Beyond Increasing Sample Sizes: Optimizing Effect Sizes in Neuroimaging Research on Individual Differences

Colin G. DeYoung[1*], Kirsten Hilger[2*], Jamie L. Hanson[3], Rany Abend[4], Timothy A. Allen[3], Roger E. Beaty[5], Scott D. Blain[6], Robert S. Chavez[7], Stephen A. Engel[1], Ma Feilong[8], Alex Fornito[9], Erhan Genç[10], Vina Goghari[11], Rachael G. Grazioplene[12], Philipp Homan[13], Keanan Joyner[14], Antonia N. Kaczkurkin[15], Robert D. Latzman[16], Elizabeth A. Martin[14], Aki Nikolaidis[17], Alan D. Pickering[18], Adam Safron[19], Tyler A. Sassenberg[1], Michelle N. Servaas[20], Luke D. Smillie[21], R. Nathan Spreng[22], Essi Viding[23], and Jan Wacker[24]

Abstract

■ Linking neurobiology to relatively stable individual differences in cognition, emotion, motivation, and behavior can require large sample sizes to yield replicable results. Given the nature of between-person research, sample sizes at least in the hundreds are likely to be necessary in most neuroimaging studies of individual differences, regardless of whether they are investigating the whole brain or more focal hypotheses. However, the appropriate sample size depends on the expected effect size. Therefore, we propose four strategies to increase effect sizes in neuroimaging research, which may help to enable the detection of replicable between-person effects in samples in the hundreds rather than the thousands: (1) theoretical matching between neuroimaging tasks and behavioral constructs of interest; (2) increasing the reliability of both neural and psychological measurement; (3) individualization of measures for each participant; and (4) using multivariate approaches with cross-validation instead of univariate approaches. We discuss challenges associated with these methods and highlight strategies for improvements that will help the field to move toward a more robust and accessible neuroscience of individual differences. ■

INTRODUCTION

We are researchers who use neuroscientific methods to investigate psychological individual differences. Humans differ in their thoughts, feelings, and behaviors, and such variations among individuals are neither random nor entirely determined by the current situation. Many individual differences, described with terms such as traits, dispositions, attitudes, abilities, and symptoms, are relatively stable over time and are caused by a mixture of genetic and environmental influences (Polderman et al., 2015). We will use the term "traits" as a generic descriptor for all such constructs. Trait levels represent the probability of particular thoughts, feelings, and behaviors; they are relatively stable within individuals over time, reasonably consistent in rank order between individuals, and typically observable in many situations. Many trait measures are useful for predicting future behavior and important life outcomes (Soto, 2019; Deary, 2012). A long tradition of research on traits has focused on identifying their causes. Regardless of the proportion of the distal causes of a trait that are genetic versus environmental, trait differences must be caused proximally by differences in brain function, because the brain governs behavior and experience.

Neuroimaging research increasingly investigates associations of psychological traits with individual differences in brain structure and function (DeYoung et al., 2022; Hilger & Markett, 2021). Our aim in this article is to discuss how best to conduct neuroimaging research on psychological individual differences to achieve robust, replicable results, in light of recent debates about sample size (e.g., Spisak, Bingel, & Wager, 2023; Marek et al., 2022; Grady, Rieck, Nichol, Rodrigue, & Kennedy, 2021). If samples are too small, estimation of parameters will be imprecise, and the chance of detecting true effects as significant (i.e.,

[1]University of Minnesota, [2]Würzburg University, [3]University of Pittsburgh, [4]Reichman University, Herzlia, Israel, [5]Pennsylvania State University, [6]Ohio State University, [7]University of Oregon, [8]Dartmouth College, [9]Monash University, Melbourne, Australia, [10]Technical University, Dortmund, Germany, [11]University of Toronto, [12]Yale University, New Haven, CT, [13]University of Zurich and ETH, [14]University of California, [15]Vanderbilt University, Nashville, TN, [16]Takeda Pharmaceuticals, Cambridge, MA, [17]Child Mind Institute, New York, NY, [18]Goldsmiths, University of London, [19]Johns Hopkins University School of Medicine, Baltimore, MD, [20]University of Groningen, [21]University of Melbourne, Australia, [22]McGill University, Montréal, Québec, Canada, [23]University College London, [24]University of Hamburg
*Shared first authorship.

statistical power) will be low. Precision and power both depend crucially on sample size, and the fact that underpowered samples yield imprecise estimates has an important consequence that is often overlooked in neuroimaging research (Nebe et al., 2023): Not only are underpowered studies more likely (by definition) to yield false negatives (Type II error, failing to detect a true effect) than adequately powered studies, but they also yield a higher proportion of significant results that are false positives (Type I error, detecting an effect that is not true) because the estimated effects fluctuate widely around the true value. Thus, significant effects in small samples are often inflated (sometimes known as a Type M or magnitude error; Gelman & Carlin, 2014) or even completely spurious, which contributes to the prevalence of unreplicable results in scientific publications (Yarkoni, 2009).

It is often said that power is defined by three things: the significance criterion (α), the effect size, and the size of the sample; however, power is also crucially defined by the statistical model being used. Neuroimaging research in general is often underpowered (Szucs & Ioannidis, 2020; Poldrack et al., 2017), and this problem is amplified in research on individual differences because the statistical models used to estimate between-person effects require larger sample sizes than those for estimating within-person effects. When studying the function of the average brain, as in typical research on task-evoked brain activity, one compares neural activity in different conditions within the same individuals—in other words, participants serve as their own controls—and this reduces noise. For example, to achieve 80% power to detect a simple bivariate correlation of $r = .20$ as significant ($\alpha = .05$; two-tailed) requires a sample of 194 participants, which is considerably larger than would be required to detect the same effect size as a difference between conditions in typical within-person designs (e.g., 49 participants required for a paired-samples t test with $d = .41$, $\alpha = .05$, two-tailed).

Achieving sufficient power is additionally challenging in neuroimaging because many statistical tests are often conducted within a single analysis. In the common case of univariate brain-wide association studies, this entails using values from voxels, vertices, or parcels across the entire brain and testing for associations of psychological variables with each neural value independently. Such multiple testing reduces power by effectively requiring a more stringent significance (α) threshold for each individual test to maintain the same overall α. (Note, however, that alpha should not be corrected simply by dividing by the number of tests, as in Bonferroni corrections, because the correlated structure of the data—e.g., in spatially adjacent voxels—makes many of the tests nonindependent; other approaches, such as controlling the false discovery rate, are needed.) Conducting many tests also increases the temptation to engage in selective reporting, in which some tests or analyses are not reported, obscuring the true burden of multiple testing by reporting results that are nominally but not actually significant. This kind of selective reporting within neuroimaging studies increases publication bias—the tendency to report only significant effects—and has contributed greatly to the proliferation of false positives and the resulting replication crisis (Stanley, 2005).

