but inferentially insignificant.

The root of our analytical shortfall is the presence of unmeasured factors, i.e., variables that aren't considered in data analysis and interpretation (Yarkoni, 2019). Scientific analysis will almost always yield unmeasured factors, but unmeasured factors inevitably lead to constraints of generality, which describes the likelihood of inferential conclusions to apply outside of the context from which they were drawn (Simons, Shoda, & Lindsay, 2017). In practice, almost every scientific discipline—particularly those dealing with human nature—relies entirely on the concept of generality to maintain relevance. If the inferential conclusions of a study don't apply in any way to circumstances apart from that particular study, then there's practically no point in conducting it to begin with. Thankfully, conclusions in most neuropsychological studies are logically generalizable to some extent due to the common thread of human nature. Despite this fundamental commonality, though, cognitive neuroscientific analysis still bears its fair share of unmeasured factors that reduce the breadth of generality.

Unmeasured factors in task-based functional localization studies can include stimulus characteristics, experimental procedures, lab conditions, and subject sample population—all of which potentially influence recorded neural responses to a substantial degree (Yarkoni, 2019). I'm not claiming that these variables go entirely unconsidered, as investigators regularly attempt to balance sensory effects of stimuli within the scanner, and sample populations are often targeted for specific characteristics. Rather, these factors are unmeasured because most statistical models simply don't account for them—design matrices include a whole host of variables that influence the probable BOLD signal, but "stimulus characteristics" is rarely one of them. Instead, stimulus characteristics and other unmeasured factors are implicitly considered fixed effects, meaning *differences in these variables would alter the nature of the analysis.* This issue spans the longitudinal breadth of psychological research, as it was originally described as the "fixed-effect fallacy" nearly 50 years ago (Clark, 1973).

One means of accounting for unmeasured factors is to incorporate them into a mixed-effects model as random effects (Friston, Stephan, Lund, Morcom, & Kiebel, 2005; Yarkoni, 2019). In short, random effects are variables that, if altered across replication studies, shouldn't alter the nature of the analysis itself. A common random effect among most task-based fMRI analyses is between-subject variability—it goes without saying that one should be able to repeat the study with a different set of subjects and obtain similar results if it is to be considered generalizable. However, a less noted but equally important perspective is that one should also be able to alter the experimental stimuli used in the task and still obtain similar results. A detailed approach to incorporating more random effects is described by (Friston et al., 2005; Yarkoni, 2019).

Unfortunately, the inclusion of multiple random effects inherently increases the variance and thus reduces the power of any statistical model. After all, if an observed BOLD effect of a task-based fMRI analysis is due to any factors other than the presence of a specific cognitive process, a random effects model would reveal less significant statistical contrasts between experimental and control conditions (Yarkoni, 2019). Not only does this discourage replication studies by constraining generality, but it makes it difficult to ask pointed research questions and still yield positive results.

As important as it is to account for unmeasured factors in fMRI analysis, though, I don't believe the incorporation of more random effects is a panacea for every problem fMRI investigators face when conducting functional localization studies. After all, modern computational developments allow the application of complex data-driven models that don't possess the same statistical constraints as the experimental contrast-based approach I've discussed thus far, yet the battle of methodological shortfalls remains far from over. Even if a statistical model were to account for every factor related to BOLD signal changes, and thresholds provided a perfect limitation of Type I and Type II errors (this is, of course, impossible), the fact remains: statistical analysis as an operationalization of experimental paradigm design. Thus, if the paradigm design is incomplete, then any statistical results (no matter how true or significant) will also be incomplete.

Experimental Paradigm Design: An Operationalization of Theoretical Perspective

The breadth of fMRI research can expand only as far as its questions allow, and its questions must adhere to the constraints of the experimental paradigms used to explore them. In this sense, the experimental paradigm of an analysis is an operationalization of the theoretical perspective expressed by the verbal hypothesis. The two primary modes of fMRI experimentation are task-based experiments and resting state functional connectivity analyses, the latter of which have only recently become mainstream. For the purposes of this paper, I will focus on task-based fMRI functional localization paradigms, which are designed to create environments that artificially induce neuropsychological phenomena in a laboratory setting.

Task-Based fMRI

Task-based paradigms rest on the premise that cognitive processes are spatially and temporally correlated with corresponding neural activity (Amaro & Barker, 2006). This premise is valid, as lesion-based case reports (such as Phineas Gage) and basic science studies of neuronal function have demonstrated that certain regions of the brain are truly associated with specific activities. To explore the relationships between location and function, these designs generally consist of several sets of stimuli and instructed responses. Experimental stimuli and instructions are designed to induce specific cognitive processes.

