

Understanding the Mind by Measuring the Brain

Lisa Feldman Barrett

Understanding the Mind by Measuring the Brain

Lessons From Measuring Behavior (Commentary on Vul et al., 2009)

Lisa Feldman Barrett

*Boston College, Massachusetts General Hospital, and Harvard Medical School*

Address correspondence to Lisa Feldman Barrett, Department of Psychology, Boston College, Chestnut Hill, MA 02467; e-mail: [barretli@bc.edu](mailto:barretli@bc.edu).

**ABSTRACT**—Throughout the history of psychology, the path of transforming the physical (muscle movements, verbal behavior, or physiological changes) into the mental has been fraught with difficulty. Over the decades, psychologists have risen to the challenge and learned a few things about how to infer the mental from measuring the physical. The Vul, Harris, Winkielman, and Pashler (2009, this issue) article points out that some of these lessons could be helpful to those of us who measure blood flow in the brain in a quest to understand the mind. Three lessons from psychometrics are discussed.

In 1862, Wilhelm Wundt tried to measure the speed of thought by tracking the discrepancy between the actual and perceived position of a swinging pendulum. By 1879, he had invented the reaction time experiment to measure the speed of perception by presenting participants with a tone or light of a particular color and measuring their latency to press or release a button in response. With these first experiments in psychology, Wundt's goal was to identify and measure the atoms of the mind—the most elemental processes that are the basic ingredients of mental life. Wundt's method remains a standard in the science of psychology today: Researchers carefully observe something physical (be it a set of muscle movements such as reaction time, a verbal response such as a self-reported experience, or a bodily response such as changes in heart rate) and record variations in these measurements across time or context. Somehow, we figure out which part of the observed variation is signal (the variation that are meaningful to us and that we want to explain) and which is noise (the variation we don't care about). We then use the physical to make inferences about the mental. We interpret the “signal” in terms of its psychological meaning and assume that the “noise” does not contaminate this interpretive process.

Throughout the history of psychology, the path of transforming the physical (muscle movements, verbal behavior, or physiological changes) into the mental has been marked with unforeseen problems. Psychologists made many mistakes along the way. Yet over the years, we have also learned a good many things about how to avoid these pitfalls as we measure behavior in various forms and guises to understand the mind. The publication of Vul, Harris, Winkielman, and Pashler (2009, this issue) provides an opportunity to highlight some of these lessons. As it turns out, the perils of inferring the

mind from measuring the brain are not all that different from those encountered when attempting to infer the mind from measuring behavior. In this commentary, I briefly discuss three lessons from classical measurement theory using test construction as an analogy. When psychologists build a test (a standardized procedure for sampling behavior), a smaller group of items are selected from a larger pool by some means, with the goal of measuring some deep psychological property or trait. As I hope you will see, there is a pretty direct parallel between “items in a test” and “voxels in the brain.” As a consequence, we modern-day neuroscientists can learn a thing or two from 20th century psychometrics.

### **LESSON 1: DISTINGUISHING RELIABILITY FROM VALIDITY**

The Minnesota Multiphasic Personality Inventory (MMPI) is a popular personality test given to help diagnose mental disorders. Created in the 1930s, the MMPI was developed using an external, criterion-based approach to test construction (Hathaway & McKinley, 1940). Researchers assembled a vast, heterogeneous pool of items, many of which had no apparent face validity. These items were then administered to samples of patients, and their friends and relatives, at the University of Minnesota Hospital. During test construction, items were selected on a purely empirical basis—those that discriminated between patient samples (the criteria) were retained in the test (regardless of their content) because they were effective (i.e., the items could reliably distinguish between groups of people). Yet, researchers went further to also assume that these items were measuring something real and important about the respondents’ mental health (i.e., that the items were valid). Items were grouped into subscales and named for the diagnostic category they best discriminated on the assumption that the items measured

the mental essence of a mental disease and could be used to diagnose it. For example, Scale 8 used to be called “the schizophrenia scale” because it contained the set of items that best classified those who held a clinical diagnosis of schizophrenia and those who did not.

