Feature Review
Building a Science of Individual Differences from fMRI
Julien Dubois¹,* and Ralph Adolphs¹

To date, fMRI research has been concerned primarily with evincing generic principles of brain function through averaging data from multiple subjects. Given rapid developments in both hardware and analysis tools, the field is now poised to study fMRI-derived measures in individual subjects, and to relate these to psychological traits or genetic variations. We discuss issues of validity, reliability and statistical assessment that arise when the focus shifts to individual subjects and that are applicable also to other imaging modalities. We emphasize that individual assessment of neural function with fMRI presents specific challenges and necessitates careful consideration of anatomical and vascular between-subject variability as well as sources of within-subject variability.

From the Group to the Individual
Brain imaging with blood oxygen level-dependent functional magnetic resonance imaging (BOLD fMRI) has been used extensively since the early 1990s to understand generic aspects of brain function, typically by averaging data across individuals to improve the signal-to-noise ratio (SNR). The statistical benefits of averaging across subjects have also been leveraged in group comparisons, for example in studies of clinical populations. However, these studies have historically fallen short of a proper characterization of brain function at the level of the individual. While the importance of a fully personalized investigation of brain function has been recognized for several years [1,2], only recent technological advances now make it possible. For example, there are advances already at the acquisition level, such as higher field strength and faster acquisition, which have led to substantial SNR improvements [3]. Attempts at interpreting individual subject fMRI measurements have become a major focus in the past 5 years or so, partly driven by the rise of ‘resting-state’ fMRI (Box 1). There is interest in examining individual differences in relation to healthy aging [4,5], personality [6], intelligence [7,8], mood [9] and genetic polymorphism [10]. On the clinical side, there are considerable efforts to use fMRI to classify individual subjects as patient or control ([11,12]; reviewed in [13,14]), to select treatment [15], or predict future outcome ([16]; reviewed in [17]).

Several issues arise when the focus shifts from group averaging to the comparison of the statistics of individual subjects. The issues can be framed in terms of key concepts from behavioral research on individual differences, namely validity and reliability. Validity asks whether the individual differences we measure with BOLD fMRI really reflect what we intend to measure. One specific concern for validity is whether we are indeed comparing functionally homologous regions across subjects. Another is whether we are indeed comparing neural function because BOLD fMRI only provides an indirect measure of neural activity. Reliability, by contrast, asks whether a finding is stable in the face of variations that should not matter. Reliability is notably hindered by relatively well-understood noise sources such as motion and subject physiology, as well as by less well-understood ones such as neuro- or vasoactive substances that we might not

¹Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

*Correspondence: jcdubois@gmail.com (J. Dubois).
Box 1. Resting-state fMRI: The Workhorse of Individual Differences Research

Resting-state fMRI, or RS-fMRI, entails imaging subjects while they lie in the scanner doing nothing and trying not to think of anything in particular [156]. Spontaneous fluctuations in activity show reproducible correlations across brain regions; regions with correlated spontaneous activity also tend to be co-activated in task fMRI, thus establishing the relevance of spontaneous correlations (functional connectivity) to brain function [157,158]. RS-fMRI has now established itself as a leading approach to the study of brain organization [156].

A major reason for the widespread adoption of the RS-fMRI approach is its minimal requirements. Most subjects can lie quietly in the scanner for 5 or more minutes [159] for more advanced analyses, up to 100 minutes of data may be required for best results [160]. Of course, careful investigations have pointed out ‘details’ that influence the data collected during the resting-state, such as whether the subject’s eyes are open or closed [161], whether they are completely awake [124], what they are actually thinking about during the run [162–164], or what task they performed immediately preceding the run [165]. Changes in the strength and directionality of functional connections have been described during runs in the same session, but also at much faster timescales (seconds to minutes) during a run (reviewed in [166]). Functional connectivity can also be estimated from task fMRI data, usually following the removal of stimulus-evoked activity [72], and shares about half its variance with the functional connectivity estimated from RS-fMRI data [167]. The remaining half of unshared variance could complicate the interpretation of individual differences in functional connectivity [94].

It remains the case that RS-fMRI provides the easiest functional data to collect and aggregate, across subject populations and sites [148], as is now done in several large efforts, such as those of the Human Connectome Project [149,168]. Individual differences research requires large sample sizes (see also main text) and thus RS-fMRI, despite imperfection on several fronts, is likely to remain the workhorse of individual differences fMRI research for the years to come (for a glimpse of the future, see Box 4).

properly take into account. Both validity and reliability can be enhanced through the use of readily available tools; we review the latest advances with respect to these issues and provide some specific recommendations for how the field can best advance (Figure 1, Key Figure).

In addition, we discuss important statistical considerations on the path to a science of individual differences from fMRI. Though correlation analysis (between individual fMRI-derived statistics and other measures of the same individuals) is overwhelmingly used in the literature, it is subject to overfitting and findings do not always generalize to other samples; instead, the out-of-sample predictive value of an fMRI-derived statistic with respect to another individual difference measure must be established. The typical sample size for fMRI studies (n = 10–50) also needs to be scaled up to n > 100 for individual differences research; larger sample sizes not only increase statistical power in general but also allow more complex models to be fit.

Validity: Are Individual Differences Attributable to Brain Function?