All of this entails that sample sizes for individual-differences research in neuroimaging need to be considerably larger than samples sizes traditionally used in this field. The big question is, "How much larger?" Recently, an influential article argued that "thousands" of participants are necessary for "studies of the associations between common inter-individual variability in human brain structure/function and cognition or psychiatric symptomatology" (Marek et al., 2022, p. 654). This important study demonstrated the limited power of common neuroimaging approaches to individual-differences research and served as a clarion call to develop more robust approaches for identifying associations between traits and neural variables. Here, we attempt to answer that call by discussing potential solutions to this problem that might not require thousands of participants. We argue that appropriate methods may allow replicable neuroimaging research on individual differences with hundreds of participants.

The key question we consider is how to increase expected effect sizes, because larger effect sizes require fewer participants to achieve the same statistical power. Fundamentally, the motivation for claims that thousands of participants are necessary comes down to effect size. Using three very large MRI samples, Marek and colleagues (2022) examined brain-wide associations of structural parameters and resting-state functional connectivity with performance tests of cognitive ability and questionnaire measures of features of psychopathology and observed that the largest 1% of replicable univariate effects were between $|r| = .06$ and .16. It is worth emphasizing that this means 99% of the replicable effects were even smaller than .06. If all expectable between-person effect sizes were indeed this low, then it might be true that samples in the thousands were always necessary, not only when conducting many statistical tests (although of course this makes the problem more acute), but even when conducting more focused studies that are not "brain wide." However, the observations of Marek and colleagues do not necessarily generalize to all individual-differences research in neuroimaging.

Expected effect sizes cannot be generalized from one set of methods to all others. Marek and colleagues (2022) drew conclusions based on analyses using some of the most common methods in the field, but these nonetheless represent only a small subset of available methods and some of them are suboptimal. Here, we discuss alternative methods, focusing on four categories of methodological improvement designed to increase expected effect sizes and, therefore, to increase power independently of sample size: (1) theoretical matching between tasks and trait constructs, (2) improving measurement reliability,

(3) individualization of measurement for each participant, and (4) pivoting from univariate to multivariate analytic approaches. Our aim is not merely to discuss arguments made by Marek and colleagues (2022), although we do address some of them directly. Rather, our aim is to consider the broader issue of improving neuroimaging methods for individual differences research.

TRAIT-RELEVANCE: MATCHING FMRI TASKS TO TRAIT CONSTRUCTS

The first strategy we recommend to increase effect sizes concerns the modality of neuroimaging assessments. Many neuroimaging studies of individual differences have relied on structural MRI or resting-state fMRI data, but effect sizes may be larger in studies using appropriate fMRI tasks (Finn, 2021). Structural and resting-state data have been widely used in individual-differences research for multiple reasons. For one, it is assumed that any trait could potentially be related to parameters derived from these imaging modalities because they do not target any specific psychological content or processes. This makes it possible to study many psychological traits in relation to the same structural and resting-state data, whereas data from any particular task seem likely to be relevant to a more limited set of traits. In addition, brain structure and resting-state functional connectivity are sometimes assumed to be more trait-like than task-evoked activity because they are independent of the situational demands of any specific task (Hilger & Markett, 2021) and have been found to demonstrate adequate retest reliability (Zuo & Xing, 2014), whereas task-induced neural activity often has poor retest reliability (Elliott et al., 2020). However, the meta-analysis of Elliott and colleagues (2020) shows that neural activity during some specific tasks does have adequate retest reliability.

The possibility of reliable signals from task fMRI is consistent with the excellent reliability of various performance-based cognitive tests (e.g., IQ tests) and with the common conceptualization of traits as tendencies to respond in consistent ways to specific classes of stimuli (DeYoung et al., 2022). Choosing the right task can provide the kind of stimuli that are particularly relevant to the processes underlying the trait in question, leading to differential associations between activation in different fMRI tasks and personality traits (Hardikar et al., 2024).

The potential relevance of structural and resting-state measures to any trait is both a strength and a weakness. Neither modality directly assesses an aspect of brain function that is transparently relevant to most psychological traits. (In addition, brain structure is ultimately relevant to psychological traits only to the degree that it influences function, so that the increment in trait prediction provided by structural variables over functional ones may not be very large; Ooi et al., 2022.)

In contrast, task-based fMRI may induce brain states directly relevant to the trait of interest whenever that trait is theorized to reflect variation in psychological processes like those involved in the task. Identifying appropriate tasks requires at least some minimal amount of theory regarding the processes underlying the trait in question, and we encourage researchers to consider the theoretical background of any trait they are studying (DeYoung et al., 2022). Theoretically informed research can potentially increase effect sizes by identifying likely associations among traits, underlying psychological processes, and the brain systems that support them.

That task-based fMRI may lead to meaningful between-person effects has been shown in studies considering task-induced neural activation (e.g., Tetereva, Li, Deng, Stringaris, & Pat, 2022) as well as in studies focused on functional connectivity during tasks (e.g., Greene, Gao, Scheinost, & Constable, 2018). In fact, task-based brain–behavior associations are consistently stronger than resting-state-based associations in cross-validation studies of very large samples (Chen et al., 2022; Ooi et al., 2022; Feilong, Guntupalli, & Haxby, 2021; Sripada, Angstadt, Rutherford, Taxali, & Shedden, 2020; Greene et al., 2018). For example, working memory is well-established as a cognitive function involved in general cognitive ability, and several of the studies cited in the previous sentence show that associations of intelligence with neural functioning during working-memory tasks are stronger than associations with resting-state or structural data.

Marek and colleagues acknowledged that fMRI task data may yield larger effect sizes than resting-state or structural data. Notably, in their Extended Data Figure 3, they reported that the correlation between cognitive ability and activation of the dorsal attention network during a working-memory task was .34. They dismissed this finding by characterizing working-memory performance as a "confound" that needs to be controlled for in analysis (yielding a much smaller correlation of .14), but the plausible causal arrangement of the three variables—activation of the dorsal attention network, working-memory performance, and general cognitive ability—is not one of confounding. Working-memory performance is a relatively stable trait strongly correlated with general cognitive ability and thought to be a crucial process facilitating that ability (Kovacs & Conway, 2016). Therefore, it should act as a mediator between neural activity and general cognitive ability, rather than as a confound. A correlation of .34 can readily be detected in samples considerably smaller than a thousand. We would not recommend assuming that task-based effects in general will be this large for the purposes of power calculations, but any effect larger than .16 is larger than all of the replicable univariate structural and resting-state effects reported by Marek and colleagues (2022) and can be detected in hundreds of participants if the multiple testing burden is not too high. We suspect that such effects may be relatively common, given the right pairings of traits with fMRI tasks.