Researchers and clinicians had to learn the hard way that the MMPI subscales did not give a direct and unencumbered window on the mental illnesses that they were designed to measure. There is an important difference between a test’s effectiveness and its meaning—a set of items can be consistently valuable in discriminating two groups without meaningfully measuring any psychological property of interest (Burish, 1984). What if the schizophrenic patients differed from controls in some other systematic (but spurious) way that allowed the items to discriminate between the groups and allowed the test to be effective even though these items did not measure schizophrenia per se? The observation that a group of schizophrenic patients score higher on a group of items does not in and of itself warrant the conclusion that the responses to the items measure schizophrenia. Vul et al. demonstrate a parallel observation about the relation between voxels and mental states (and even about weather stations and stock performance).

The distinction between effectiveness and meaningfulness can be framed in formal psychometric terms: An external, criterion-based approach to psychological measurement confounds estimates of *reliability* (the repeatability of a measurement) and *validity* (the psychological meaning of a measurement). This is because an external criterion is being used to determine both. As a result, it is easy to think we are measuring one (validity) when, in fact, we are measuring other (reliability).

In their article, Vul et al. observe that some cognitive neuroscience investigations of social processes have confused reliability and validity. If we use an external criterion (such as looking at negative and neutral pictures) to identify those voxels showing a significant change in blood oxygenation level dependent (BOLD) response, and then we correlate the negativity of the slides to the magnitude of this change, it is ambiguous whether the correlation reflects a reliability coefficient or a validity coefficient. We all know, however, that this sort of mistake is not limited to functional imaging studies of social phenomena. Examples can be found here and there throughout the imaging literature. And as the MMPI example shows, this mistake in psychological measurement is not limited to functional imaging studies per se. No matter what the measurement domain, the practice of using some external criterion (be it performance on a task or self-reports) to identify signal from noise blurs the boundaries between estimating what is reliable (and potentially effective) and what is valid (and psychologically meaningful), leading to confusions when interpreting changes in BOLD signal.

## **LESSON 2: THE ELUSIVE NATURE OF “ERROR”**

Notions of reliability and validity can be phrased in terms of this equation:  $X = T + E$ . Albeit somewhat simple, this little equation might be the single most important equation in any field that attempts to infer something mental from the measurement of something physical. It states that observed scores (the actual numbers generated during measurement, denoted as  $X$ ) have two parts: that which is consistent and repeatable (the “true score” variance, denoted as  $T$ ) and that which is random and not repeatable (“error”, denoted as  $E$ ). Reliability is the proportion of variance in observed scores ( $X$ ) that is accounted for by consistent variance ( $T$ )—it is the proportion of variance that is

repeatable on another occasions. Validity refers to the psychological meaning of reliability variance. Now that we have measured something consistently, what is it? What are the numbers a measurement of?

As every observed measurement includes some signal ( $T$ ) and some noise ( $E$ ), the trick is figuring out which part is which. There are various ways to accomplish this. A set of measurements taken at one point in time (Time 1) can be correlated with those exact measurements taken at another point in time (Time 2). This is called test–retest reliability. When a group of respondents take the MMPI twice, separated by some interval, the correlation coefficient that results is interpreted as a test–retest (or stability) coefficient. The assumption from classical measurement theory is that only consistent variance can correlate with something else because error is random and will fluctuate from Time 1 to Time 2. Half of the measurements at Time 1 can be correlated with the other half taken at Time 1 on the assumption that the two halves are equivalent.<sup>1</sup> This is called split–half reliability. When responses from half the MMPI items are correlated with responses from the other half within a single group of respondents, the resulting correlation coefficient is interpreted as a split–half reliability (or consistency) coefficient. The split–half logic can be extended to examine the consistency of responses across all possible combinations of MMPI items to compute coefficient alpha, a very common estimate of internal consistency.

No single form of reliability is more “true” than any other. The various ways of computing reliability estimate different aspects of repeatability or consistency. Furthermore, no form of reliability tells us what is being measured—a set of observations (i.e., responses to MMPI items, BOLD responses in voxels) could be measuring one thing

(homogeneous variance) or one hundred things simultaneously (heterogeneous variance). Reliability only tells us that we are measuring consistently, either within a measurement moment (split-half reliability or internal consistency) or across measurement moments (test-retest reliability).