In the previously mentioned n-back task, for example, participants are shown a series of simple images. When they see images that appeared *n* times previously in the series, they press a button. This stimulus-response pattern is repeated several times throughout the scan. The premise is that the task requires participants to consistently rely on working memory to successfully follow instructions, so the neural activation detected by the scanner should represent the necessary components of working memory (Kirchner, 1958; Redick & Lindsey, 2013). An example of the n-back task is shown in **Figure 4** (Ragland et al., 2002)

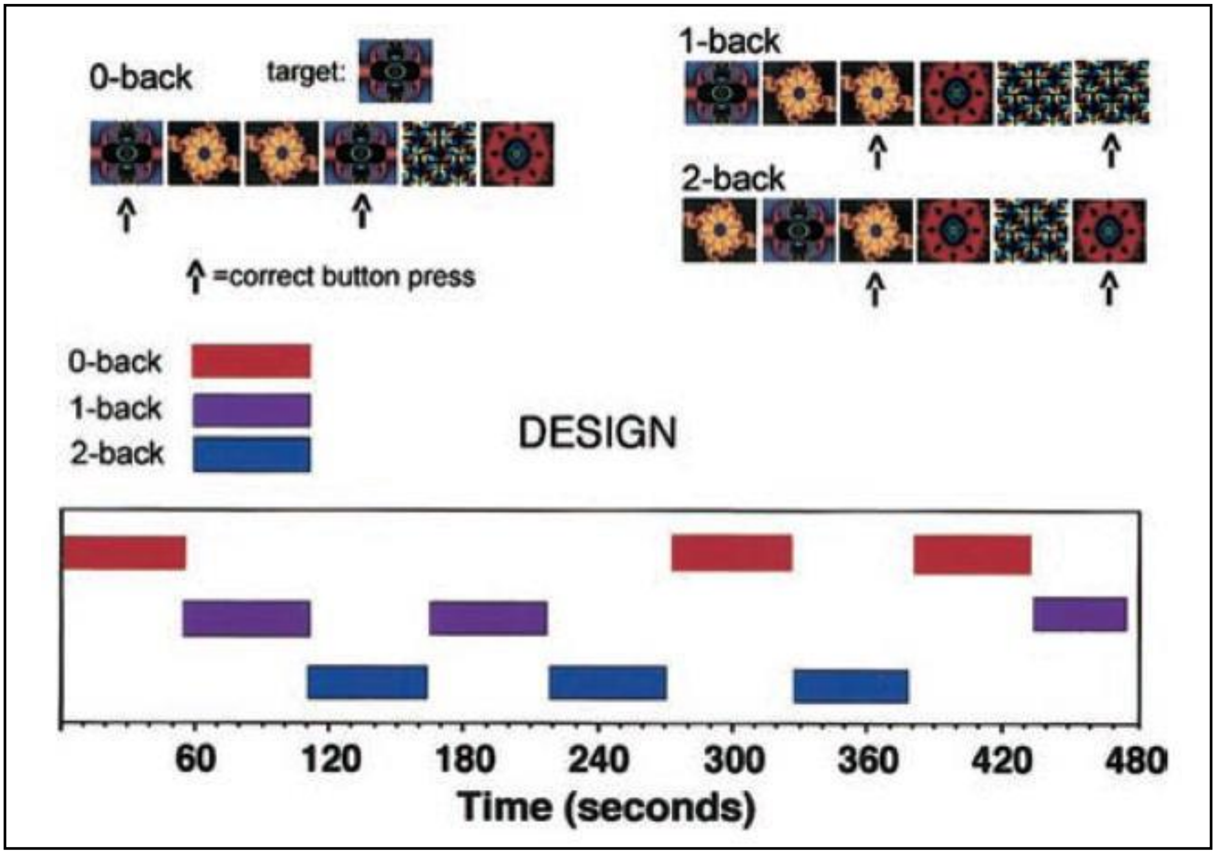

Figure 4. Demonstration of n-back Design Using Fractals as Target Stimuli. *This is an example of the n-back task-based experimental design. Subjects are required to memorize a target stimulus and indicate when it appears again 'n' times later in the series. Stimuli are presented as a block design with three conditions over the course of 480 seconds.*

Like any experiment, task-based fMRI paradigms rely on a control condition to provide desirable experimental contrasts. Control conditions can be simple fixation (i.e., staring at the screen) or special tasks that help control for irrelevant components of the experimental stimuli. For example, a control condition for the n-back task may be passively viewing the image series without attempting to recall previous images—when activation in the task is compared to activation in the control, the only differences should be those due to the use of working memory. In the example from (Ragland et al., 2002), the control condition is the 0-back task, where participants need only identify a target stimulus without actively recalling the series.