Despite our enthusiasm for task-based fMRI, we want to be clear that we are not suggesting individual-difference

researchers should abandon structural and resting-state neuroimaging studies. The remaining three categories of strategies we endorse are applicable to all neuroimaging modalities.

IMPROVING MEASUREMENT RELIABILITY

After deciding what question to address, an important strategy for increasing effect sizes is to improve the reliability of both trait and neural measures, because the joint reliability of two measures sets an upper bound on the possible strength of association between them (Nikolaidis et al., 2022). Many recent articles have discussed reliability in the context of human neuroscience (e.g., Haines, Sullivan-Toole, & Olino, 2023; Nebe et al., 2023; Nikolaidis et al., 2022), so our discussion here is not intended to be exhaustive. However, we will highlight some opportunities for improving reliability for three different types of assessment: neural measures, behavioral tasks, and questionnaires. (In addition, our suggested methods in the next section also tend to increase reliabilities of neural measures.)

The reliability of neural parameters can be improved by a variety of means, both in fMRI data acquisition and in subsequent data processing and analysis. For acquisition, we recommend the use of multi-echo sequences. Multi-echo fMRI significantly improves whole-brain temporal signal-to-noise ratio and reduces signal dropout in typically problematic regions along the ventral-anterior surface of the brain (Lynch et al., 2020; Kundu et al., 2017). Furthermore, it enables a biophysically based removal of noise from fMRI data sets during preprocessing because of the known echo-time dependence of the BOLD signal. This has been shown to improve reliability substantially (Lynch et al., 2020; Kundu et al., 2017). More costly but also effective is simply increasing the amount of fMRI data for each task (or resting-state scan) that is collected for each participant (Cho, Korchmaros, Vogelstein, Milham, & Xu, 2021; Noble et al., 2017). In data processing, reliability can be improved by modeling the hierarchical structure of neural parameters, using machine learning methods, or generating aggregates from multiple measures (Blair, Mathur, Haines, & Bajaj, 2022; Schubert, Löffler, & Hagemann, 2022).

In task-based fMRI, one pitfall to avoid is exclusively selecting ROIs for individual-differences research by using group-level fMRI contrasts to identify regions where a task significantly activates the brain relative to a control condition (DeYoung et al., 2022). The problem with this approach is that group-level contrasts ensure identification of ROI where a sufficient number of individuals show activation for the group average to be significantly different from the control condition. This approach risks identifying ROIs with less individual variation in brain activity relative to other brain regions. The most important regions for a given trait may be ones that are not significant at the group level, precisely because different brains respond differently to the task in those regions. Because the reliability of measures of individual differences depends on variability, focusing on robust within-persons effects works against reliability at the between-person level, a phenomenon described as the "reliability paradox" (Hedge, Powell, & Sumner, 2017). Beyond using group-level contrasts to select ROIs, researchers can select ROIs from functional networks or anatomical regions indicated as relevant for the trait of interest by theory or prior empirical evidence, or they can use other relevant individual-difference variables that are not involved in the focal hypothesis to identify regions where those predict neural variables (e.g., using performance during a scanned working memory task to identify regions where performance predicts activation, then using activation levels in those regions to predict other behavioral traits, such as intelligence; DeYoung, Shamosh, Green, Braver, & Gray, 2009).

The reliability paradox applies not only to neural variables but also to behavioral variables extracted from tasks. Some tasks, such as those that make up standard intelligence tests, have been designed specifically to optimize the assessment of individual differences, but many experimental tasks used in neuroimaging have not. Instead, those tasks have usually been designed to minimize between-person variability to aid in studying typical function as the group average in within-person designs. Researchers should investigate the degree to which tasks used in neuroimaging are reliable as measures of individual differences and take steps to improve them (Blair et al., 2022). Sometimes better options are already available; for example, new versions of the Stroop and Flanker tasks have recently been designed to improve measurement of individual differences, and they show excellent internal consistency and retest reliability (Burgoyne, Tsukahara, Mashburn, Pak, & Engle, 2023).

For questionnaires, reliability is most often measured as internal consistency (i.e., the degree to which individual items correlate with each other), but, for constructs that are conceived as relatively stable features of individuals, retest reliability (i.e., the degree to which a sample's rank order is consistent over time) is an even more relevant metric (Nikolaidis et al., 2022). Another important consideration is that measures may differ in their reliability across the range of the variables they are assessing, which requires more sophisticated methods to detect. Although Marek and colleagues reported adequate reliability for their measure of psychopathology, analysis of the same data using item response theory showed that it was inadequate for assessing individual differences in the lower range of the scales, where most healthy individuals score (Tiego et al., 2023). Measures should be investigated to make sure they are appropriate for the population being studied.

In addition, although it may be tempting to use short forms of questionnaires, longer measures generally have better validity and reliability (Credé, Harms, Niehorster, & Gaye-Valentine, 2012). Using multiple informants is also

valuable, as they provide incremental validity and reduce the biases introduced by individual raters. This principle can also be extended to the use of multiple measurement modalities for the same trait (e.g., questionnaire and behavioral task), although identifying adequately parallel measurements across modalities can be challenging (Joyner & Perkins, 2023). Whenever multiple measures of the same variable are collected, measurement can often be improved by modeling constructs as latent variables representing the shared variance of multiple indicators.

INDIVIDUALIZATION OF MEASURES

The third strategy we recommend to increase effect sizes is to improve measurement through a family of procedures known as individualization. A serious measurement challenge arises from the uniqueness of every human brain. Standard procedures to align individual brains to a common anatomical template cannot handle variations in which an anatomical feature is present or absent. For example, in ACC, people vary in whether they have only one sulcus or two, and the second sulcus (known as the paracingulate sulcus [PCS]), if present, can be very short, or it can extend the entire length of ACC. Warping a brain with a PCS to a template without one (or vice versa) causes inaccuracy in subsequent comparisons across individuals because the presence or absence of PCS has important consequences for the brain's functional and structural organization (Amiez, Wilson, & Procyk, 2018; Fornito et al., 2008). Such structural idiosyncrasies can be taken into account to improve measurement (Miller, Voorhies, Lurie, D'Esposito, & Weiner, 2021; Voorhies, Miller, Yao, Bunge, & Weiner, 2021).