Once we know how much consistent variance is contained in  $X$ , we can then ask what this variance refers to in psychological terms. This is the question of validity. Just as there are many different types of reliability, so too are there many different types of validity. For example, criterion validity refers to a measurement's ability to predict or estimate some other measurement. When we ask whether some items on the MMPI predict a certain kind of diagnosis under certain circumstances, we are asking a question about criterion validity. When we ask if BOLD responses in a particular set of voxels predict a behavior or self-report, we are asking a question about criterion validity. In these cases, the resulting correlation coefficient (or  $t$  test, which can be transformed into a correlation coefficient) is interpreted as a validity coefficient. Criterion validity does not really tell us much about the mind—it does not tell us why a set of items or voxels predicts what they do, only that they do predict it. The question “why?” is answered with an estimation of construct validity. Construct validity refers to a measurement's ability to assess an idealized psychological process or state and only that process or state (Cronbach & Meehl, 1955). When we ask questions about the function of any localized set of voxels, we are asking a question about the construct validity of those BOLD responses. In principle, construct validity can only be established by showing that a measurement is associated with an interlocking set of variables (a nomological net) that is dictated by theory; it can never be established with a single validity coefficient.

Furthermore, construct validity must show that a measurement is consistently related to a set of criterion measures (i.e., it must show convergent validity) and that it is specific to that construct (i.e., it must show discriminant validity).

Just as there is no “true” measure of reliability, there is no “true” measure of validity. A measurement can have criterion validity without having construct validity. Scale 8 on the MMPI can differentiate groups of patients without measuring the essence of schizophrenia (albeit we now know that there is no essence to schizophrenia or to any other mental illness). Similarly, BOLD responses in the amygdala can be effective in predicting ratings of distress when viewing negative pictures, but this does not mean that the amygdala’s function is to compute or represent distress.

Now, as it turns out,  $X = T + E$  is an overly simple equation. And this makes things confusing, particularly when it comes to figuring out what kind of error we are dealing with. In fact, the equation should really be re-written as  $X = (T + E_c) + E_r$ , where  $E_c$  refers to systematic error variance (or the variance of something that we do not care about that is inadvertently measured with some consistency) and  $E_r$  refers to random error variation that is not repeatable.  $T + E_c$  refers to the variance in an observed measurement ( $X$ ) that is repeatable and is estimated with some form of reliability analysis. Because there are two kinds of error that can be found in observed measurements ( $X$ ), separating signal ( $T$ ) from noise ( $E_c$  or  $E_r$ ) can be even trickier than was first assumed.

There are instances when random error ( $E_r$ ) inadvertently masquerades as true score variance (noise is being treated as signal). In such cases, we run the risk of overestimating an observed correlation between a BOLD response and its criterion relative to its true population value. This is what it means to “capitalize on chance.” This

is the risk that is inherent when we estimate reliability and validity with the same data coming from the same sample using the same (or strongly related) statistical comparisons. And as Vul et al. illustrate, this risk is real and potent with functional imaging data. When the method used to select measurements (whether items or voxels) is not independent of the subsequent tests performed on those measurements, random error (variance that exists only in this measurement moment) creeps into the estimate of  $T$  and can correlate with the dependent variables of interest, inflating the magnitude of the statistical result. So, if we use an external criterion (such as looking at affectively negative and neutral pictures) to identify voxels showing a significant change in activation, and then we correlate the negativity of the slides to the change in BOLD response in those voxels, the resulting correlation coefficient will likely be inflated from its true population value. And if we then interpret this correlation as a validity coefficient, we have almost certainly capitalized on chance. This kind of mistake cannot be dealt with by making corrections for multiple comparisons. Such comparisons cannot protect us from the fact that when the exact same measurements are taken at another point in time or on another sample, the magnitude of the correlation will shrink. This kind of shrinkage is a well-known problem in regression analysis (because regression coefficients are, not coincidentally, mathematically equivalent to correlation coefficients of one type or another).