Experimental paradigms are typically highly controlled and carefully designed to provide precisely what neuroimagers desire to see. However, I believe the major shortfall of fMRI investigation is rooted in experimental paradigm design.

Achilles Heel of Experimental Paradigm Design

The most prevalent yet silent weakness of experimental paradigm design is failure to operationalize the theoretical perspective, an issue that spans the ocean of previous and current fMRI functional localization literature. If we aren't answering what we're asking, we aren't making progress—we are digressing at best. If we proceed under the impression of progress via statistically significant analyses and experimental replications, we are realistically regressing. *Thus, failure to capture the breadth of our hypotheses via paradigm design is the fundamental source of all problems discussed in this paper.*

For instance, the brain regions derived as the working memory network via the n-back experimental paradigm are suitably involved in working memory for a specific n-back task with a special set of stimuli given a certain statistical contrast; without external validation, this result is not generalizable to all working memory function apart from this particular task. Thus, a conclusion along the lines of, *"The dorsolateral prefrontal cortex is a fundamental component of successful working memory encoding,"* is not applicable to the results of a paradigm such as shown in (Ragland et al., 2002), as this task is specific to arbitrary stimuli presented in a single operationalization of working memory. Any neural correlates determined to be significant components must be considered in light of the context. Further conclusions regarding the working memory encoding network must be drawn from separate studies with different parameters.

Granted, cognitive processes are not always far removed from the realm of our operationalization. Certain neuropsychological experiences are relatively easy to induce in the experimental context (e.g., pain), leading to a more established connection between verbal hypotheses and experimental operationalization. In these cases, we can generally worry less about the disconnect between our questions and means of pursuing answers. Other neuropsychological phenomena, however, can be much harder to artificially invoke within the laboratory, rendering operationalization more difficult. Lieberman et. al. point out the difficulties that many psychologists face along these lines when attempting to analyze social and emotional activity within the brain (Lieberman & Cunningham, 2009). I would contend that these difficulties generalize to nearly all psychological constructs apart from basic sensorimotor functions, at the very least because the time-courses of complex thoughts are bound to vary between subjects, whereas sensorimotor responses are generally consistent across healthy populations (Lieberman & Cunningham, 2009). Additionally, psychological processes apart from sensorimotor functions tend to occur in light of and in concert with other processes, as this is how we typically function in daily life. Thus, the ability to evoke relevant neural responses in the experimental context typically decreases as the complexity of the research question increases.

At a certain point (rather, at every point), we must consider whether our experimental paradigms are actually addressing the questions we are asking. As straightforward as this sounds, the potential disconnect between theoretical perspectives and experimental operationalization is often and unfortunately overlooked. This has much to do with the convention of the general scientific community—once an operationalization is determined valid by replication, the breadth of its theoretical perspective is rarely questioned. Under closer analysis, though, this convention ultimately defeats the purpose of its existence.

Replicability: A Validation of Operationalization?

In general, replicability describes a paradigm's tendency to yield similar results across similar contexts. As a scientific convention, this is essential to the corroboration of perspectives over time and space—fundamentally, though, replication is a test of methodology, not theory. If a thousand different investigators each independently validate a task-based fMRI paradigm by conducting analyses and obtaining similar neural correlates, we can determine that the paradigm itself is stable and trustworthy of producing those results. The paradigm is not, however, guaranteed to be a valid operationalization of any specific verbal hypotheses. Each of the thousand investigators may or may not have asked the same question, and they may or may not have interpreted the results similarly in light of their questions. Even if they each ask the same question and interpret their results the same way, all we know is that the paradigm consistently yields a certain pattern of results—we cannot infer that the paradigm adequately addresses the research question on the grounds of its replicability alone.

One of the leading voices addressing the disconnect between established experimental paradigms and our understanding of neuropsychological phenomena is Jacqueline Sullivan, Associate Professor of Philosophy at the University of Western Ontario.

