CellPress

Trends in Cognitive Sciences

Review

Striving toward translation: strategies for reliable fMRI measurement

Maxwell L. Elliott,[1,*] Annchen R. Knodt,[1] and Ahmad R. Hariri[1]

fMRI has considerable potential as a translational tool for understanding risk, prioritizing interventions, and improving the treatment of brain disorders. However, recent studies have found that many of the most widely used fMRI measures have low reliability, undermining this potential. Here, we argue that many fMRI measures are unreliable because they were designed to identify group effects, not to precisely quantify individual differences. We then highlight four emerging strategies [extended aggregation, reliability modeling, multi-echo fMRI (ME-fMRI), and stimulus design] that build on established psychometric properties to generate more precise and reliable fMRI measures. By adopting such strategies to improve reliability, we are optimistic that fMRI can fulfill its potential as a clinical tool.

Highlights

Since its introduction in 1992, fMRI has rapidly matured to become a powerful tool in neuroscience, allowing researchers to noninvasively map the functional organization of the average human brain and probe the brain bases of behaviors from the simple to the complex.

However, the translation of fMRI to clinical applications as well as the study of individual differences in brain function has been limited to date. This limitation, in part, reflects the inadequate reliability of many of the most commonly used fMRI measures. Reliability is a prerequisite for the valid measurement of brain function in individuals.

We highlight four emerging strategies, each with roots in psychometrics, that have the potential to improve measurement reliability, thereby advancing the potential clinical utility of fMRI: (i) extended aggregation; (ii) reliability modeling; (iii) multi-echo fMRI; and (iv) stimulus design.

Can we reliably measure individual differences in brain function?

Cognitive neuroscience has revolutionized our understanding of how the brain supports behavioral functions ranging from basic sensory to complex cognitive processes. Based on these fundamental insights into brain and behavior, an emerging program of translational neuroscience seeks to identify individual differences in these patterns and, in so doing, inform the development of clinical biomarkers that can be used to predict disease risk, prioritize interventions, and improve treatment. Central to these efforts is fMRI, as it affords the noninvasive measurement of brain activity in behaving humans across the lifespan. In recent years, fMRI studies of individual differences in clinically meaningful domains have proliferated, alongside expectations for clinical applications [1]. This expansion of individual-differences research using fMRI has triggered questions about its readiness to fulfill the measurement properties necessary for **clinical translation** (see Glossary), central among which is **reliability**.

Psychometrics has long established that reliability is the necessary first step toward validity. For example, to investigate how brain function makes a super-ager resilient in the face of neurodegeneration or to tailor brain stimulation to an individual's unique **functional topography**, we must first be able to reliably **measure** idiosyncrasies in brain function. To establish reliability, repeated measurements of brain function must produce converging estimates in the absence of significant changes in the individual (e.g., disease progression, exposure to treatment). Recently, we reported that many of the most widely adopted task-fMRI measures of brain activity during clinically relevant behavior (e.g., episodic memory, executive control) have low test–retest reliability and are, therefore, unable to serve as clinical biomarkers in their current state [2]. Results from similar studies have also pointed to low reliability in other widely used fMRI measures including functional connectivity measures generated from short scans [3–5]. Fortunately, methods to improve reliability have long been developed and employed in the allied field of psychology (e.g., personality or cognitive assessments). However, these methods have yet to be fully adopted in fMRI research.

In this review, we begin by describing historical trends that contributed to the widespread use of unreliable fMRI measures in individual-differences research. Then we highlight four

[1]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA

*Correspondence: maxwell.elliott@duke.edu (M.L. Elliott).

emerging strategies (extended aggregation, reliability modeling, ME-fMRI, and stimulus design), each with roots in psychometrics, that could enable researchers to reliably measure individual differences in brain function with fMRI (Figure 1, Key figure). We conclude that, despite several false starts and dead ends, there is a bright future for a cumulative, translational neuroscience of individual differences using fMRI, but that for this future to be realized we must upend *status quo* approaches in favor of psychometrically sound principles for measure development.