Not only do different brains differ in structure, but also the localization of functions relative to the brain's anatomical landmarks differs from person to person. This means that, even if structural alignment were perfect, comparing brains based merely on location would remain suboptimal. Neuroimaging studies often use canonical brain atlases or parcellations to identify ROI and define brain networks (Moghimi et al., 2022), and these schemes often rely in part or entirely on functional information to parcellate cortex (which is appropriate given the primacy of function for psychology). However, using the same standard parcellation for all participants means that the parcel boundaries will not precisely reflect the relevant functional boundaries for any participant (Chong et al., 2017; Mueller et al., 2013).

To overcome this problem, we recommend methods that individualize standard parcellations by optimizing the boundaries of each parcel for each participant. These include group prior individualized parcellation (GPIP; Chong et al., 2017) and multisession hierarchical Bayesian modeling (MS-HBM; Kong et al., 2019). Both of these methods have been found to increase effect sizes in individual-differences research, as compared with using the same atlas or parcellation scheme without individualization (Sassenberg et al., 2023; Kong et al., 2021). Notably, individualized parcellation is preferable to using dual regression following an independent components analysis (an earlier strategy to deal with the same problem) because, unlike dual regression, individualized parcellation retains the benefit of canonical atlases in allowing comparison of the same parcels across individuals and samples (DeYoung et al., 2022).

Individualization can be taken even a step further than shifting boundaries of parcels to identifying different collections of voxels that encode the same information in different brains. Even within a given brain region that is well aligned through GPIP or MS-HBM, information may be encoded differently in different people. A technique known as hyperalignment identifies different sets of voxels with similar patterns of neural activity for each participant and treats them as the relevant neural unit of analysis. Hyperalignment increases effect sizes relative to other methods (Feilong et al., 2021; Haxby, Guntupalli, Nastase, & Feilong, 2020), considerably more than the increase generated by individualized parcellation, although both methods can be used together.

An older method of individualization is the use of functional localizers, which are fMRI tasks that reliably activate a particular brain system and thus can be used to identify a specific region or regions activated by that task in each participant before correlating parameters derived from those regions with measures of psychological traits. This is a powerful method for theory-driven research and may be especially valuable for focal hypothesis testing that maximizes power by minimizing the multiple-testing burden. However, it is also limited in that it can only identify regions that are relevant to the particular task used, rather than being able to individualize the whole brain. Individualized parcellation using the methods described above can match the functional localization of neural activity by tasks, even when the individualized parcellation is derived from resting-state data (Uddin et al., 2023; Kong et al., 2021; Chong et al., 2017).

Finally, people vary not only in the spatial layout of brain function but also in the timing of brain function as measured by fMRI. An early fMRI study revealed marked variability in the hemodynamic response function (HRF) across individuals and brain regions (Aguirre, Zarahn, & D'Esposito, 1998). Although fMRI research typically assumes a canonical HRF, methods are available for estimating the idiosyncratic shape of the HRF for each participant separately (Singh, Wang, Cole, Ching, & Braver, 2022; Singh, Braver, Cole, & Ching, 2020). Variability in the HRF is related to variation in vasculature more generally, and methods to estimate and control for such differences have been shown to improve fMRI signals dramatically (Kazan et al., 2016).

Up to this point, we have focused on individualizing neural data, but it is worth noting that individualization of psychological measurements is also sometimes possible, for example, using computerized adaptive testing

(Wainer, Dorans, Flaugher, Green, & Mislevy, 2000) or fitting computational models to each participant's trial-by-trial task data. The latter is useful in part because different individuals can employ different strategies when performing the same task. For example, studies using learning tasks can estimate the degree to which participants engage in model-free versus model-based learning (e.g., Kool, Gershman, & Cushman, 2017).

MOVING FROM UNIVARIATE TO MULTIVARIATE APPROACHES

Our fourth suggestion is to increase effect sizes by transitioning from univariate to multivariate analytic approaches. Multivariate analyses involve using multiple variables to predict the criterion variable and follow naturally from the premise that psychological traits are determined by many neural parameters. These variables could all be of the same type (e.g., activation values of individual voxels, as in multi-voxel pattern analysis, or parameters derived from resting-state EEG; Thiele, Richter, & Hilger, 2023), or they could involve parameters from multiple measurement modalities, such as structural and fMRI, from which parameters can be combined in the statistical model (Rasero, Sentis, Yeh, & Verstynen, 2021; Jiang et al., 2020).

It is unsurprising that multivariate models using many predictors can yield better overall predictions than univariate models using a single predictor. In a commentary on the work of Marek and colleagues (2022), Spisak and colleagues (2023) showed that multivariate brain-wide association studies effects are larger than univariate effects and can be replicable even when identified in smaller samples. Although they suggest these samples can be as small as 75, we would not recommend samples that small, given the resulting lack of precision of parameter estimates. Indeed, Spisak and colleagues' analyses show that multivariate effects in such small samples may be replicable in the sense of producing a significant effect in the same direction, but the size of that effect is often substantially different. Nonetheless, they showed that samples of 300–500 generally yield multivariate effects reasonably replicable in magnitude as well as significance, when they are cross-validated to prevent overfitting.

Tervo-Clemmens and colleagues (2023) rejected Spisak and colleagues' (2023) conclusion, but their exchange makes it clear that the argument between the two research teams is based largely on a terminological disagreement about the proper way to use the phrase "out of sample." Before considering their different uses, it is important to understand the difference between more traditional statistical approaches that optimize the fit of the statistical model in all of the data at once to best explain variance within one sample, and predictive approaches, such as those in machine learning, in which the model is fit in one sample (or subset of one sample), and then the parameters from that model are applied in another sample (or subset of the same sample). In the predictive approaches, the data are divided into a training set, in which the model parameters are identified, and a test set, in which the model parameters identified in the training data are used to predict the criterion variable in the test data, to determine the generalizability of the model. The effect size reported is from the test data (often computed as the correlation between the predicted values of the criterion and the observed values) because it is less biased by overfitting than the effect size from the training data. The predictive approach can be implemented in three ways (Thiele et al., 2024; Yarkoni & Westfall, 2016): (1) A sample can be split into training and test sets in an iterative manner, repeatedly fitting a model in one part of the sample and then testing it in the rest of the sample (as in *k*-fold cross-validation), with the reported effect size being the average result from the test sets across iterations. (2) A sample can be split just once into test and training sets, without iteration, such that the test set is never used as part of the training set (lock-box validation). (3) For the most stringent test of model generalizability, the model parameters identified in the complete original sample can be used in an entirely new sample, differing in various characteristics (external cross-validation). In the predictive approach, finding a significant effect in the test data is not considered replication because replication requires an entirely new independent sample beyond both the training and the test data.