Sullivan has spent years challenging the status quo of neuroscience research by analyzing paradigm validity and reliability on a case-by-case basis (J. Sullivan, 2007). One of her most noted works is her analysis of the development and modern use of the Morris water maze, a classic learning and memory task that has stood the test of time and analysis for over 35 years (D'Hooge & De Deyn, 2001; Morris, 1984). The paradigm itself has been historically employed as a quantitative and qualitative measure of spatial memory, but Sullivan points out that studies have never quite captured "precisely 'what' investigators who train rodents in the water maze are actually investigating," (J. A. Sullivan, 2010). Rather, we know that the paradigm has proved replicable over thousands of different contexts, and the maze theoretically provides a good model for learning and memory analysis in general, but that's about all we know. The specific neuropsychological components captured by the Morris water maze (e.g., spatial learning, navigation, cognition, etc.) typically become clumped into broad phrases like "spatial memory" or simply go unaddressed entirely (J. A. Sullivan, 2010). As it stands, *we actually don't know precisely what cognitive faculty the water maze delineates.*

Though most of Sullivan's work lies outside of fMRI functional localization as described in this paper, the same issue applies to experimental paradigms throughout neuroimaging literature. Most identifiable cognitive processes are associated with standard paradigmatic approaches that, due to their replicability, are largely accepted by the scientific community. The consensus indicates that paradigm replicability has become a misunderstood validation of experimental operationalization—thousands of replications cannot establish the breadth of verbal hypotheses a paradigm realistically addresses. At this stage, it's likely that several standardized paradigms account for cognitive faculties beyond or entirely apart from those listed as intentional targets. Whether we acknowledge it or not, the range of potential fMRI-based verbal hypotheses is dramatically limited by our ability to operationalize underlying psychological constructs.

Moving Forward

Considering the seemingly unapproachable disconnect between verbal hypotheses and experimental operationalizations, we can easily default to the fact that this is simply what the field of functional task-based neuroimaging looks like right now—we can't just throw out everything we've done up to this point to accommodate an error of scientific reasoning that's a prevalent threat to *every* field. Inasmuch as it would be highly impractical to "restart" our approach to task-based fMRI experimentation, I agree. We have doubtlessly made considerable progress in our understanding of functional localization since the birth of fMRI nearly thirty years ago. And, our experimental operationalizations have indeed shifted both paradigmatically and statistically in monumental ways to better accommodate the breadth of our hypotheses. But, we can't simply pretend to make progress from here without addressing the shortfalls of our experimental approach. The further we allow our theoretical perspectives to shift from the bounds of our methodology, the greater a disservice we are committing to the field of neuroimaging. So, how can we move forward in fMRI experimentation such that we establish meaningful, reproducible, and relevant inferential conclusions regarding functional localization?

Operationalization and Inference: Bridging the Gap

The key lies in acknowledging and bridging the gaps between our verbal hypotheses, experimental paradigms, and statistical analyses. The first step is tailoring our research questions to fit into the grand scheme of the human experience rather than vice versa. Consider the behavior of the human brain—it's no secret that cognitive processes occur in an incredibly expansive and complex concerted fashion (Bargmann & Marder, 2013; Gazzaley, Cooney, McEvoy, Knight, & D'Esposito, 2005; Helfrich, Breska, & Knight, 2019; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999; Marek, Sun, & Sah, 2019; Merrikhi, Clark, & Noudoost, 2018). One can rationally claim that no psychological process realistically functions in complete isolation from other processes, since the human experience does not control for cognitive states the way fMRI investigators do. Rather, our brains perceive the world in concert, approaching and processing the abundance of stimuli with a complex mixture of top-down and bottom-up processes in both serial and parallel circuits. Of course, top-down directed attention selectively modulates the induction and maintenance of situationally relevant cognitive processes, but even this is an example of concerted processing—at no point do our brains operate on a simple unidimensional scale of a singular process.

If we're truly interested in exploring the neural correlates of a particular neuropsychological phenomenon, we should do so in the context which envelops it. "Controlling" for this context through tight experimental paradigmatic contrast is like removing a fish from the water to examine how it swims. The example sounds silly, but we effectively do the same thing via conventional experimental paradigms that utilize highly controlled conditions to generate significant and replicable statistical contrasts. Tight experimental contrasts are generally viewed with greater respect as they theoretically isolate the cognitive processes in question, but I believe that tighter experimental designs can potentially widen the chasm between operationalization and inference when approaching complex psychological processes. To echo an earlier point, fMRI analysis only yields the neural correlates of the operationalization of the psychological construct in question, and tightening a design by adding stricter conditions only limits the breadth of potential inference based on that operationalization.