A very brief history of individual-differences research with fMRI

The 1990s

In March 1992, the first studies to map human brain function with MRI were sequentially reported by Kwong *et al.* and Ogawa *et al.* [6,7]. In each study, patterns of activity in the visual cortex were measured using the blood-oxygen level-dependent (BOLD) signal by implementing a within-subject design to contrast activity between alternating blocks of darkness and visual stimulation. These simple yet powerful experiments demonstrated the potential of fMRI to noninvasively measure human brain activity, sparking a flurry of research further 'mapping' functions of the human brain. As these early experiments were typically designed for cognitive neuroscience research and required costly, technologically demanding infrastructures, they often relied on two critical design features: experimental manipulation and group averaging. Experimental manipulation was achieved by presenting tightly controlled stimuli in structured patterns so that the BOLD signal for a condition of interest could be experimentally contrasted with the BOLD signal during a baseline condition. By carefully constraining stimulus features (e.g., visual angle, size, instructions) and timing (e.g., block and event-related presentations), researchers could find patterns of brain activation that reflected specific, experimentally contrasted differences in task conditions (e.g., working memory versus passive viewing of visual objects). In tandem, group averaging was utilized to reduce the inherent noisiness of individual-level BOLD data to elicit robust group effects between conditions of interest, thereby allowing inferences about the functions of the 'average human brain'. Tasks were explicitly designed and optimized to consistently evoke within-subject effects using experimental manipulation and group averaging. Using these core tools, the first decade of fMRI started with a technological trigger that was followed by iterative development, widespread adoption, and increasing expectations for the translation of fMRI into the advancement of our understanding and treatment of brain disorders ranging from depression and schizophrenia to Alzheimer's disease (Figure 2).

The 2000s

During its second decade, fMRI expanded in both breadth and depth. More powerful scanners provided measures of brain activity with ever-greater spatial and temporal resolution [8]. Simultaneously, fMRI became increasingly employed in research with unique populations (e.g., children, brain disorders). It was during this period of expansion that some investigators began adopting fMRI tasks, originally developed to elicit robust within-subject effects, to probe between-subjects individual differences. Due to the high cost of fMRI and the infancy of the technology, these investigations of individual differences were often a secondary aim of studies, opportunistically explored after the primary experimental cognitive neuroscience questions [9]. The logic of this approach was straightforward and alluring: if an fMRI task experimentally elicits activation in a targeted brain region during a psychological process of interest, variability in the magnitude of that activity between individuals may drive variability in related behaviors and clinical endpoints. In this way, investigators attempted to simultaneously map behaviorally relevant brain activation and associate variability in this activation with individual differences in behavior. For example, work by us and others demonstrated that, when

Glossary

Between-subjects variance: variability in a measure due to differences between individuals (i.e., individual differences). In fMRI, this is often represented by individual differences in the magnitude of activation or functional connectivity.

Classical test theory: an early psychometric framework to improve the reliability of tests by conceptualizing observed scores from a measurement device as the sum of error score variation and true score variation.

Clinical translation: the process of using scientific insight to inform the diagnosis, treatment, and prevention of clinical disorders. fMRI could one day be more routinely used to assess risk and prioritize treatments for brain disorders in individuals.

Ecological validity: the extent to which an experimental paradigm generalizes to the intended setting of interest. In fMRI, ecological validity increases as stimuli and tasks better capture real-world experiences and challenges faced by individuals (e.g., naturalistic stimuli).

Error score: the part of a measurement that is driven by random noise, measurement error, artifacts, and bias.

Functional topography: the spatial layout of brain functions and network organization across the cortex that has both general properties (e.g., the location of primary sensory cortex) and highly individualized patterns (see Figure 3 in the main text).

Generalizability theory: a psychometric framework that expands classical test theory with tools to disambiguate the multitude of sources of true score and error score variance. In fMRI, this allows researchers to measure variance that is driven by sources including the scanner, the time of day, and the task.

Item-response theory: a psychometric framework that expands on classical test theory by providing tools to assess, design, and select individual items for a test instead of focusing on the observed scores that are generated from the test itself.

Latent variable: a variable that is not directly observed but instead is inferred from other measurements.

Measure: a standard unit used to express the size, amount, or degree of something. Reliability is a property of a measure and will thus vary as the properties of a measure vary. In fMRI,

averaged across participants, the amygdala exhibits increased activity when participants view threat-related facial expressions compared with neutral visual stimuli [10,11]. In response to the extensive animal literature demonstrating the critical importance of the amygdala in fear learning and stress-related dysfunction, we and others hypothesized that variability in the magnitude of this threat-related amygdala activity would map onto individual differences in related behaviors such as anxiety and depression. These links were indeed reported alongside many similar studies of individual differences across a wide variety of fMRI tasks, behavioral traits, and clinical symptoms [12,13].