A Common Space for Mapping Function

How can we tell whether individual differences in metrics such as BOLD activations or functional connectivity are actually related to differences in the underlying neural activity or communication, respectively? One ubiquitous problem arises in matching different brains such that functionally meaningful comparisons across subjects are possible in the first place. In a typical fMRI analysis pipeline, both structural and functional data from individual brains are spatially warped to a common anatomical space. The most widely used common space [18] is the MNI152 atlas, to which subjects’ brains are anatomically warped via a volumetric transform ([19] for review). Such registration is appropriate for subcortical structures which are inherently volumetric; by contrast, the cortex is a 2D structure and volumetric alignment does not properly align folding patterns across subjects. Although switching to cortical folding-based inter-subject alignment [20,21] has been shown to somewhat reduce functional mismatch [22,23] (but see [24]), this has not yet become common practice. There are several reasons for this, from the burden of generating accurate cortical surfaces for each brain to the additional complexities of working with surface data (which until recently were not handled well by leading fMRI analysis software). Recent improvements to Freesurfer’s automated cortical surface reconstruction pipeline [25], together with the release of a new file format that combines surface and volumetric data (Connectivity Informatics Technology Initiative, CIFTI) and software for visualization and analysis
(the Connectome Workbench), could foster wider adoption of surface analysis. Nevertheless, there is still no guarantee that functionally similar vertices (see Glossary) correspond spatially across subjects after such alignment: individual differences may thus invalidly arise from incorrect alignment of function across subjects.

A Multimodal Common Cortical Topography
Watching a movie with a rich variety of visual, auditory, and social percepts elicits time-locked brain activity that is correlated across subjects in many brain regions (inter-subject correlation, ISC) [26]. After cortical folding-based inter-subject alignment, the cortical mesh can be further warped to maximize inter-subject correlations during movie viewing [27]. In the absence of brain activity that is time-locked to a shared external stimulus, it is also possible to instead improve the vertex-wise match of functional connectivity patterns across subjects [28] using resting-state data (see also [29,30]). A multimodal surface matching (MSM) framework [31] was recently introduced that not only performs similarly to, and faster than, state-of-the-art algorithms based solely on cortical folding [21], but that can also flexibly accept other types of data (e.g., fMRI) to further inform registration (Figure 2A). The framework can, for example, accept any combination of geometric (shape-based), myelin [32], task activation, retinotopy [33], and functional and structural connectivity features. In theory, the combination of all this information in the MSM framework should yield the best possible inter-subject alignment, with the constraint that neighboring vertices must remain neighbors. However, the optimal set of modalities to include, the associated cost functions, and the relative weights assigned to each modality remain under investigation, and further validation of an improved alignment of function will be required before widespread use can be prescribed [31].

From a Common Cortical Topography to a Common Representational Space
Another recently introduced technique, ‘hyperalignment’, goes one step further, ignoring topological constraints altogether and using only representational geometry across subjects – in other words matching multivariate spatial patterns of activity (Figure 2B) [34]. As before [27], the fMRI data used in this study [34] were obtained in response to a full-length movie. The selection of voxels/vertices that are fed into the hyperalignment algorithm is the sole spatial constraint; data from each subject are eventually projected into a common space that can violate topology. The original implementation of the technique relied on pre-selecting a region of interest (ROI) [34]; a more recent whole-brain version of hyperalignment uses a searchlight centered on each vertex, then combines the resulting transformation matrices into a large, whole-brain transformation matrix that can be used to project the cortical data from each individual subject onto a common representational space [35]. Ongoing work is extending the hyperalignment algorithm to make use of resting-state data, instead of movie data, in an effort to maximize applicability to extant datasets. In theory, whole-brain hyperalignment provides the most thorough correction for anatomical variability. Because hyperalignment is based on an invertible transformation of the data, individual differences are not lost at any stage of the process. Nevertheless, the validity of individual differences measured in the common space is not assured because individual differences of interest may also be mixed into the transformation matrix for each subject. We would encourage researchers to explore hyperalignment for their datasets, and to report results obtained with and without hyperalignment, such that we can accumulate further evidence for its most appropriate application. In general, we would recommend that investigators try more than one approach to alignment, and report all of them, so we can see which might work best for which kinds of questions.

ROI Analysis
Many of the problems with alignment can be largely circumvented if a functional ROI is isolated in each individual subject. Functional localizers have traditionally been used to improve validity [36]: in an independent MRI run, a task designed to activate a ROI is performed, and the ROI is represented as a brain-activity pattern. The dissimilarity of two patterns corresponds to the distance between their points in the representational space. The geometrical arrangement of multiple points in representational space is a characteristic of the representation. Retinotopy: mapping of visual input from the retina to neurons in the visual cortex. Techniques exist for non-invasive retinotopic mapping with fMRI.
Run: in the context of a fMRI experiment, a run designates an uninterrupted collection of fMRI data, resulting in a sequence of successive volumes.
Statistical parametric map: the result of a univariate (voxel-by-voxel, or vertex-by-vertex) statistical analysis of fMRI data, using any standard statistical test; it consists of the values of the statistic at each voxel or vertex.
Vertex: the cortical surface is represented as a tessellation of triangles. The vertices of the surface mesh are the vertices of its triangular tiles.
defined in each individual from a statistical parametric map (SPM) as one of the clusters that exceeds a given threshold. However, the choice of a statistical threshold in single subjects is often a thorny issue; recent approaches can avoid arbitrary thresholds by using mixture models that fit the noise in each subject’s data [37] or by using multiple spatial scales to define clusters of activation (threshold-free cluster enhancement, TFCE) [38]. Another issue is of ensuring that the individually-defined ROIs are indeed well-matched across subjects, a problem for which approaches such as the group-constrained subject-specific (GCSS) algorithm [39] provide elegant solutions. Functional localizers have a long and successful track-record and can provide functional ROIs in individual subjects with only a few minutes of scanning, for processes ranging from face perception [40] to theory of mind [41,42]. However, they quickly become inefficient if one wants to map many different processes in a study [43]. At present we would encourage investigators to continue using well-vetted localizers when studying specific psychological