Furthermore, the risk of capitalizing on chance exists whenever test–retest reliability is low. For example, we could observe a large coefficient alpha (strong internal consistency) or strong split–half reliability in a set of measurements that have lower stability across time (low test–retest reliability) because something unexpected or

undesired is influencing all the responses within a single measurement moment. If this is true, then it is not clear that we can avoid capitalizing on chance by splitting a data set in half, so that half of the data from all participants is used to determine reliability (i.e., to select voxels for analysis) and the other half of the data can be used to estimate validity (i.e., to determine what the BOLD activity in those voxels refers to or means), as Vul et al. suggest. As discussed further in Lesson 3 (later in this article), the measurements from the first and second halves of a study (from the same participants) are not, strictly speaking, independent and therefore cannot really be used for cross-validation. To estimate the degree of shrinkage in a correlation coefficient that is inherent in inadvertent instances of capitalizing on chance, it is necessary to split a sample in half, using one set of participants for voxel selection and another set for validity estimation. This procedure is followed routinely for statistical procedures such as discriminant function analysis (in which a subset of items or variables is chosen from a larger available set and then weighted to optimally predict some psychological outcome). True replication can only be achieved with different sets of participants.

Estimating reliability and validity separately can reduce the risk of capitalizing on chance, but it does not protect against spurious correlations (relationships that exist because of some third, irrelevant cause). Spurious correlations can occur when stable, non-random errors ( $E_c$ ) are inadvertently estimated as part of  $T$  (when there is some consistent variance in our measurements that we do not care about or are not interested in). As long as the systematic error is shared between the observed measurements and their criterion, then estimates of validity become spuriously inflated because the magnitude of the correlation coefficient reflects something other than what we believe it

does (such as method variance; Campbell & Fiske, 1959). This causes us to make mistakes about what physical measurements mean in psychological terms. For example, the MMPI uses a true–false response format. If we were to give a group of respondents another test that required them to make true–false judgments, scores on the two tests would be more highly correlated because they share a response format and this would be mistakenly estimated as true score variance. Similarly, if we use ratings of negative pictures to identify those voxels showing a significant change in activation, and then we take any other measurement (such as a self-report measure of momentary distress) that uses a similar rating scale, both will have similarly high (or low) correlations with the change in BOLD response in part because they share a similar response format. To separate the systematic variance into that which we care about and that which we do not care about, we must use multiple measures of a construct and analyze the data with structural equation modeling (e.g., Barrett & Russell, 1998).

### **LESSON 3: WITHIN-SUBJECT DEPENDENCIES**

Now, it might seem as if we can avoid spurious correlations by making sure that our estimates of construct validity involve the use of a third measure that is relatively independent and free of method variance or other unwanted shared features. With the MMPI, perhaps the construct validity of the schizophrenia scale items could have been quickly confirmed by correlating patients' scores on the schizophrenia scale with another objective criterion, such as ratings of hallucination severity provided by the diagnosing clinicians. But this kind of a criterion would not completely resolve the spurious correlation problem. The validity coefficients would probably be high, for sure, but the risk remains that the correlations could be inflated by systematic error variance. This is

due to the fact that, in reality, the third measure (the additional criterion) was not truly independent from the original criterion used to select the items in the first place. Even asking the patients themselves to report on their severity of hallucinations would be problematic as a criterion for Scale 8, because responses to two different scales from the same subject would not be statistically independent from one another. A similar problem is evident when estimating the reliability and validity of BOLD activity with data from different measures sampled from the same participants. If we use an external criterion (such as looking at negative or neutral pictures) to identify those voxels showing a significant change in activation, and then we take any other measurements (such as a self-report measure of momentary distress, reaction times to judge the slides, or even trait ratings of neuroticism) and correlate them with the change in BOLD response in those voxels, there is always a risk that the resulting correlation coefficient will be inflated from its true population value if all the data are taken from the same sample of participants.

Multiple measurements sampled from the same individuals are never truly independent from one another. Within-subject dependencies exist even when the observed measurements are supposed to be measuring very different psychological domains in different modalities. Since the mid 1990s, behavioral scientists have been statistically modeling within-subject dependencies (using hierarchical linear modeling or multilevel regression modeling; e.g., Laurenceau, Barrett, & Pietromonaco, 1998).<sup>2</sup>

The within-subject dependencies are even more complicated in neuroimaging experiments. Neurons are nested within columns that are nested within voxels that are nested within brain areas that are nested within individual brains that also produce the

behavioral estimates that are measured as criteria. Furthermore, the BOLD signal from different voxels that are close in proximity to one another are made even more dependent as a function of preprocessing procedures (such as smoothing). The fact that there are dependencies in the multiple measurements taken from a single individual means that measures share some variance over and above what is caused by the psychological construct of interest, which in turn inflates the magnitude of correlation coefficients (be they reliability or validity estimates). For example, some factor that is irrelevant to the psychological domain of interest (such as hormonal changes related to circadian rhythms or to menstrual cycles in women, or blood volume changes related to hydration) could lead a host of measurements to have spuriously high correlations. In the behavioral realm, these data dependencies have been shown to matter quite a bit, and the final statistical results that are reported often depend on whether these dependencies are modeled or not. It seems critically important, then, to better deal with these dependencies in a statistical sense when trying to determine the psychological meaning of a physical measurement, especially when those measurements are based on the brain (for a start, see Lindquist & Gelman, 2009).