The 2010s

In the 2010s, two emerging trends simultaneously attracted more attention and scrutiny to translational fMRI. First, spurred by the maturation and promise of fMRI, large consortium studies (e.g., the Human Connectome Project [14], the UK Biobank [15]) were founded with the explicit goal of using fMRI tasks from experimental cognitive neuroscience to measure individual differences in brain function. With large samples and broad phenotyping, these studies created an open-access canvas for fMRI researchers to test the generalizability, replicability, and reliability of individual-differences findings from prior work in small samples. Second, amid the emerging replication crisis in psychology [16,17], a wave of studies found critical limitations in mainstream fMRI practices, calling into question many individual-differences findings. Among these, researchers noted that statistical circularity was common in fMRI analyses, leading to inflated effect sizes [18]. Statistical methods for mapping brain activity were also found to be too liberal in determining statistical significance [19] and widely implemented statistical approaches to generalize from experiments to the 'real world' were found to be inadequate [20,21]. Relatedly, others noted that the small sample sizes of most fMRI studies ($n \leq 100$) left them underpowered to detect realistic, uninflated effect sizes [22,23]. Furthermore, investigators discovered that many fMRI findings were confounded by group differences in head-motion and physiological artifacts during scanning [24,25]. Finally, our group and others found that many of the most commonly used fMRI measures had low test–retest reliability [2,3,5,26], which fundamentally undermines a measure's ability to index individual differences or serve as a clinical biomarker. Collectively, these findings highlighted previously underappreciated limitations of fMRI research that called into question the ability of many fMRI measures to validly measure individual differences using *status quo* methodologies.

The present

These observations have already sparked methodological innovations to directly address many of these limitations, including more accurate methods for statistical inference, larger samples, multivariate modeling to boost reliability, motion censoring, and advanced data-processing techniques [27–34]. However, reliance on short scans (i.e., 5–10 min) in small samples ($n \leq 100$) as well as rigid stimulus control and group averaging continue to limit the ability of fMRI to reliably measure individual differences in brain function that represent interpretable and tractable mechanisms of risk, pathophysiology, and treatment response. Even with advanced approaches for artifact reduction and statistical inferences, commonly used fMRI methods often generate unreliable measures. This unreliability thus continues to represent a fundamental threat to our ability to realize a rigorous translational neuroscience of individual differences. Given that reliability is a minimum, necessary prerequisite for valid individual-differences research, a growing contingent of fMRI researchers have sought to build a new framework for translational neuroscience by asking a fundamental question: under what conditions can fMRI generate reliable, individual-specific measures of brain function with sufficient precision to inform clinical practice? We next highlight four complementary strategies that have emerged in response to this question, as well as the psychometric principles that underlie their utility.

each combination of task length, stimulus type, task demands, etc. generates different measures with different reliabilities.

Naturalistic stimuli: experimental stimuli that are designed to approximate the rich complexity of everyday experience as opposed to the rigidly controlled stimuli that are common in experimental cognitive neuroscience. In fMRI, naturalistic stimuli can include complex audio (e.g., speeches, stories) and visual (i.e., pictures and movies) modalities.

Psychometrics: a specialist field focused on developing and improving the measurement of psychological constructs such as personality and cognitive ability.

Reliability: the consistency of a measure when repeatedly assessed under similar conditions. Low reliability statistically limits the ability to detect associations between fMRI measurements and outcome measures (e.g., schizophrenia status, working memory capacity).

True score: the targeted part of a measurement that is free of noise, error, contamination, or bias.

Within-subject variance: variability in a measure within an individual due to different conditions. In fMRI, this is often represented in differences in the pattern of activation between task and control conditions.