Figure 1. This pipeline may not fit all situations, given the multidimensionality and multi-component nature of fMRI data. (A) Data $Y_{raw}$ are acquired while a subject undertakes a particular task (movie with stars). Minimal preprocessing is applied (e.g., HCP pipelines [25] shown in the box to the right). (B) Subcortical areas are volumetrically warped to match the MNI template, while the cortex is registered to a surface template using cortical folding patterns or a multi-modal surface matching approach. A choice is then made between conducting a region-of-interest or whole-brain analysis. Several options for group-informed definitions of functional parcels in individual subjects are available. Finally, hyperalignment can be performed to project data into a common representational space. (C) The preprocessed data $Y_{preproc}$ are modeled as a weighted linear combination of known signal (except for RS-fMRI) and confounds of no interest, plus unmodeled noise. All parameter estimates are normalized for vascular differences before further analysis. A statistic $S$ (possibly multidimensional) is derived for this subject. (D) More than 100 subjects participate in the experiment, possibly with slight variations (e.g., at two different sites, black and gray scanners). (E) The fMRI-derived individual statistic is used together with confound variables to predict an individual measure of interest $IV$, for example the subject’s IQ (prediction) or patient/control (classification). The same analysis is conducted without the fMRI-derived statistic. Performance for the full model is compared to performance for the null model, using a permutation test, to establish the unique predictive power of the fMRI-derived statistic. Text in gray: methods awaiting further validation.

Key Figure

Proposed Generic Analytical Pipeline for Individual Differences Research in fMRI

(A) Acquire and preprocess fMRI data $Y_{raw}$

- 1. GDC (large gradient nonlinearity)
- 2. Realignment
- 3. Slicetime correction if TR>2s
- 4. Field inhomogeneity correction
- 5. Conjugation to T1w
- 6. Grand mean scaling
- 7. Temporal filtering

(B) Map data to a common space

- Volumetric registration to template (e.g., MNI152)
- Folding-based registration to template (e.g., averaged for cortex
- Multi-modal surface matching (MSM) to template (e.g., HCP group average)

(C) Derive fMRI statistic $S$

$Y_{preproc} = \beta_{S}X + \beta_{null}X + \epsilon$

(Modal fit of $S$ or other constrained basis set)

(D) Repeat many times ($n>100$)

(E) Establish predictive value of $S$ for an Independent variable of interest $IV$

$IV = f_{null} (S, site, motion, other individual measures)$

$IV = f_{full} (S, site, motion, other individual measures)$

Areal classifier

Random forest

Statistical analysis

Classification (cross-validated)

False alarm rate

Permutation statistics

Prediction (cross-validated)

Observed $IV$

$AUC_{FULL} > AUC_{NULL}$

Permutation statistics

$R^2_{FULL} > R^2_{NULL}$

Permutation statistics
processes; in other contexts, emerging methods for single-subject functional parcellation from resting-state data and other modalities (Box 2) should be considered. Pending further validation against known function, these methods may replace functional localizers altogether in the not-too-distant future.

The Ever-Lurking Plumbing Issue
Differences in vasculature cause differences in the hemodynamic response. A review of all the potential effects of vasculature differences is outside the scope of this article (see [44–46]); we focus here instead on existing solutions at the level of acquisition and data analysis that are rarely implemented but may prove crucial for valid individual differences research.

Capturing Response Shape Variability
The hemodynamic response function (HRF) is a model of the BOLD activity triggered by a neural event of infinitesimal duration. Most fMRI analyses assume a fixed-shape HRF (the canonical HRF).
Box 2. Functional Parcellation of the Brains of Individual Subjects

It is now well established that resting-state functional connectivity (RSFC) can extract networks in the brain that subserve shared psychological functions [145]. Accordingly, there has been interest in using RSFC to define functional parcels or ROIs in the brains of single subjects. There are two schools of thought regarding functional parcellation: those which allow the parcels to overlap spatially, and those that prefer to tile the brain with non-overlapping parcels. The first school typically uses independent component analysis (ICA) to decompose the RS-fMRI signal into statistically independent non-Gaussian sources (spatial maps) via a linear and instantaneous mixing process corrupted by additive Gaussian noise components [149]. The second school typically uses some variation of a clustering or region growing algorithm that results in parcels that have homogeneous patterns of whole-brain connectivity [148, 170–178], or an algorithm that detects transitions in whole-brain RSFC [174, 175].

There are two approaches to establishing a one-to-one correspondence between the parcels in different subjects that enhance validity beyond whole-brain alignment techniques. One is to perform RSFC-based parcellation in individual subjects independently with no prior information [148, 171], then use some clustering or matching algorithm to establish one-to-one correspondence between parcels; this can prove a rather difficult problem, with for example splitting of parcels. The other, preferred approach is to start from a set of group-level parcels and discover their instantiation in individual subjects. The initial group parcellation is of course critical, and may come from analyzing averaged or concatenated individual data, or from a previously published result. Projecting it into individual brains is typically done using dual regression in the case of IC components [178], and different approaches have been proposed for other parcellation schemes – for example a template-matching procedure which assigns each voxel or vertex from an individual subject to a parcel (e.g., [177, 201]), or an iterative assignment procedure as in [178].