Marek and colleagues (2022) and Tervo-Clemmens and colleagues (2023) use the phrase "out of sample" to refer to any test of a model in data not used to train the model (including the use of a different subset of the same sample as the test set), whereas Spisak and colleagues use "out of sample" to refer to testing the model in an entirely new sample, excluding cases where the test set is a subset of the same sample. These different uses of "out of sample" diverge in *k*-fold cross-validation (which was used by Spisak et al., 2023) because the training and test data are both subsets of the same sample. This effect size is "out of sample" in Marek and colleagues' sense, but "in sample" in Spisak and colleagues'. From our perspective, Spisak's perspective more directly addresses the question of how many participants are needed in total to identify effects with sufficient precision that they will be replicable in subsequent studies in other samples. In other words, how large must a single sample be to yield accurate results after cross-validation in that sample? The exchange between Spisak and colleagues and Tervo-Clemmens and colleagues makes clear that replicable results considerably larger than $|r| = .16$ can be achieved using multivariate prediction in samples in the hundreds rather than the thousands, as long as effect sizes are estimated using appropriate cross-validation procedures, rather than in data for which the model was optimized.

This conclusion is supported by the existing literature, especially in relation to cognitive ability. One study of the same sample used by Spisak and colleagues found that neural activation during various tasks predicted general cognitive ability in multivariate models with $r \approx .30$ (Sripada et al., 2020). Similarly high multivariate correlations between task connectivity and cognitive ability were found in another large sample (Chen et al., 2022). Multivariate effects can be even larger when combined with individualization approaches. In the same sample analyzed by Spisak and colleagues (2023), Feilong and colleagues (2021) were able to predict cognitive ability using estimates of functional connectivity based on hyperalignment, with average multivariate effect sizes of $r = .53$ for task data and $r = .44$ for resting-state data.

When transitioning to multivariate analysis, it is important to keep in mind that not all multivariate methods provide equally good prediction or equally generalizable results. Some yield larger effects than others (Spisak et al., 2023). Even with cross-validation, effects are likely to be larger if multiple modalities of imaging data are employed (Schulz, Bzdok, Haufe, Haynes, & Ritter, 2024). Multivariate approaches are often understood as wholly exploratory rather than hypothesis driven, and this impression is reinforced by their association with brain-wide analyses. However, multivariate methods need not be applied brain-wide and can easily be used in theoretically driven research, in which multiple parameters are derived, for example, from particular brain regions or systems of interest. When multivariate approaches are designed to facilitate interpretable insights rather than only maximizing prediction, they can also be used to test hypotheses and provide evidence for or against specific psychological models (Thiele et al., 2024). For example, a study of 257 participants used many of the strategies we recommend and identified significant multivariate correlations (ranging from .18 to .47) of a trait measure of autobiographical memory with functional connectivity between two theoretically identified brain regions (hippocampus and temporal pole) and the default network (Setton, Mwilambwe-Tshilobo, Sheldon, Turner, & Spreng, 2022).

Conclusion

Our proposed strategies for increasing effect size are summarized in Table 1, organized according to the typical sequence of neuroimaging research. Our list is not exhaustive; anything that improves the reliability and validity of measurement should increase expected effect sizes in research linking neural variables to individual differences in psychological traits. Careful application of these methods may provide a path to identifying more robust, generalizable, and scientifically or clinically useful brain–behavior relationships in samples with hundreds of participants. The debate about how large samples sizes should be is currently prominent in the field, but focusing exclusively on increasing sample sizes ignores other ways to achieve higher statistical power. Increasing effect size contributes independently to power and is often easier and cheaper than increasing sample size.

Evident in the research we reviewed above is that multivariate effect sizes are generally larger than univariate effect sizes, and shifting to multivariate methods with cross-validation and samples in the hundreds seems likely to be very effective for increasing effect sizes, statistical power, and replicability. However, smaller replicable effects can also yield important conceptual insights, and univariate research will continue to be valuable, especially for theory-driven hypothesis testing. For univariate research where the focal analysis involves only a single statistical test, we suggest using a sample size of at least 200, provided there is reason to expect an effect size of at least $r = .20$ (and such expectations should never be based on a single study). Note, however, first that this is just a loose heuristic that does not consider the exact statistical model (such as the inclusion of standard covariates like sex and head motion) and, second, that only a small fraction of published neuroimaging research on individual differences includes only a single focal test. If univariate research involves conducting multiple tests, then the sample size needs to be adjusted upward accordingly (or researchers must be confident that their expected effect sizes are even larger, which will probably be rare). We have mainly contrasted increases in sample size with increases in effect size as two ways to improve statistical power, but it is important to keep in mind a third way: reducing the multiple-testing burden by using theory to devise more focused hypotheses.

In conclusion, our suggestions for increasing effect size can help neuroimaging researchers to conduct robust research on psychological individual differences in situations where it is difficult to amass thousands of participants for a single study. Marek and colleagues' (2022) impressive work showed that common approaches to investigating univariate associations of traits with resting-state or structural MRI data are likely to require thousands of participants. If this were true for all approaches to neuroimaging research on psychological traits, it would be truly daunting, especially given the cost of neuroimaging. This issue is especially critical for those with less access to resources, such as early-career researchers or researchers in developing countries. (Such researchers can certainly benefit from the increasing availability of open data, but many research questions require new data acquired with specific acquisition parameters, during appropriate tasks, or within particular samples.) Fortunately, the situation we face as a field is not quite so dire. There are many available ways to improve our methods and increase effect sizes, leading to sample requirements that are larger than traditional norms in neuroimaging research but still less than a thousand.

Table 1. Recommendations to Increase Effect Sizes in Neuroimaging Research on Individual Differences, Organized by Stages of Research