Key figure

Emerging strategies to generate more reliable fMRI measures

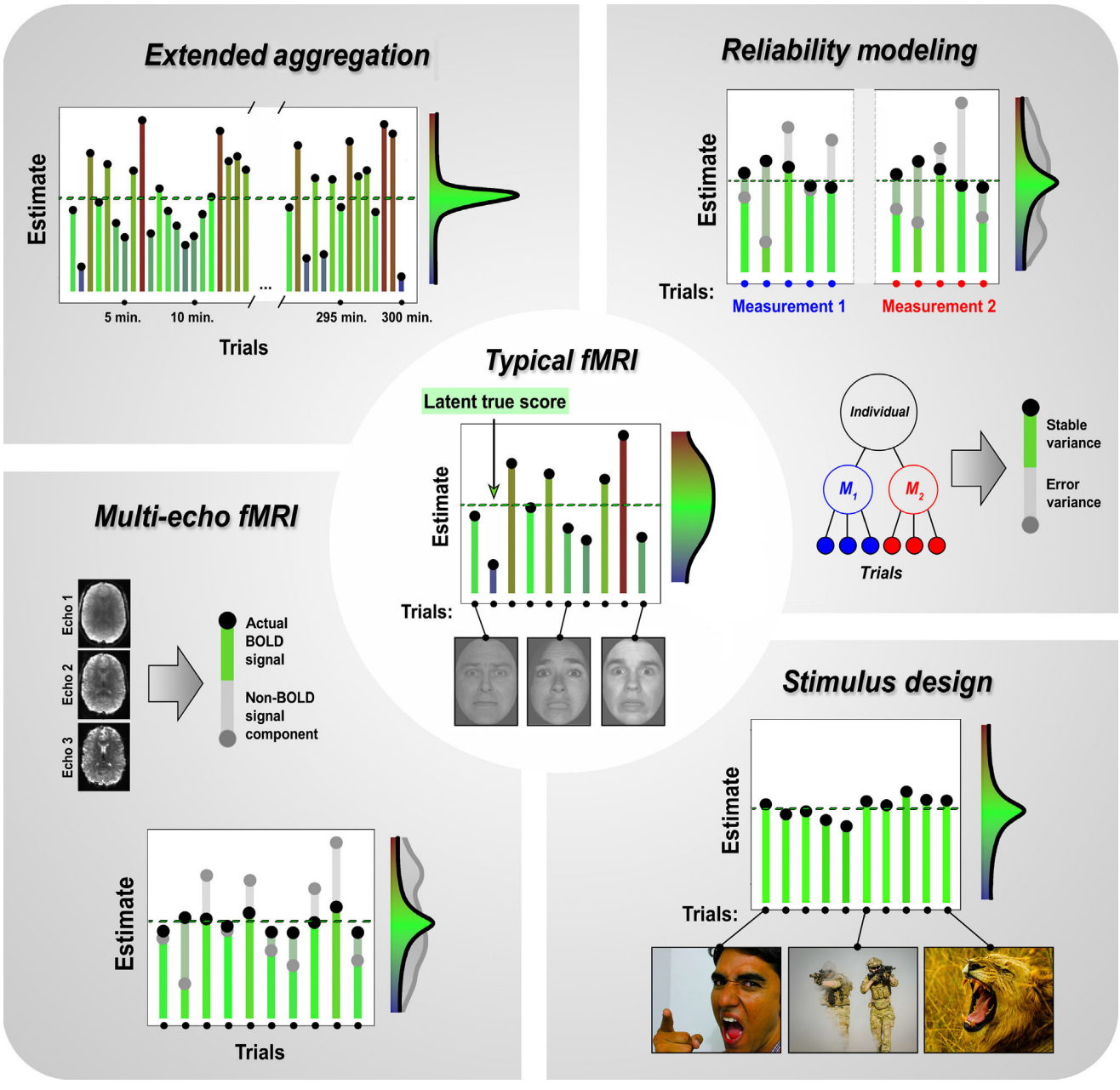

Figure 1. A typical fMRI study generates measures of activation or functional connectivity by averaging over time (often 5–10 min of data). For example, black-and-white photos of facial expressions may be shown in blocks. Then, a single regressor is fitted to all face blocks to generate an average estimate of brain activation for each individual. Many of the most commonly used activation and functional connectivity measures from such short, 'typical' fMRI studies are unreliable (represented by a wide error variability around the true score). The reliability of fMRI measures can be dramatically improved by extended aggregation of hours, instead of minutes, of

(Figure legend continued at the bottom of the next page.)