Including information from multiple modalities has proved beneficial to refine inter-subject whole-brain alignment, as demonstrated in the MSM framework [31]. Using multiple modalities can also be instrumental in parcellating an individual brain, as demonstrated recently: after defining a multimodal group parcellation from the HCP data (including architecture, such as thickness and myelin content, task fMRI, RS-fMRI connectivity, and RS-fMRI derived topographic maps), a machine-learning approach can be used to find the instantiation of each group-defined parcel in single subjects, correctly accounting for variations in anatomy [200].

However, the shape of the HRF is well known to vary across subjects and brain regions [44, 47]. Factors such as vasodilatory signaling, blood vessel stiffness, neurovascular coupling delay, venous transit time, and the time constant of autoregulatory feedback contribute to this variability [48, 49]; these factors are particularly affected by aging and disease [50]. Inferences about individual differences in BOLD response magnitude are not valid when the true shape of the HRF differs across subjects. To account for these variations, the use of multiple, orthogonal basis functions to more accurately model the HRF has been proposed (reviewed in [51]). Popular constrained basis sets include the canonical HRF plus its temporal and dispersion derivatives [52], or a basis function set based on singular value decomposition [53]. Releasing all constraints, the finite impulse response (FIR) basis set has one free parameter for every time-point following stimulation for every event type modeled [54]. A caveat of increasing the number of free parameters is the concurrent increase in variance of the HRF produced, as well as the cost to compute it.

When the shape of the HRF is fixed, the amplitude of the BOLD response is easily compared across subjects on the basis of a single parameter estimate. Modeling HRF shape faithfully complicates the comparison of response amplitude across subjects: the BOLD response is now represented across several parameters, which must be jointly compared. Some solutions have been proposed for this joint comparison, but none is widely accepted yet [51, 55]. Recent work combining estimation of the HRF shape with detection of activation in the same optimization seems a step in the right direction [56, 57], although these approaches currently assume that HRF shape is similar across subjects in a given brain region, which may not be a valid assumption when investigating individual differences.

HRF shape variability is problematic even for resting-state fMRI because differences in response shape across subjects can result in differences in functional connectivity estimates. With current analytical practices of low-pass filtering the data (considering only fluctuations slower than 0.1 Hz), the effects of small differences in response shape should be small [45]. However, as investigators start looking at higher-frequency fluctuations [58], HRF shape variability may limit
validity in interpreting individual differences. Methods are emerging to deconvolve the spontaneous signal [59,60] and thus possibly account for these variations. In our view, further development of how best to quantify and incorporate HRF variability should be of the highest priority for individual differences research because all subsequent measures depend on getting this right in the first place.

Accounting for Vascular Differences across Subjects: Calibration and Normalization
The BOLD signal is a function of changes in cerebral blood flow (CBF), cerebral blood volume, and the cerebral metabolic rate of oxygen (CMRO2) [61]; it also depends on the baseline physiological state (hematocrit, oxygen extraction fraction, blood pressure). Both hemodynamic coupling and baseline physiology differ across subjects [45,62]. Two main approaches have been put forward to try to control for differences in these parameters across subjects: calibration and normalization.

The calibrated fMRI technique has been around for many years, and can in theory control for differences in both baseline physiology and hemodynamic coupling across subjects. The technique has not seen widespread adoption, chiefly because it is difficult to implement (e.g., requiring concurrent measurement of BOLD and CBF and inhalation of CO2 [63,64]) and is still imperfect [65].

Several approaches have been proposed to normalize for differences in baseline physiology across subjects. Many rely on a companion scan. For example, the BOLD response to hypercapnia, induced through administration of CO2 [66] or by using a breath-hold challenge [67], can be used as a normalization factor (Figure 3A). Alternatively, whole-brain venous oxygenation levels can be measured with a special pulse sequence and used to normalize the BOLD response [68]. A more easily applicable option is to use the amplitude of low-frequency fluctuations in resting-state fMRI data (RS-ALFF) [69,70] as a normalization factor; indeed RS-ALFF reflects naturally-occurring variations in cardiac rhythm and in respiratory rate and depth [71], and approximates the BOLD response to a hypercapnic challenge (Figure 3A). In fact, one does not even need to acquire a separate resting-state scan. In the same way that functional connectivity can be derived from the residuals of a general linear model (GLM) for task-based fMRI data [72], the amplitude of low-frequency fluctuations in the residuals of task-based fMRI data (GLMres-ALFF) can also be used to rescale the BOLD signal change; this ‘vascular autorescaling’ (VasA) technique was even shown to outperform RS-ALFF-based normalization [73] (Figure 3B). Using data from the same fMRI run, as VasA does, is also desirable because it avoids contamination of the data with noise from a separate run.

While all these methods were developed with the aim of improving group statistics by reducing inter-subject vascular variance, they should be considered for valid assessment of neural individual differences (Figure 3C) [74]. As with any manipulation of the data, there are potential caveats: calibration and normalization methods may remove individual differences of neural origin, for instance if baseline blood oxygenation were linked to neural activity [62]. Before such techniques are routinely applied, more examples of their successful application will be necessary, ideally vetted by an independent measure of neural function [74].