Say, for instance, we wish to explore language processing: if the brain requires modulatory attention networks to successfully process verbal language stimuli, we should consider those "unrelated" networks with the same perspective as those we're targeting rather than contrast them out of the results. After all, the brain apparently considers these extraneous networks necessary and natural to complete the task at hand. Additionally, there's quite a possibility that integrative capacities such as attention modulate processes differently according to the circumstances, so they may demonstrate nuanced differences in activation that would only be observed if accounted for by the experimental contrasts.

I realize that this "uncontrolled" experimental approach can easily grow out of hand and too messy for any beneficial inferential conclusions. If everything appears significant, nothing appears significant. I should clarify that I do not believe that abandoning all experimental contrast in neuroimaging is the best step forward; however, I do think we can remodel our paradigms to encompass the generality of our hypotheses without sacrificing the necessary level of experimental control.

To do so, perhaps we should consider designing paradigms with variation in mind (Yarkoni, 2019). This would involve looser, less precise designs that emphasize and even embrace variability. Granted, experimental variability is a scary thing, as it implies loss of control and typically renders statistical analyses much less impressive (depending on the methods). Though the emphasis on variability may sound counterintuitive at first, this approach ultimately leads to more generalizable conclusions, whether positive or negative. And, as investigators attempt to overcome the "replication crisis" of the current age, generality is all too sought-after. As a result, the proposition of intentionally implementing experimental variability is on the rise among modern neuroimagers (Hamilton & Huth, 2018; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Yarkoni, 2019).

Consider Hamilton and Huth, who have spent the last few years pointing out that fMRI scanning and analytical technologies have advanced tremendously, yet our approach to asking deep, impactful questions has remained relatively unchanged (Hamilton & Huth, 2018). In short, we've defaulted to highly controlled, simplified stimuli such as randomly presented images, words, sounds, etc. that allow for tight experimental contrasts. While simplified stimuli have previously proven highly useful in several neuroimaging contexts, we've reached a point at which our questions are too real to engage using the same mindset (David, Vinje, & Gallant, 2004). Conclusions based on former practices may already be outdated and incomplete, regardless of how many times we replicate them. To address the shortfall of experimental paradigm design and gradual obsoletion of simplified stimuli, Hamilton and Huth argue for the implementation of natural stimuli, particularly in neuroimaging studies of language (Hamilton & Huth, 2018).

Paradigm Design: Natural Stimuli

What qualifies a stimulus as natural? According to Hamilton and Huth, the naturalness of a stimulus can be gauged along a spectrum by three sensible questions:

- Would the person reasonably be exposed to this stimulus outside of the experimental setting?
- Does the stimulus appear in the same context as it would in real life?
- Is the subject's motivation for perceiving and understanding the stimulus one that would occur outside of the experimental stimulation, i.e., in real life?

Stimuli that meet these standards are considered more natural than those that don't. As expected, though, these standards induce a great deal of experimental variability compared to tightly controlled stimuli. For instance, a stimulus that appears in the same context as it would in real life must contain several variables of differing degrees and domains. Additionally, subject motivation to perceive and understand is highly dependent upon stimulus content, so natural stimuli would introduce a whole host of semantic influences to the paradigm. These effects are entirely counterintuitive to the mainstream approach of task-based functional localization, which has historically involved an experimental contrast that holds all variables equal across conditions except for a single variable of interest. This former approach is statistically powerful as it minimizes confounding variables, but Hamilton argues that it rarely yields relevant inferential conclusions (Hamilton & Huth, 2018). Ultimately, the nuanced differences in processing patterns of two unrealistic experimental conditions aren't applicable to any realistic human context, especially for complex cognitive functions. Natural stimuli, however, induce neural responses that capture neuropsychological phenomena much closer to their realistic entirety.

In the context of language studies, natural stimuli appear less as isolated words or sentences and more as complete narratives (Lerner, Honey, Silbert, & Hasson, 2011; Wehbe et al., 2014). Narratives possess the advantage of ensuring that each sentence occurs in a natural context that ultimately delivers a story worth listening to. When sampled from real-world sources like popular books or radio shows, the stimuli meet each of Hamilton and Huth's naturalness criteria. And, as expected, the use of natural stimuli in experimental paradigms has resulted in several holistic improvements of the functional localization of language processing.