Strategy	*Examples*	*Relevant References*	*Article Section*
(1) Study design			
Selecting tasks for fMRI acquisition that induce/require behavior that is relevant for the behavioral constructs of interest	Using a working-memory task during fMRI acquisition to study neural correlates of intelligence	Greene et al. (2018)	2
Using tasks during fMRI assessment that were developed to measure individual differences instead of tasks optimized for research on within-person effects	Stoop and Flanker tasks that were specifically developed to detect individual differences	Burgoyne et al. (2023)	3
Using longer questionnaires instead of abbreviated forms	A measure of personality traits that includes 10 items for each construct rather than two	Credé et al. (2012)	3
Using multiple informants and multiple measurement modalities	Using self-, parent, and teacher ratings of children's behavioral problems; measuring impulsivity using tasks as well as questionnaires	Joyner and Perkins (2023)	3
Choosing behavioral measures appropriate for the population being studied	Avoiding instruments optimized only for making clinical distinctions when assessing the general population	Tiego et al. (2023)	3
Using computerized adaptive testing (CAT) to improve questionnaire or cognitive test assessment	Using CAT to reduce the number of items needed for high quality assessments of intelligence or symptoms of psychopathology	Wainer et al. (2000)	4
(2) Data acquisition			
Using multi-echo fMRI to improve signal-to-noise ratio and reduce signal dropout		Kundu et al. (2017); Lynch et al. (2020)	3
Increasing the amount of fMRI data per participant	Lengthening scan time or conducting multiple scan sessions longitudinally	Cho et al. (2021); Noble et al. (2017)	3
Modeling the hierarchical structure of neural parameters		Schubert et al. (2022)	3
Using machine-learning methods to increase the reliability of neural parameters		Blair et al. (2022)	3
Taking into account categorical differences in brain morphology	Considering whether each participant has one or two sulci in ACC	Miller et al. (2021); Voorhies et al. (2021); Amiez et al. (2018); Fornito et al. (2008)	4
Using individualized parcellation of the brain	Applying GPIP or MS-HBM	Mueller et al. (2013); Chong et al. (2017); Kong et al. (2019)	4

Using hyperalignment	Intelligence can be predicted more accurately using hyperalignment than with nonindividualized methods	Feilong et al. (2021); Haxby et al. (2020)	4
(3) Data processing			
Modeling the idiosyncratic shape of the HRF for each participant		Sigh et al. (2020, 2022)	4
Individualization of psychological measures	Identifying the extent to which participants use different strategies in the same tasks (e.g., model-based vs. model-free strategies in learning tasks)	Kool et al. (2017)	4
(4) Formal analysis			
Using latent variable models for neural or behavioral variables		Joyner and Perkins (2023); Schubert et al. (2020); Tiego et al. (2023)	3
Using multivariate approaches with cross-validation to predict criterion variables	Using multiple variables of the same type or combining variables from different modalities in a prediction model	Thiele et al. (2023, 2024); Jiang et al. (2020); Rasero et al. (2021)	5

fMRI = functional magnetic resonance imaging.

Acknowledgments

Authors were supported by the following sources of funding.

Corresponding authors: Kirsten Hilger, University of Würzburg, Marcusstraße 9–11, Germany, or via e-mail: kirsten.hilger@uni-wuerzburg.de or Colin DeYoung, Department of Psychology, 75 East River Rd, Minneapolis, MN 55455 USA, or via e-mail: cdeyoung@umn.edu.

Author Contributions

Colin G. DeYoung: Conceptualization; Visualization; Writing—Original draft; Writing—Review & editing. Kirsten Hilger: Conceptualization; Visualization; Writing—Original draft; Writing—Review & editing. Jamie L. Hanson: Conceptualization; Writing—Review & editing. Rany Abend: Writing—Review & editing. Timothy A. Allen: Writing—Review & editing. Roger E. Beaty: Writing—Review & editing. Scott D. Blain: Writing—Review & editing. Robert S. Chavez: Writing—Review & editing. Stephen A. Engel: Writing—Review & editing. Ma Feilong: Writing—Review & editing. Alex Fornito: Writing—Review & editing. Erhan Genç: Writing—Review & editing. Vina Goghari: Writing—Review & editing. Rachael G. Grazioplene: Writing—Review & editing. Philipp Homan: Writing—Review & editing. Keanan Joyner: Writing—Review & editing. Antonia N. Kaczkurkin: Writing—Review & editing. Robert D. Latzman: Writing—Review & editing. Elizabeth A. Martin: Writing—Review & editing. Aki Nikolaidis: Writing—Review & editing; Alan D. Pickering: Writing—Review & editing; Adam Safron: Writing—Review & editing; Tyler A. Sassenberg: Writing—Review & editing; Michelle N. Servaas: Writing—Review & editing; Luke D. Smillie: Writing—Review & editing; R. Nathan Spreng: Writing—Review & editing; Essi Viding: Writing—Review & editing; Jan Wacker: Writing—Review & editing.

Funding Information

Kirsten Hilger, German Research Foundation ([DFG] https://dx.doi.org/10.13039/501100001659), grant number: HI 2185-1/1. Roger E. Beaty, National Science Foundation Graduate Research Fellowship Program (https://dx.doi.org/10.13039/100023581), grant numbers: DRL-1920653, and DUE-2155070.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *Neuroimage*, *8*, 360–369. https://doi.org/10.1006/nimg.1998.0369, PubMed: 9811554

Amiez, C., Wilson, C. R., & Procyk, E. (2018). Variations of cingulate sulcal organization and link with cognitive performance. *Scientific Reports*, *8*, 13988. https://doi.org/10.1038/s41598-018-32088-9, PubMed: 30228357

Blair, R. J. R., Mathur, A., Haines, N., & Bajaj, S. (2022). Future directions for cognitive neuroscience in psychiatry: Recommendations for biomarker design based on recent test re-test reliability work. *Current Opinion in Behavioral Sciences*, *44*, 101102. https://doi.org/10.1016/j.cobeha.2022.101102

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and measurement of attention control. *Journal of Experimental Psychology: General*, *152*, 2369. https://doi.org/10.1037/xge0001408, PubMed: 37079831

Chen, J., Tam, A., Kebets, V., Orban, C., Ooi, L. Q. R., Asplund, C. L., et al. (2022). Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nature Communications*, *13*, 1–17. https://doi.org/10.1038/s41467-022-29766-8, PubMed: 35468875

Cho, J. W., Korchmaros, A., Vogelstein, J. T., Milham, M. P., & Xu, T. (2021). Impact of concatenating fMRI data on reliability for functional connectomics. *Neuroimage*, *226*, 117549. https://doi.org/10.1016/j.neuroimage.2020.117549, PubMed: 33248255

Chong, M., Bhushan, C., Joshi, A. A., Choi, S., Haldar, J. P., Shattuck, D. W., et al. (2017). Individual parcellation of resting fMRI with a group functional connectivity prior. *Neuroimage*, *156*, 87–100. https://doi.org/10.1016/j.neuroimage.2017.04.054, PubMed: 28478226

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the big five personality traits. *Journal of Personality and Social Psychology*, *102*, 874. https://doi.org/10.1037/a0027403, PubMed: 22352328

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*, 453–482. https://doi.org/10.1146/annurev-psych-120710-100353, PubMed: 21943169

DeYoung, C. G., Beaty, R. E., Genç, E., Latzman, R. D., Passamonti, L., Servaas, M. N., et al. (2022). Personality neuroscience: An emerging field with bright prospects. *Personality Science*, *3*, 1–21. https://doi.org/10.5964/ps.7269, PubMed: 36250039