Reliability: Individual Differences or Unmodeled Noise?
Once validity is maximized (inasmuch as current technology allows), it is also crucial to ensure that individual differences measured with fMRI are not merely attributable to unaccounted-for noise in the measurements. The reliability of fMRI has been inspected closely in recent years [75–77]. Of course there is no such thing as ‘the’ reliability of fMRI because different derived statistics are differentially affected by noise in the raw data and thus have different reliabilities [78] (for an overview of studies addressing the reliability of specific fMRI-derived measures see Table S1 in
The reliability of individual fMRI measures can be derived from extant literature, details of the experimental procedures [79] and of the preprocessing applied to the raw data [80,81]. For these reasons, we believe that it is crucial to provide an assessment of reliability in each and every study that looks at individual differences, under the specific pipeline used for the analysis. At present, individual differences in fMRI are always studied with respect to an independent measure of the same individuals. The reliability of interest is that of the relationship between the fMRI-derived statistic and the independent variable (see also [78,82]). There are several flavors of reliability which are relevant to individual differences research, with direct analogs in behavioral testing. Test–retest reliability quantifies how variable the established relationship is in the same sample of subjects, under the same conditions (stimulus, scanner, time of day, analysis) at an appropriate time-interval (Figure 4A). It is also important to ensure that the relationship is robust to the exact preprocessing performed on the raw data, which can be conceived of as ‘inter-rater’ reliability (Figure 4B). In addition, because we are interested in brain function rather than the low-level properties of a given stimulus, the relationship must also be robust to the exact experimental conditions, known in psychology as ‘parallel forms’ reliability (Figure 4C). The relationship should further hold for a different sample of subjects (from the same population) (Figure 4D), and across scanners.

Though many measures of reliability have been proposed over the years [75], some directly addressing the ratio of inter- to intra-subject variance [83], a predictive framework inspired by machine-learning is best suited to establish the reliability needed for individual differences research, as we discuss further below (see ‘Choosing Prediction over Correlation’). Reliability is ensured when a relationship discovered in a sample of subjects at site 1 using stimulus 1 and...
Figure 4. Reliability for Individual Differences Research. (A) Test–retest reliability. The same subjects are tested with the same stimuli in the same scanner, at an appropriate time-interval given the function of interest. (B) Inter-rater reliability. The same data are preprocessed in slightly different ways which are thought to be interchangeable (e.g., using physiological regressors or ICA to remove physiological noise). (C) Parallel forms reliability. The same subjects view slightly different stimuli which are thought to involve the same brain processing. (D) Out-of-sample reliability. A different set of subjects undergo the same experiment in the same scanner, and the data are analyzed in the same way. (E) Putting it all together for individual differences research: A result (e.g., relationship between fMRI-derived statistic and neuropsychological score) should be reliably obtained for different subjects at different sites, possibly using slightly different stimuli and preprocessing steps.

Sources of Within-Subject Variance We Acknowledge and Strive To Correct For
There are several sources of within-subject, inter-session variance that are well known, and which fMRI researchers strive to eliminate at acquisition time and/or correct for during preprocessing of the data [84,85]. Scanner-related noise, artifacts, and drift are unavoidable but are fairly simple to address through artifact rejection and temporal filtering. We briefly review here two main sources of within-subject variance which are more problematic and the state-of-the-art in addressing them: subject motion and subject body physiology.

Motion
Subject motion in the scanner leads to poor data quality, for obvious reasons: in fast echo planar imaging (EPI) sequences, motion not only disrupts spatial encoding but also disrupts the physical phenomena that the MRI signal relies on (e.g., resonance frequency, relaxation time between samples). Artifacts occur with frame-to-frame movements of a few tenths of a millimeter or less [86], meaning that they affect all datasets to some extent. The classical approach to correcting for motion artifacts has been to first estimate subject motion through rigid-body realignment of brain volumes, then to regress the calculated head motion parameters out of the fMRI timecourse to correct for any residual effects of motion [87]. Other researchers also include the global signal timecourse, the white matter timecourse, and the cerebrospinal fluid (CSF) timecourse as confound regressors [88]. This regression approach was recently shown to incompletely correct for motion artifacts in the context of resting-state fMRI analyses [86,89,90]. Additional corrections have been proposed, for example the removal of independent components related to motion [91], or the complete removal of motion-affected frames, known as scrubbing [86] (92) for comprehensive review.

Thoroughly correcting for motion artifacts is important to ensure test–retest reliability: a given subject may move more in one session than in another. Nonetheless, a discussion of motion would have been equally appropriate in the previous section on validity: some subjects are
generally more fidgety than others in the scanner (some have argued that this constitutes a trait in and of itself, with a neurobiological basis [93]). Motion arguably contributes more to inter-subject variance than to intra-subject variance (e.g., men tend to exhibit more head movements than women [90], older people move more than younger people [94], and people with autism move more than controls [95]). Motion artifacts have complex effects on fMRI statistics, and incompletely correcting for them can lead to erroneous conclusions in individual differences research [86,95–97].

Physiology
In a previous section we pointed out how differences in vasculature may affect the validity of inter-individual comparisons. To complicate matters further, breathing and heart rate affect fMRI measurements. For instance, motion of the chest wall with respiration results in magnetic field changes [38], normal alterations of the depth or rate of breathing lead to variations in arterial CO₂ [99] and subsequent bloodflow changes [100], and pulsatile bloodflow due to heartbeat leads to fluctuations in signal intensity in arteries, arterioles, and other large vessels [101].