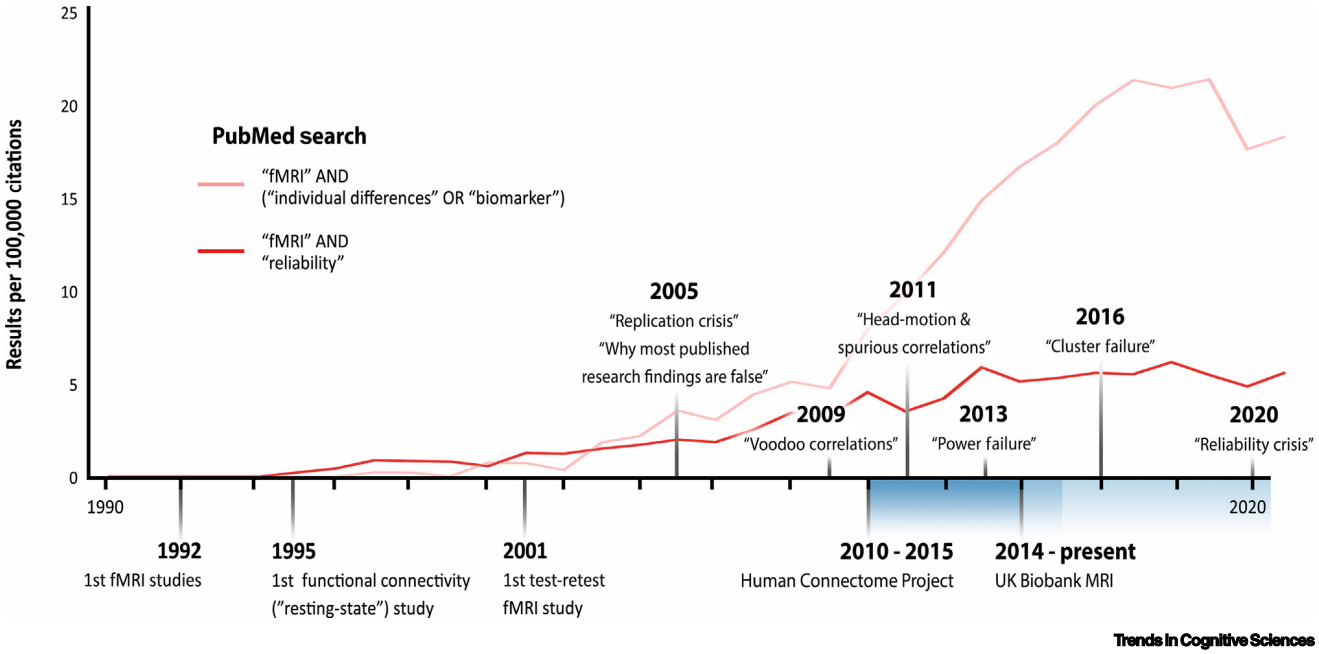

Figure 2. A timeline of fMRI research. A highly select list of events that have helped to shape the current state of fMRI research on individual differences and clinical translation [2,16,18,19,24,100]. Blue shading represents the windows of data collection for the Human Connectome and UK Biobank studies. In addition, the rise of individual differences and biomarker research with fMRI in the scientific literature is plotted, in pink, along with fMRI studies that mention reliability, in red. While this is a simplification of fMRI history, it is clear that the number of publications using fMRI to investigate between-subjects questions rose rapidly in the two decades following its origin before plateauing, while fMRI studies that consider reliability have constituted a smaller fraction of fMRI studies. We have plotted the normalized proportion (search results as a fraction of all PubMed citations in that year) of PubMed search results per year from 1990 to 2020, using esperr.github.io/pubmed-by-year to account for the fact that there has been a rapid rise in the overall number of scientific publications throughout the past century.

Building a reliable neuroscience of individual differences

Reliability can be improved through extended aggregation

The BOLD signal is surrounded by noise originating from thermal, physiological (e.g., motion, respiration), and miscellaneous non-physiological (e.g., scanner drift) sources [35] and represents only a small fraction of the variance in fMRI data, on the order of 5–20% [35]. To tackle the challenge of isolating the subset of BOLD variance driven by reliable individual differences, a growing contingent of 'precision fMRI' (pfMRI) research has adopted a tried-and-true principle from **classical test theory**: collect more data per person. Reliability tends to increase as assessment length increases because there are more opportunities for random, unstructured **error score** variability to cancel itself out. When this happens, **true score** variation constitutes a larger portion of the measurement, resulting in higher reliability. For example, single-item measures are often dominated by noise and item-specific variance that cancels itself out as additional items are added and an aggregate score across many items is used. Typically, in psychometrics, 'assessment length' refers to the number of items on a questionnaire or survey; however, the same principles apply to fMRI scan length. Across a wide variety of cohorts, scanners, and study designs, it has been