DeYoung, C. G., Shamosh, N. A., Green, A. E., Braver, T. S., & Gray, J. R. (2009). Intellect as distinct from openness: Differences revealed by fMRI of working memory. *Journal of Personality and Social Psychology*, *97*, 883–892. https://doi.org/10.1037/a0016615, PubMed: 19857008

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., et al. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science*, *31*, 792–806. https://doi.org/10.1177/0956797620916786, PubMed: 32489141

Feilong, M., Guntupalli, J. S., & Haxby, J. V. (2021). The neural basis of intelligence in fine-grained cortical topographies. *eLife*, *10*, e64058. https://doi.org/10.7554/eLife.64058, PubMed: 33683205

Finn, E. S. (2021). Is it time to put rest to rest? *Trends in Cognitive Sciences*, *25*, 1021–1032. https://doi.org/10.1016/j.tics.2021.09.005, PubMed: 34625348

Finn, E. S., Scheinost, D., Finn, D. M., Shen, X., Papademetris, X., & Constable, R. T. (2017). Can brain state be manipulated to emphasize individual differences in functional connectivity? *Neuroimage*, *160*, 140–151. https://doi.org/10.1016/j.neuroimage.2017.03.064, PubMed: 28373122

Fornito, A., Wood, S. J., Whittle, S., Fuller, J., Adamson, C., Saling, M. M., et al. (2008). Variability of the paracingulate sulcus and morphometry of the medial frontal cortex: Associations with cortical thickness, surface area, volume, and sulcal depth. *Human Brain Mapping*, *29*, 222–236. https://doi.org/10.1002/hbm.20381, PubMed: 17497626

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651. https://doi.org/10.1177/1745691614551642, PubMed: 26186114

Grady, C. L., Rieck, J. R., Nichol, D., Rodrigue, K. M., & Kennedy, K. M. (2021). Influence of sample size and analytic approach on stability and interpretation of brain–behavior correlations in task-related fMRI data. *Human Brain Mapping*, *42*, 204–219. https://doi.org/10.1002/hbm.25217, PubMed: 32996635

Greene, A. S., Gao, S., Scheinost, D., & Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, *9*, 2807. https://doi.org/10.1038/s41467-018-04920-3, PubMed: 30022026

Haines, N., Sullivan-Toole, H., & Olino, T. (2023). From classical methods to generative models: Tackling the unreliability of neuroscientific measures in mental health research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*, 822–831. https://doi.org/10.1016/j.bpsc.2023.01.001, PubMed: 36997406

Hardikar, S., McKeown, B., Turnbull, A., Xu, T., Valk, S. L., Bernhardt, B. C., et al. (2024). Personality traits vary in their association with brain activity across situations. *Communications Biology*, *7*, 1498. https://doi.org/10.1038/s42003-024-07061-0, PubMed: 39533085

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*, e56601. https://doi.org/10.7554/eLife.56601, PubMed: 32484439

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1, PubMed: 28726177

Hilger, K., & Markett, S. (2021). Personality network neuroscience: Promises and challenges on the way toward a unifying framework of individual variability. *Network Neuroscience*, *5*, 631–645. https://doi.org/10.1162/netn_a_00198, PubMed: 34746620

Jiang, R., Calhoun, V. D., Cui, Y., Qi, S., Zhuo, C., Li, J., et al. (2020). Multimodal data revealed different neurobiological correlates of intelligence between males and females. *Brain Imaging and Behavior*, *14*, 1979–1993. https://doi.org/10.1007/s11682-019-00146-z, PubMed: 31278651

Joyner, K. J., & Perkins, E. R. (2023). Challenges and ways forward in bridging units of analysis in clinical psychological science. *Journal of Psychopathology and Clinical Science*, *132*, 888–896. https://doi.org/10.1037/abn0000879, PubMed: 37843543

Kazan, S. M., Mohammadi, S., Callaghan, M. F., Flandin, G., Huber, L., Leech, R., et al. (2016). Vascular autorescaling of fMRI (VasA fMRI) improves sensitivity of population studies: A pilot study. *Neuroimage*, *124*, 794–805. https://doi.org/10.1016/j.neuroimage.2015.09.033, PubMed: 26416648

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., et al. (2021). Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cerebral Cortex*, *31*, 4477–4500. https://doi.org/10.1093/cercor/bhab101, PubMed: 33942058

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost–benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*, 1321–1333. https://doi.org/10.1177/0956797617708288, PubMed: 28731839

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*, 151–177. https://doi.org/10.1080/1047840X.2016.1153946

Kundu, P., Voon, V., Balchandani, P., Lombardo, M. V., Poser, B. A., & Bandettini, P. A. (2017). Multi-echo fMRI: A review of applications in fMRI denoising and analysis of BOLD signals. *Neuroimage*, *154*, 59–80. https://doi.org/10.1016/j.neuroimage.2017.03.033, PubMed: 28363836

Lynch, C. J., Power, J. D., Scult, M. A., Dubin, M., Gunning, F. M., & Liston, C. (2020). Rapid precision functional mapping of individuals using multi-echo fMRI. *Cell Reports*, *33*, 108540. https://doi.org/10.1016/j.celrep.2020.108540, PubMed: 33357444

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*, 654–660. https://doi.org/10.1038/s41586-022-04492-9, PubMed: 35296861

Miller, J. A., Voorhies, W. I., Lurie, D. J., D'Esposito, M., & Weiner, K. S. (2021). Overlooked tertiary sulci serve as a meso-scale link between microstructural and functional properties of human lateral prefrontal cortex. *Journal of Neuroscience*, *41*, 2229–2244. https://doi.org/10.1523/JNEUROSCI.2362-20.2021, PubMed: 33478989

Moghimi, P., Dang, A. T., Do, Q., Netoff, T. I., Lim, K. O., & Atluri, G. (2022). Evaluation of functional MRI-based human brain parcellation: A review. *Journal of Neurophysiology*, *128*, 197–217. https://doi.org/10.1152/jn.00411.2021, PubMed: 35675446

Mueller, S., Wang, D., Fox, M. D., Yeo, B. T., Sepulcre, J., Sabuncu, M. R., et al. (2013). Individual variability in functional connectivity architecture of the human brain. *Neuron*, *77*, 586–595. https://doi.org/10.1016/j.neuron.2012.12.028, PubMed: 23395382

Nebe, S., Reutter, M., Baker, D. H., Bölte, J., Domes, G., Gamer, M., et al. (2023). Enhancing precision in human neuroscience. *eLife*, *12*, e85980. https://doi.org/10.7554/eLife.85980, PubMed: 37555830