A common approach to correcting for these artifacts is to collect independent measurements of the heart rate (with a pulse oximeter) and respiration (with a respiratory belt), and regress them out of the fMRI signal using one of several available algorithms (reviewed in [85]); however, these corrections are not routinely applied. Interestingly, it has been found that many of these model-based corrections actually reduce test–retest reliability for functional connectivity analyses [102,103]. This has been interpreted to mean that a large fraction of functional connectivity reflects basic physiological signals [102] (another validity issue, e.g., [104]). A more optimistic take is that these model-based approaches still need some tweaking – either the models are not physiologically accurate or the measurements of heart rate and respiration are suboptimal (too noisy, or not measuring the variable of interest). Further development of the models (e.g., [105]) and of new and better MR-compatible measurement devices such as a continuous blood pressure monitoring device is needed [85].

Other approaches only rely on the fMRI data without the need for separate measurements (reviewed in [85]; see also [106–108]). Most of these are based on decomposing the data into components, for example, using independent component analysis (ICA) (but see [109]). Some of these methods have been shown to improve reliability [110]. However, a fair comparison with model-based physiological regression has not been conducted yet – typically, decomposition-based denoising techniques target a wider range of artifacts, including motion.

In decomposition-based methods, identifying noise components is not always trivial [111]; some components may be a mixture of signal and noise, and there is a risk of ‘throwing the baby out with the bathwater’. Automated algorithms, such as ICA-FIX [108], rely on prior manual classification of components, and are therefore not immune to this criticism. A promising new approach to teasing apart signals of neural origin from artifacts was recently introduced: using multi-echo EPI [112] it appears possible to distinguish a genuine BOLD effect from artifacts by looking at the timecourse of MR decay. The technique could lead to improvements in both the validity and reliability of measurements [113]. Finally, its combination with recent developments in simultaneous multi-slice acquisition [114] could make it a viable imaging protocol in terms of temporal and spatial resolution [115]; we are thus eagerly awaiting further validation and improvements of this technique (see also [116] for a related approach).

Other Sources of Intra-Subject Variance
Another Physiological Rhythm: Vasomotion
Vasomotion refers to the spontaneous changes in tone of blood vessels, independently of heart rate and respiration. Vasomotion leads to low-frequency oscillations in the BOLD signal [117].
There is some evidence that vasomotion may be localized to specific regions of cortex [118], making it a potentially serious confound. As with previous artifacts, vasomotion may differ systematically across subjects (validity). It is as yet unclear how vasomotion should be dealt with.

**Baseline Physiological State**

Baseline physiological state was already discussed in the validity section, but baseline physiology is also variable for a given subject. Going for a run an hour or less before scanning may lead to alterations of baseline physiology [119,120]. Anxiety or stress [121], extended exposure to high altitudes [122], recent sleep quality [123,124], or phase of the menstrual cycle (for women) [125] are all examples of poorly understood vascular and neural factors that may hinder the reliability of fMRI measurements if they are not properly measured and incorporated into a full model [126].

**Neuromodulators and Vasoactive Substances**

The effects of caffeine on fMRI measurements have been studied extensively. While it was once held that a dose of caffeine was a way to enhance the BOLD response [127], subsequent studies have muddied this picture [128,129]. Many complex effects of caffeine have been reported over the years, such as enhanced linearity of visually evoked BOLD responses [130]. The complexity likely stems from the double action of caffeine on neural activity and on hemodynamics ([131] for review).

Several other routinely consumed substances such as alcohol, nicotine, illicit drugs, and prescription medication, and even foods and food supplements, also have complex neural and vascular effects on fMRI measurements, and these are beyond the scope of this review. Using arterial spin labeling (ASL) and fMRI to obtain a better understanding of these effects, a field of research known as pharmacological MRI [132], may be a further important ingredient towards a reliable science of individual subjects from fMRI.

**Further Considerations for an fMRI Science of Individual Differences**

**Choosing Prediction over Correlation**

Currently the only way to interpret a fMRI-derived statistic is to relate it to another individual measure in the same set of subjects, such as their age, gender, test scores indexing aspects of intelligence and personality, or other measures of behavior (Table S2). By far the majority of fMRI studies of individual differences use correlation analysis to establish such a relationship. Correlation analysis relies on in-sample population inference and does not directly ensure the generalizability of the established relationship to out-of-sample individual subjects [133,134]. Shifting to a predictive framework is necessary to ensure generalizability and to interpret fMRI-derived statistics at the individual subject level [17,135] (for a more in-depth discussion of adopting a predictive machine-learning inspired framework, and the proper use of training, validation, and test datasets, see [136]). To fully control for remaining confounds, fMRI-derived statistics should be included in a full model alongside other potential predictors, and the unique predictive power of fMRI features should then be assessed through their selective removal from the model (as demonstrated in [16]) (Figure 1E).

**Increasing Sample Size**

It is now widely recognized that the small sample sizes routinely used in fMRI studies ($n = 10–50$) have low statistical power; given current reporting practices, this leads to inflated estimates of effect size [137,138] (Figure 5C) and thus poor replicability (see Box 3 for an overview of replicability-enhancing practices). fMRI research is not the first to face this challenge [143]. The current heuristic for studies using correlation analysis is a sample size of at least 100 (see [139] and Figure 5A for a more quantitative recommendation). In addition, replication in an independent sample is a worthy precaution, and this should be a requirement for studies that are not based on a strong prior hypothesis.
Larger samples are also beneficial in a predictive framework: a larger number of examples guards somewhat against overfitting. As to replication in an independent sample, it is a defining feature of a predictive framework. Model selection is conducted on the basis of training data, and independent testing data is reserved until the model is finalized. The fully trained model is tested once, and once only (as in machine learning competitions [140]), on the test data to establish generalizability (but see the recently introduced ‘reusable holdout’ algorithm for a viable alternative [202]).