For instance, naturally designed experimental paradigms give rise to the detection of a much vaster anatomical network of areas engaged in various aspects of language processing (Cogan et al., 2014; Obleser, Eisner, & Kotz, 2008). Namely, the process that has famously been perceived as left-lateralized according to simplified stimuli studies demonstrates a high degree of bilateral function according to natural stimuli studies (Berwick, Friederici, Chomsky, & Bolhuis, 2013; Geschwind, 1970; Jung-Beeman, 2005). This likely arises due to the necessitation of semantic comprehension of thoughts and stories, which reflects language processing as it occurs in the natural context.

Another benefit of applying natural stimuli is the increase of experimental efficiency (Hamilton & Huth, 2018). Though the inclusion of multiple variables appears to reduce inferential validity, it actually allows investigators to compare the relative contributions of several variables within a single set of stimuli instead of conducting numerous conventional experiments with highly controlled stimuli (Wehbe et al., 2014). Despite the benefits, though, these approaches are far removed from the traditional methods of single-variable adjusted contrasts—so much so that the analytical methods must differ entirely from previous standards.

Natural Stimuli: Statistical Analysis

As a result, one of the greatest obstacles to conducting statistical analysis of BOLD responses to natural stimuli is the lack of clear contrasts—conventional analyses such as t-tests, f-tests, or ANOVA aren't applicable to paradigms that allow so much variability within and across conditions. For some, this is reason enough to avoid the practice overall. However, there are several other statistical models that prove successful in the realm of natural stimuli and are widely available with common statistical software packages. One of the most common methods is the linearized encoding model, which circumvents issues of variability via machine learning processes (Hamilton & Huth, 2018; Holdgraf et al., 2017; Naselaris, Kay, Nishimoto, & Gallant, 2011). Rather than contrasting BOLD response between two strict conditions, the linearized encoding model accounts for known stimulus features and "learns" to detect the neural responses based on these features. The trained model is then applied to a validation dataset to determine how well it accounts for the BOLD responses of each stimulus feature—the resulting statistical contrast reveals how well the model predicts actual responses (Naselaris et al., 2011). If the model is significantly accurate, then the parameters estimated for the natural stimulus features must be accurate. (This is analogous to accurately determining *ß* -weights of the GLM.) When features are not known *a priori,* other viable methods of analyzing natural stimuli data include convex non-negative matric factorization and inter-subject correlation (Hamilton & Huth, 2018).

Whatever model is applied to fMRI functional localization data, the results of natural paradigms are more likely to generalize beyond the context of the experiment than traditional paradigmatic approaches because they account for much more variation than highly controlled contrasts. Though the multitude of variables can easily become entangled, the confounds can be extracted by careful regression analysis within encoding models or otherwise (Hamilton & Huth, 2018). For investigators who find too little control within natural stimulus paradigms, the stimuli themselves can be manipulated within reason to produce a somewhat familiar analytical contrast, such as changing the pitch of a sentence or delivering stimuli according to a different time scale (Lerner et al., 2011; Tang, Hamilton, & Chang, 2017).

Implementing Natural Stimuli in other Contexts

Admittedly, speech and reading comprehension are perhaps the simplest among complex cognitive processes to study using natural stimuli. It's practically easier to engage a subject naturally within a scanner by reading to them than by instructing them to memorize something or feel a specific emotion. Thus, studies of language processing currently remain at the forefront of natural stimulus paradigm employment. Given the practical limitations of fMRI analysis, the most effective way to generalize Hamilton and Huth's perspective of experimental design to other contexts of neuroimaging is likely through the use of videos and other naturally engaging stimuli that elicit a genuine motivation to apply the targeted cognitive process as one would in real life. These stimuli would, of course, bear the same limitations and introduce the same variability as natural language stimuli, if not more. However, I believe the benefits of invoking and analyzing realistic, generally applicable neural responses are well worth the increased cost of less powerful statistical contrasts or more complex computational methods.