Nikolaidis, A., Chen, A. A., He, X., Shinohara, R., Vogelstein, J., Milham, M., et al. (2022). Suboptimal phenotypic reliability impedes reproducible human neuroscience. *BioRxiv*, 2022–2007. https://doi.org/10.1101/2022.07.22.501193

Noble, S., Spann, M. N., Tokoglu, F., Shen, X., Constable, R. T., & Scheinost, D. (2017). Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cerebral Cortex*, *27*, 5415–5429. https://doi.org/10.1093/cercor/bhx230, PubMed: 28968754

Ooi, L. Q. R., Chen, J., Zhang, S., Kong, R., Tam, A., Li, J., et al. (2022). Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *Neuroimage*, *263*, 119636. https://doi.org/10.1016/j.neuroimage.2022.119636, PubMed: 36116616

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., et al. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*, 702–709. https://doi.org/10.1038/ng.3285, PubMed: 25985137

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., et al. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*, 115–126. https://doi.org/10.1038/nrn.2016.167, PubMed: 28053326

Rasero, J., Sentis, A. I., Yeh, F. C., & Verstynen, T. (2021). Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *PLoS Computational Biology*, *17*, e1008347. https://doi.org/10.1371/journal.pcbi.1008347, PubMed: 33667224

Sassenberg, T. A., Burton, P. C., Mwilambwe-Tshilobo, L., Jung, R. E., Rustichini, A., Spreng, R. N., et al. (2023). Conscientiousness associated with efficiency of the salience/ventral attention network: Replication in three samples using individualized parcellation. *Neuroimage*, *272*, 120081. https://doi.org/10.1016/j.neuroimage.2023.120081, PubMed: 37011715

Schubert, A. L., Löffler, C., & Hagemann, D. (2022). A neurocognitive psychometrics account of individual differences in attentional control. *Journal of Experimental Psychology: General*, *151*, 2060–2082. https://doi.org/10.1037/xge0001184, PubMed: 35130011

Schulz, M. A., Bzdok, D., Haufe, S., Haynes, J. D., & Ritter, K. (2024). Performance reserves in brain-imaging-based phenotype prediction. *Cell Reports*, *43*, 113597. https://doi.org/10.1016/j.celrep.2023.113597, PubMed: 38159275

Setton, R., Mwilambwe-Tshilobo, L., Sheldon, S., Turner, G. R., & Spreng, R. N. (2022). Hippocampus and temporal pole functional connectivity is associated with age and individual differences in autobiographical memory. *Proceedings of the National Academy of Sciences, U.S.A.*, *119*, e2203039119. https://doi.org/10.1073/pnas.2203039119, PubMed: 36191210

Singh, M. F., Braver, T. S., Cole, M. W., & Ching, S. (2020). Estimation and validation of individualized dynamic brain models with resting state fMRI. *Neuroimage*, *221*, 117046. https://doi.org/10.1016/j.neuroimage.2020.117046, PubMed: 32603858

Singh, M. F., Wang, A., Cole, M., Ching, S., & Braver, T. S. (2022). Enhancing task fMRI preprocessing via individualized model-based filtering of intrinsic activity dynamics. *Neuroimage*, *247*, 118836. https://doi.org/10.1016/j.neuroimage.2021.118836, PubMed: 34942364

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, *30*, 711–727. https://doi.org/10.1177/0956797619831612, PubMed: 30950321

Spisak, T., Bingel, U., & Wager, T. D. (2023). Multivariate BWAS can be replicable with moderate sample sizes. *Nature*, *615*, E4–E7. https://doi.org/10.1038/s41586-023-05745-x, PubMed: 36890392

Sripada, C., Angstadt, M., Rutherford, S., Taxali, A., & Shedden, K. (2020). Toward a "treadmill test" for cognition: Improved prediction of general cognitive ability from the task activated brain. *Human Brain Mapping*, *41*, 3186–3197. https://doi.org/10.1002/hbm.25007, PubMed: 32364670

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys*, *19*, 309–345. https://doi.org/10.1111/j.0950-0804.2005.00250.x

Szucs, D., & Ioannidis, J. P. (2020). Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage*, *221*, 117164. https://doi.org/10.1016/j.neuroimage.2020.117164, PubMed: 32679253

Tervo-Clemmens, B., Marek, S., Chauvin, R. J., Van, A. N., Kay, B. P., Laumann, T. O., et al. (2023). Reply to: Multivariate BWAS can be replicable with moderate sample sizes. *Nature*, *615*, E8–E12. https://doi.org/10.1038/s41586-023-05746-w, PubMed: 36890374

Tetereva, A., Li, J., Deng, J. D., Stringaris, A., & Pat, N. (2022). Capturing brain-cognition relationship: Integrating task-based fMRI across tasks markedly boosts prediction and test-retest reliability. *Neuroimage*, *263*, 119588. https://doi.org/10.1016/j.neuroimage.2022.119588, PubMed: 36057404

Thiele, J. A., Faskowitz, J., Sporns, O., & Hilger, K. (2024). Can machine learning-based predictive modelling improve our understanding of human cognition? *PNAS Nexus*, *13*, 519. https://doi.org/10.1101/2023.12.04.569974

Thiele, J. A., Richter, A., & Hilger, K. (2023). Multimodal brain signal complexity predicts human intelligence. *eNeuro*, *10*, 1–18. https://doi.org/10.1523/ENEURO.0345-22.2022, PubMed: 36657966

Tiego, J., Martin, E. A., DeYoung, C. G., Hagan, K., Cooper, S. E., Pasion, R., et al. (2023). Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. *Nature Mental Health*, *1*, 304–315. https://doi.org/10.1038/s44220-023-00057-5, PubMed: 37251494

Uddin, L. Q., Betzel, R. F., Cohen, J. R., Damoiseaux, J. S., De Brigard, F., Eickhoff, S. B., et al. (2023). Controversies and progress on standardization of large-scale brain network nomenclature. *Network Neuroscience*, *7*, 864–905. https://doi.org/10.1162/netn_a_00323, PubMed: 37781138

Voorhies, W. I., Miller, J. A., Yao, J. K., Bunge, S. A., & Weiner, K. S. (2021). Cognitive insights from tertiary sulci in prefrontal cortex. *Nature Communications*, *12*, 5122. https://doi.org/10.1038/s41467-021-25162-w, PubMed: 34433806

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. https://doi.org/10.4324/9781410605931

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 294–298. https://doi.org/10.1111/j.1745-6924.2009.01127.x, PubMed: 26158966

Zuo, X. N., & Xing, X. X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: A systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews*, *45*, 100–118. https://doi.org/10.1016/j.neubiorev.2014.05.009, PubMed: 24875392