## **CONCLUSION**

Without taking a stand on all the articles discussed in Vul et al. , in large part because I have not read them all myself, I have tried to show with a simple discussion of classical measurement theory that Vul et al. tell an important cautionary tale about the pitfalls of translating measurements of the brain into knowledge about the mind. More important, I have tried to show that these pitfalls are correctable: (a) don't estimate reliability and validity of the BOLD response simultaneously with the same statistic on

the same data, (b) ensure that error (whether random or systematic) is not mistakenly estimated as true score variance through replication, and (c) model the dependencies in measurements that are collected on the same individuals or at least consider those dependencies when interpreting your data.

If there is one true adage in psychology, it is that past behavior is a great predictor of future behavior. Over the last 70 years, the MMPI has been the subject of tremendous study and scientific effort to fix the glaring problems surrounding its inception. Items have been replaced to reflect changes in diagnostic practice. The test has been renormed so that it reflects a broader population than just those people living in Minneapolis and the surrounding area in the 1930s. There has been an attempt to take a more theoretically driven (deductive) approach to test construction. And, of course, the various forms of reliability and validity have been estimated separately for the various subscales of the test. Every year, the MMPI contributes to the millions of clinical assessments that are performed in hospitals, mental health clinics, and research labs in many countries around the world. There is no doubt in my mind that imaging research is following the same path. In 70 years, when someone writes the history of how measurements of the brain eventually translated into knowledge about the mind, psychologists and neuroscientists will marvel at how far we have come.

***Acknowledgments***—Thanks to both the members of my lab and Moshe Bar’s lab for stimulating discussion of the Vul et al. (2009) article and some of the commentaries. I especially thank Maria Gendron, Jennifer Fugate, Yang-Ming Huang, Kristen Lindquist, Spencer Lynn, Elizabeth Kensinger, and Julie Norem for their comments on an earlier

draft of this article. Preparation of this manuscript was supported by the National Institutes of Health Director's Pioneer Award (DP1OD003312), grants from the National Institute of Aging (AG030311) and the National Science Foundation (BCS 0721260; BCS 0527440), and a contract with the Army Research Institute (W91WAW).

## REFERENCES

- Barrett, L.F., & Russell, J.A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*, 967–984.
- Burish, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, *39*, 214–227.
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.
- Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
- Duncan, T.E., Duncan, S.C., Okut, H., Strycker, L.A., & Hix-Small, H. (2003). A multilevel contextual model of neighborhood collective efficacy. *American Journal of Community Psychology*, *32*, 245–252.
- Hathaway, S.R., & McKinley, J.C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Personality*, *10*, 249–254.
- Laurenceau, J.P., Barrett, L.F., & Pietromonaco, P.R. (1998). Intimacy as a process: The importance of self-disclosure and responsiveness in interpersonal exchanges. *Journal of Personality and Social Psychology*, *74*, 1238–1251.

Lindquist, M.A., & Gelman, A. (2009). Correlations and multiple comparisons in functional imaging: A statistical perspective. *Perspectives on Psychological Science*, 4, xx–xx.

Raudenbush, S.W., & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1–17.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4, xx–xx.

<sup>1</sup>The issue of whether neural responses at the beginning and end of a study are equivalent (reflecting the same psychological process) is somewhat questionable (given issues like habituation and repetition suppression). But a discussion of this issue goes beyond the scope of the present article.

<sup>2</sup>In the mid-1980s, behavioral scientists were measuring data dependencies in other nested data structures, such as when children who are nested within classrooms that are nested within schools (Raudenbush & Bryk, 1986), and when measuring people who are nested within families that are nested within neighborhoods (Duncan, Duncan, Okut, Strycker, & Hix-Small, 2003).