Conclusion

In short, the breadth of interpretative inference is principally limited by the operationalization upon which inferential conclusions rest. This applies broadly to scientific philosophy in general, but it speaks highly to the current direction of task-based fMRI functional localization research—we've moved too far with too little grounds to do so. To clarify, though, this paper is by no means a dismissal of previous neuroimaging work. Were it not for the monumental progress allowed by highly controlled experiments, simplified stimuli, and tightly run statistical contrasts, we would never have come close to the current stage of functional localization. Rather, I believe we have reached a point at which the field of neuroimaging is geared to explore the depths of the working components of the human brain more fully than we would have thought possible. To achieve this potential, though, we must ensure that our experimental operationalizations remain in sync with our theoretical perspectives—only then will our inferential conclusions possess the merit we desire to assign them. As is stands, the use of natural stimuli within variable experimental paradigms is one potential means of mending the disconnect between question and answer. Regardless of how we move forward, though, the logical principles of inference must move forward with us. So long as they do, I have great hopes for the future of neuroimaging research.

Acknowledgments

I’d like to thank Dr. Kristina Visscher and Dr. Cristin Gavin for invaluable conversational contribution toward the direction and completion of this project.

References

Amaro, E., & Barker, G. J. (2006). Study design in fMRI: basic principles. *Brain and Cognition*, *60*(3), 220–232. doi:10.1016/j.bandc.2005.11.009

Bargmann, C. I., & Marder, E. (2013). From the connectome to brain function. Nature Methods, 10(6), 483–490. doi:10.1038/nmeth.2451

Bennett, C M, Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. Neuroimage, 47, S125. doi:10.1016/S1053-8119(09)71202-9

Bennett, Craig M, Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. Social Cognitive and Affective Neuroscience, 4(4), 417–422. doi:10.1093/scan/nsp053

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. Trends in Cognitive Sciences, 17(2), 89–98. doi:10.1016/j.tics.2012.12.002

Castillo, S. J. (2018, July). Neural Correlates of Visual Episodic Encoding: an fMRI Analysis. Poster Presentation presented at the Neuronal Networks in Epilepsy and Memory, Louisiana Tech University.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of verbal learning and verbal behavior, 12(4), 335–359. doi:10.1016/S0022-5371(73)80014-3

Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., & Pesaran, B. (2014). Sensory-motor transformations for speech occur bilaterally. Nature, 507(7490), 94–98. doi:10.1038/nature12935

D'Hooge, R., & De Deyn, P. P. (2001). Applications of the Morris water maze in the study of learning and memory. Brain Research. Brain Research Reviews, 36(1), 60–90. doi:10.1016/S0165-0173(01)00067-4

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. The Journal of Neuroscience, 24(31), 6991–7006. doi:10.1523/JNEUROSCI.1422-04.2004

Fisher, R. A., Sir,-. (1926). 048: The Arrangement of Field Experiments.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. Neuroimage, 24(1), 244–252. doi:10.1016/j.neuroimage.2004.08.055

Gazzaley, A., Cooney, J. W., McEvoy, K., Knight, R. T., & D'Esposito, M. (2005). Top-down enhancement and suppression of the magnitude and speed of neural activity. Journal of Cognitive Neuroscience, 17(3), 507–517. doi:10.1162/0898929053279522

Geng, Y., Zhao, W., Zhou, F., Ma, X., Yao, S., Hurlemann, R., … Kendrick, K. (2018). Oxytocin enhancement of emotional empathy: generalization across cultures and effects on amygdala activity. BioRxiv. doi:10.1101/307256

Geschwind, N. (1970). The organization of language and the brain. Science, 170(3961), 940–944. doi:10.1126/science.170.3961.940

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. Neurosurgery clinics of North America, 22(2), 133–9, vii. doi:10.1016/j.nec.2010.11.001

Goodwin, G. M., & Norbury, R. (2016). The amygdala and fear. In Stress: concepts, cognition, emotion, and behavior (pp. 305–310). Elsevier. doi:10.1016/B978-0-12-800951-2.00037-6

Hamilton, L. S., & Huth, A. G. (2018). The revolution will not be controlled: natural stimuli in speech neuroscience. Language, cognition and neuroscience, 1–10. doi:10.1080/23273798.2018.1499946

Harlow, J. M. (1999). Passage of an iron rod through the head. 1848. The Journal of Neuropsychiatry and Clinical Neurosciences, 11(2), 281–283. doi:10.1176/jnp.11.2.281

Helfrich, R. F., Breska, A., & Knight, R. T. (2019). Neural entrainment and network resonance in support of top-down guided attention. Current opinion in psychology, 29, 82–89. doi:10.1016/j.copsyc.2018.12.016

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. Frontiers in Systems Neuroscience, 11, 61. doi:10.3389/fnsys.2017.00061