data for each individual. These measures are far more precise and reliable because random error score variability cancels out over more trials (represented by the precise density plot around the true score). Reliability modeling can improve reliability by separating stable variability, which is consistent across multiple measurements (green distribution), from error variance (gray distribution), which is transient. Multi-echo fMRI improves reliability by utilizing multiple echoes to separate non-blood-oxygen level-dependent (BOLD) error variability (gray distribution) from the BOLD signal of interest (green distribution). Finally, the reliability of fMRI measures can be improved through stimulus design (i.e., designing stimuli to evoke more reliable between-subjects variance). For example, the stimuli chosen here are colored, visually striking, emotionally rich images that are more complex and relevant to everyday life (i.e., naturalistic and ecologically valid) than the tightly controlled black-and-white photographs of faces shown for a typical fMRI study. Such stimuli could be static images or dynamic movies selected based on their ability to generate reliable individual differences.

shown that the reliability of fMRI measures tends to increase as scan length increases [3–5,36,37]. Illustrating this point, reliability gains are particularly pronounced when data from multiple scan sessions on different days are combined, suggesting that unwanted variance due to transient factors like time of day, head positioning, wakefulness, and scanner effects often obscure stable individual differences hidden within measures from short, single-session scans [38,39]. By collecting many hours of data from a small number of subjects (Figure 3), pfMRI studies have further demonstrated that reliable, individual-specific signatures are present in the spatial organization of brain networks as well as their temporal structure and timescale [36,37,40–43]. Furthermore, pfMRI has helped to uncover new cortical networks that were previously obscured by group averaging [43–45] and begun to move toward clinical applications in the guidance of transcranial magnetic stimulation [46,47], the detection of recovery from traumatic brain injury [48], and the measurement of individual-specific cortical reorganization [49,50].