Sample sizes in excess of 100 are still difficult to achieve for a small research group funded by a typical research grant; thus, despite this recommendation, underpowered correlation studies without internal replication will continue to be published. While it might seem that the cumulative output from many underpowered studies should eventually converge to a reliable conclusion through meta-analysis, this is not the case because of the strong bias to publish only significant findings [136,141] (Figure 5C). The reporting of all results regardless of their significance in the null hypothesis significance testing (NHST) framework, which could be implemented through initial pre-registration to ensure the quality of the methods [142], would be a way for studies with small sample sizes to contribute unbiased information for meta-analysis.

To achieve larger datasets in fMRI, data collected at different institutions can be aggregated in a shared online resource [144], as pioneered by the 1000 Functional Connectomes Project [145]. Several data-sharing platforms are currently available (see [146]), for example openfMRI [147]. For the accumulation of data across sites, standardized procedures must however be implemented. Resting-state fMRI is an obvious candidate for data aggregation given minimal instructions and requirements but, as we noted earlier, ‘small’ details must still be carefully

---

**Figure 5. A Predictive Framework for a More Reproducible Science.** These plots are provided for illustrative purposes only. They are based on simple simulations described in the supplemental material. (A) The ‘cone of confidence’ for correlation analysis, as in [139]. The true effect size in the population is $r = 0.1$ (green), with SNR = 0.1. The light-gray traces (1000 shown, of 100 000 generated) are bootstrap samples from this population. As sample size increases, the range of correlation values observed in the samples decreases. In red, the 95% confidence interval for the null hypothesis $r = 0$. (B) Same representation for the $R^2$ of a leave-one-out prediction analysis, using the same simulated samples. The average of all bootstrap samples is shown in dark gray; it asymptotes to the true effect size (green line) for larger sample sizes. (C) Power of the correlation and prediction analyses (dark blue), for $\alpha = 0.05$. Proportion of significant results ($P < 0.05$) which overestimates the effect size (light blue). Prediction analysis is less likely to overestimate effect size than correlation analysis; it even tends to underestimate it, making it a conservative analysis.
Box 3. Increasing the Reproducibility of fMRI Research

The replicability of fMRI research has recently been called into question. Culprits for the lack of replicability are small sample sizes [138], analytical flexibility [179] which fosters p-hacking [180] without appropriate control for the overall risk of false positives [181], and the bias to publish only statistically significant results (also known as the file drawer problem [182]). This means that we are often building new research on fragile grounds (previous false-positive results), and thus wasting a large amount of resources [183]. A shift in research practices is necessary.

Reproducibility consists in obtaining the same results using the same code and data [184] (Figure I) (see also [183] for a slightly different terminology); it is a lower scientific standard than fully independent replication using a different dataset and similar analysis, but ensuring reproducibility in the first place is a surefire way to enhance replicability (although see [185] for a differing view). We note that failure to replicate a finding need not always imply that the finding was a false positive: a difference in outcome may also stem from any number of effects in execution or analysis that, unknown to the investigators attempting the replication, have actually changed the psychological process under investigation. Replication attempts help to establish the conditions under which a given finding is robust.

Following earlier recommendations for reporting fMRI methods [186], a recent white paper draft (from the OHBM Committee on Best Practices in Data Analysis and Sharing, also known as COBIDAS) is likely to be adopted by the imaging community. It identifies transparency, detailed reporting of methods, and important sharing of data and analysis pipelines (Figure I) as key ingredients for reproducible fMRI research. Data-sharing platforms were discussed in the main text [146]; there are also platforms to share code, and guidelines on how to use scripting and pipelining to foster reproducibility [183]. An example of how to implement these recommendations is to release all the data together with a virtual machine environment to allow others to execute identical analyses [126]. At this stage such diligence still requires a large amount of technical overhead and cannot be made mandatory. The enforcement of these recommendations will initially largely depend on journals following suit and requiring these ingredients for publication, or on reviewers taking them into account when evaluating their peers’ research. However, a growing set of resources may soon make this effort much more feasible and thus more widely adopted; in particular, we are closely following the efforts of the Stanford Center for Reproducible Neuroscience (http://reproducibility.stanford.edu) and of the Scientific Transparency Project (http://post.stanford.edu), both aiming at providing web-based platforms to generate reproducible workflows and leverage high-performance computing for the analysis of neuroimaging data.

![Figure I. From Reproducibility to Replicability: Sharing of Data and Executable Analysis Pipelines. Adapted from [184].](image)

controlled (Box 1). Although simple task fMRI paradigms (e.g., finger tapping [148]) would also be good candidates, they may be too inefficient in that they are restricted to a narrow set of processes. We view movie fMRI as particularly well suited for data aggregation, given its ability to capture richer representational information while better matching subjects’ mental states compared to the resting-state (Box 4), and at the same time preserving similarly minimal instructions and requirements.