Huettel, S. A., Song, A. W., & McCarthy, G. (2014). Functional magnetic resonance imaging (Third edition.). Sunderland, Massachusetts, U.S.A.: Sinauer Associates, Inc., Publishers.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453–458. doi:10.1038/nature17637

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. Trends in Cognitive Sciences, 9(11), 512–518. doi:10.1016/j.tics.2005.09.009

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron, 22(4), 751–761. doi:10.1016/s0896-6273(00)80734-5

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. Journal of experimental psychology, 55(4), 352–358. doi:10.1037/h0043688

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. The Journal of Neuroscience, 31(8), 2906–2915. doi:10.1523/JNEUROSCI.3684-10.2011

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. Social Cognitive and Affective Neuroscience, 4(4), 423–428. doi:10.1093/scan/nsp052

Lindquist, M. A., Meng Loh, J., Atlas, L. Y., & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. Neuroimage, 45(1 Suppl), S187–98. doi:10.1016/j.neuroimage.2008.10.065

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., & Brovelli, A. (2012). Multivoxel pattern analysis for FMRI data: a review. Computational and mathematical methods in medicine, 2012, 961257. doi:10.1155/2012/961257

Marek, R., Sun, Y., & Sah, P. (2019). Neural circuits for a top-down control of fear and extinction. Psychopharmacology, 236(1), 313–320. doi:10.1007/s00213-018-5033-2

Merrikhi, Y., Clark, K., & Noudoost, B. (2018). Concurrent influence of top-down and bottom-up inputs on correlated activity of Macaque extrastriate neurons. Nature Communications, 9(1), 5393. doi:10.1038/s41467-018-07816-4

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. Journal of Neuroscience Methods, 11(1), 47–60. doi:10.1016/0165-0270(84)90007-4

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. Neuroimage, 56(2), 400–410. doi:10.1016/j.neuroimage.2010.07.073

O'Driscoll, K., & Leach, J. P. (1998). No longer Gage": an iron bar through the head. Early observations of personality change after injury to the prefrontal cortex. BMJ (Clinical Research Ed.), 317(7174), 1673–1674.

Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. The Journal of Neuroscience, 28(32), 8116–8123. doi:10.1523/JNEUROSCI.1290-08.2008

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academy of Sciences of the United States of America, 87(24), 9868–9872. doi:10.1073/pnas.87.24.9868

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron, 72(5), 692–697. doi:10.1016/j.neuron.2011.11.001

Poline, J.-B., & Brett, M. (2012). The general linear model and fMRI: does love last forever? Neuroimage, 62(2), 871–880. doi:10.1016/j.neuroimage.2012.01.133

Ragland, J. D., Turetsky, B. I., Gur, R. C., Gunning-Dixon, F., Turner, T., Schroeder, L., … Gur, R. E. (2002). Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. Neuropsychology, 16(3), 370–379.

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: a meta-analysis. Psychonomic Bulletin & Review, 20(6), 1102–1113. doi:10.3758/s13423-013-0453-9

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on psychological science : a journal of the Association for Psychological Science, 12(6), 1123–1128. doi:10.1177/1745691617708630

Sprooten, E., Rasgon, A., Goodman, M., Carlin, A., Leibu, E., Lee, W. H., & Frangou, S. (2017). Addressing reverse inference in psychiatric neuroimaging: Meta-analyses of task-related brain activation in common mental disorders. Human Brain Mapping, 38(4), 1846–1864. doi:10.1002/hbm.23486

Sullivan, J. (2007). Reliability and Validity of Experiment in the Neurobiology of Learning and Memory (Doctoral dissertation). University of Pittsburgh. Retrieved from http://d-scholarship.pitt.edu/8449/1/JASullivan06.29.07.pdf

Sullivan, J. (2015). Experimentation in cognitive neuroscience and cognitive neurobiology. In J. Clausen & N. Levy (eds.), Handbook of Neuroethics (pp. 31–47). Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-4707-4_108

Sullivan, J. A. (2010). Reconsidering "spatial memory" and the Morris water maze. Synthese, 177(2), 261–283. doi:10.1007/s11229-010-9849-5

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. Science, 357(6353), 797–801. doi:10.1126/science.aam8577

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. Plos One, 9(11), e112575. doi:10.1371/journal.pone.0112575

Yarkoni, T. (2019). The Generalizability Crisis. doi:10.31234/osf.io/jqw35

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nature Methods, 8(8), 665–670. doi:10.1038/nmeth.1635