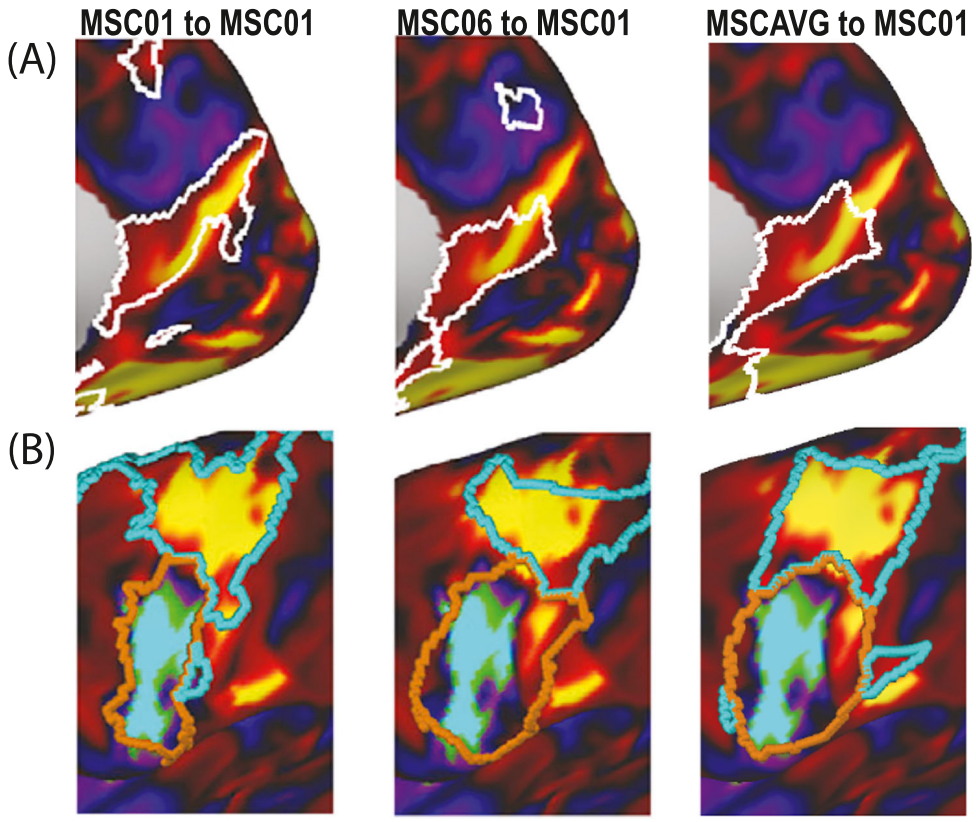

Figure 3. Precision fMRI can reveal reliable, individual-specific features of brain function. By collecting hours of data from each individual, the Midnight Scan Club (MSC) and other 'deep-phenotyping' studies have demonstrated that fMRI can resolve highly detailed individual-specific patterns of brain function that are lost during group averaging. (A) With enough data, precise alignment can be identified between the boundaries of a retrosplenial functional connectivity network (white outline) and a map of task activation in the retrosplenial cortex of a single MSC participant (i.e., MSC01 to MSC01). Critically, this alignment is individual specific and thus does not hold when the same functional connectivity network from a different individual (MSC06) is mapped onto the task activation of MSC01 (i.e., MSC06 to MSC01). Furthermore, the group-average functional connectivity map also maps poorly onto the task activation of MSC01 (i.e., MSCAVG to MSC01). (B) Similarly, the hand (in cyan) and face (in orange) functional connectivity networks can be mapped onto hand and face activation maps in the motor cortex that are highly specific to an individual and obscured by group averaging. Such precision fMRI (pfMRI) reveals that precise, individual-specific estimates of individual differences in brain function are possible but often missed when unreliable measures and group averaging are used. Adapted, with permission, from [41].

Emerging findings from pfMRI indicate that, with enough data, idiosyncratic functional organization may be the rule, not the exception [36]. This suggests that the traditional use of short fMRI scans and group averaging is holding back many translational neuroscience efforts because individual differences are obscured underneath a sea of unaccounted-for variability and noise. However, the current pfMRI methodology requires many hours of data from each individual to achieve high levels of precision [36,37]. In many anatomical targets for translational neuroscience (e.g., amygdala, accumbens, orbitofrontal cortex), signal dropout compounds unreliability, requiring even more data [51–54]. Therefore, the participant burden of pfMRI is high for most developmental and clinical samples for whom it is particularly challenging to lie still in a scanner for hours [24,55]. Thus, in its present form, pfMRI has not been widely pursued in population neuroscience efforts, which will be critical in realizing the broad translational value of fMRI [36,56,57]. However, to date, pfMRI has largely achieved greater reliability through the relatively crude approach of using aggregation to allow unstructured variance to cancel itself out over time [58]. To the extent that we understand the generative source of the true score variability that we want to measure (i.e., BOLD signal) and the error score variability that we want to remove (i.e., noise), alternative strategies may be able to more efficiently achieve precise and reliable measurement with shorter scans and lower participant burden.

Reliability can be improved by modeling stable variability

Translational neuroscience efforts are often focused on measuring stable biomarkers of disease risk, status, and prognosis [1]. However, many of the most widely used fMRI modeling approaches mix stable and transient variability by reducing a large number of fMRI measures to a single average estimate for each individual. Namely, fMRI studies often reduce regional brain function to a single estimate of activation or functional connectivity. In task fMRI, for example, this is frequently done by fitting a single regressor or contrast of interest that represents the alternating structure of a task between control and experimental conditions (e.g., a boxcar model). Similarly, functional connectivity estimates are typically generated by correlating activity across the entire fMRI scan. These modeling approaches were originally designed for experimental cognitive neuroscience, where the **between-subjects variance** is a source of error to be minimized to maximize the statistical power to estimate within-subject experimental effects and group averages. However, with only a single estimate per individual (i.e., task contrast beta or edge functional connectivity), stable, individual-specific variance cannot be separated from transient sources of **within-subject variance** (e.g., fluctuations in thoughts, emotional states, or attention) and noise [59,60].

Recent research suggests that the reliability of task-activation and functional connectivity measures can be substantially improved by explicitly isolating stable variance with tools designed for repeated measures (e.g., **latent variable** and hierarchical Bayesian modeling) [21,38,61–64]. Critically, these modeling approaches can be applied both when multiple scans are available from each individual and when only a single scan is available. This is because fMRI scans intrinsically comprise many estimates of brain activity or connectivity. For example, multiple activation estimates can be generated by fitting regressors to the first and second halves of an fMRI scan separately (i.e., split-half analysis) or, at a finer-grained level, by fitting separate regressors to each trial within a scan [21,45,61]. Similarly, multiple functional connectivity estimates can be generated by splitting a single scan in half thereby generating two functional connectivity estimates or, in the extreme, by generating covariance estimates for every fMRI volume or data point [38,65]. Once multiple estimates are generated for each individual, tools from repeated-measures modeling can be used to separate 'stable components' of fMRI variance from transient variance and noise [64]. Collectively, such modeling has been found to boost the reliability of activation and functional connectivity measures, especially from short fMRI scans, by as much as 60% [38,61]. Moreover, these stable components exhibit higher heritability and larger behavioral associations, further boosting translational value [38,61,62,66–68].