Standardized procedures are of course easiest to implement in the context of large-scale projects such as the IMAGEN project (2000 subjects) [149], the WashU–UMinn Human Connectome Project (1200 subjects) [150], or the Cambridge Center for Ageing and Neuroscience (approximately 700 subjects) [151]. The latter two projects are aimed at acquiring state-of-the-art data and distributing it as a data-mining resource for fMRI researchers around the world. Projects such as these are akin to the accelerators used by particle physicists or the large telescopes used by astronomers: a few sites in the world

One promising class of stimuli, which affords tighter control on internal states than the resting-state while still being scalable (e.g., for aggregation of data across centers), comprises movies [26] and narrated stories [187]. These are naturalistic, highly-engaging stimuli which are very effective at driving brain activations and encompass a great diversity of mental states. Public datasets relying on story/movie data are in the making: for example, the studyforest project recently released a high-quality 7T fMRI dataset of subjects listening to the narrated movie ‘Forrest Gump’ (1994) [188]. Efforts to extract and distribute annotations that can be used in analyses have also begun [189]. We present here three main methods to study the representations triggered by such stimuli across subjects.

Inter-Subject Correlation

Because all subjects watch the same movie, the timecourse of stimulus-related activity in their brains should look similar [190]. As one would expect, activity in early sensory cortices is similar across subjects, whereas activity in association cortices is less similar [26]. Dynamic inter-subject correlation analysis in light of a rich featural description of the stimuli is a promising avenue of research to capture rich information about individual differences in brain representations.

Representational Geometry

Different stimuli are represented in the brain along several dimensions, and representational geometry denotes the arrangement of stimuli in that space and the distances between them [191]. While representational geometry and representational similarity analysis have so far been exploited mostly to assess different models of brain function [192–194], they can also be leveraged to compare representational geometry across individual subjects [195,196]. Representational geometry can be derived from movie or story data, provided that the events of interest are properly labeled [189].

Voxel-Wise Modeling

Voxel-wise modeling consists in using high-dimensional featural models of natural stimuli as encoding models to make sense of fMRI data [197]. When a model is found to satisfactorily predict the data out-of-sample, it can then be interpreted. A recent study using movies with annotated semantic features generated detailed semantic maps of each subject’s brain [198]. These maps represented those movie features that were most predictive of the activations observed at each voxel in the brain. These semantic maps capture a tremendous amount of information about the brain activity of each subject, and thus interrogating them with respect to individual differences is a very exciting endeavor. Such work is already well underway [199].

acquire the best possible data, and these data are subsequently probed (for many years) by the best analysts around the world.

Towards Normative fMRI Research?

A key concept in psychology is that of a ‘norm’, in other words the distribution of a score in the population of interest. Scores on psychological tests are usually standardized with respect to their distribution in a large normative sample (the size of which depends on the distribution of the score [152]), and this allows their direct interpretation without recourse to another measure. In the future, we may be able to interpret the fMRI-derived statistic of an individual subject in light of its distribution in a normative sample, provided that absolute reliability is achieved; this approach could lead, for clinical imaging, to a more biologically informed science of human neuropsychiatric disease [153,154] and a better basis for personalized medicine.

Concluding Remarks

Individual differences in brain function are key to understanding healthy differences based on personality, gender, age, or culture. They are also crucial for personalized medicine approaches to neuropsychiatric diseases. Recent technical advances have increased the sensitivity of functional MRI and set the stage for a characterization of brain activity at the level of momentary mental events in individual subjects. We now face key challenges of reliability and validity on the path to an fMRI-informed science of individual differences. We already have some tools to address these challenges: we propose a pipeline for the measurement and analysis of individual differences with fMRI as shown in Figure 1, on the basis of current knowledge (see Outstanding Questions). While interpretation of fMRI-derived measures currently relies on independent

Outstanding Questions

Will the newer methods that we describe here and recommend (highlighted in gray in Figure 1) stand the test of time?

Which of the concerns reviewed here (inter-subject alignment, hemodynamic variability across subjects, sources of noise) are the most important, and which can often be ignored in practice?

How do the many corrections we have reviewed here interact with one another? For example, does normalization become superfluous when the fMRI is properly modeled?

How long will BOLD-fMRI still be the norm for non-invasive functional imaging of the whole brain in awake humans? Are other more valid non-invasive imaging techniques conceivable in the future?

What other modalities can be combined with fMRI to provide a richer set of measures and enhance validity, besides physiological recordings? Options include eye-tracking, EEG, and fNIRS.

Should pre-registration be required for fMRI studies of individual differences below a minimum sample size?

What role will behavioral and genotypic data play in the future? Will these always be required, or can we explore individual differences in a data-driven way based on fMRI data alone?

Will fMRI-derived individual differences lead us to revise clinical diagnostic categories?
measures of behavior, psychological scores, and neuropsychiatric diagnosis, we are hopeful that in the near future they will stand on their own in light of normative data. No matter the strategy, increasing sample size dramatically is necessary, and data-sharing efforts together with standardized procedures and reproducibility-enhancing practices will be key to the process. Beyond the current emphasis on resting-state fMRI data, which is easy to perform and aggregate, the use of naturalistic stimuli that capture rich inter-individual differences in cognition is a direction worth exploring.

Acknowledgments
This work was funded in part by a NARSAD grant from the Brain and Behavior Research Foundation (to J.D.) and a Conte Center from the National Institute of Mental Health (NIMH) to R.A.). The authors thank Rebecca Schwarzlose, Michael Miller, Alex Huth, Swaroop Guntupalli and two anonymous reviewers for useful comments on the manuscript.

Appendix A Supplemental information
Supplemental information related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.tics.2016.03.014.

References
Trends in Cognitive Sciences


144. Van Horn, J.D. and Gazzaniga, M.S. (2013) Why share data? Lessons learned from the IMRDC. Neuroimage 82, 677–682
165. Tung, K.-C. et al. (2013) Alterations in resting functional connectivity due to recent motor task. Neuroimage 78, 316–324
181. Simmons, J.P. et al. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1593–1606