These methods illustrate a measurement principle that may appear counterintuitive: splitting fMRI data into multiple noisier estimates can generate more reliable measures, at the latent variable level, than can be achieved through simple aggregation across the constituent parts. Furthermore, this insight is consistent with recent structural MRI findings that multiple, rapid, lower-resolution scans can generate more precise estimates of brain structure (e.g., cortical thickness) than a single, longer, higher-resolution scan [69]. However, it is important to note that such reliability modeling is not a panacea and cannot replace careful measurement. In its simplest form, such modeling will, by design, absorb all forms of stable variance including stable artifacts like head motion, respiration, and vascular dynamics [29,70,71]. Therefore, to the extent to which these physiological artifacts are imperfectly removed during modeling, they will also be absorbed by the stable component and continue to corrupt the validity of brain–behavior associations.

Reliability can be improved by removing physiological artifacts

Non-BOLD sources of variability, like head motion, are often stable features of individuals [24,70–73]. Therefore, such sources will not necessarily be removed through aggregation or latent variable modeling, because their variance is nonrandom and insidiously mimics individual differences of interest. Furthermore, mainstream data-processing techniques often fail to fully remove these physiological artifacts [24,74–76]. ME-fMRI represents an emerging, biophysically principled approach to isolate and remove noise and non-BOLD sources of variance from fMRI data [52]. To do so, ME-fMRI collects multiple whole-brain images during each excitation pulse (i.e., multiple echoes) instead of the single image that is typically collected (i.e., single echo). This allows the removal of many physiological artifacts because the BOLD signal decays across echoes while non-BOLD artifacts and noise do not [52,77].

As would be expected, improved isolation of the BOLD signal with ME-fMRI generates more precise measurements of task activation and functional connectivity and improves the statistical power [52,78,79]. Furthermore, ME-fMRI substantively reduces signal dropout in regions of particular interest for translational neuroscience (e.g., amygdala, accumbens, orbitofrontal cortex) because the echoes can be optimally weighted based on regionally specific rates of fMRI signal decay [51,52]. Of particular importance, early findings suggest that ME-fMRI allows reliable, precise mapping of individual differences in brain function with much shorter scans. For example, 10 min of ME-fMRI data have been found to generate more stable estimates of functional connectivity than 30 min of single-echo data [51].

While ME-fMRI's widespread adoption been slowed by technological limitations, recent developments in scanner hardware and software (e.g., parallel and multiband imaging) now allow ME-fMRI to be acquired on most scanners with minimal loss of the spatial or temporal resolution typical of single-echo data [51,80]. Given the improved measurement precision already offered, as well as its likely continued development, ME-fMRI represents another promising strategy for translational neuroscientists to prioritize the reliable measurement of individual differences by isolating true sources of individual differences in the BOLD signal from non-BOLD but stable physiological artifacts and noise. ME-fMRI additionally allows innovative study designs that are of interest for translational neuroscience but methodologically challenging for single-echo fMRI. These include measurement of brain function during slow-onset drug-administration paradigms as well as the mapping of rapid, stimulus-driven effects in naturalistic paradigms [81–83]. However, careful data-cleaning practices are still required with ME-fMRI, because stable individual differences in non-neural BOLD effects (e.g., breathing patterns throughout a scan) can still confound individual-differences research [75].

Reliability can be improved by designing stimuli to evoke individual differences

As already noted, the vast majority of fMRI tasks were designed to experimentally manipulate within-subject, group-averaged effects, not to optimally evoke between-subjects individual differences in brain function [2,59,84]. Thus, another strategy to improve fMRI measurement of individual differences is to design new tasks from the ground up with the explicit goal of optimizing reliability and precision. In particular, there is a largely untapped opportunity to adopt psychometric tools from **item-response theory** and **generalizability theory** to select stimuli based on their ability to evoke reliable individual differences (see [85–87] for early steps in this direction). Constructing new tasks and stimuli from the ground up will, admittedly, require time-consuming and expensive fMRI pilot studies to assess large batches of stimuli and task items, test their psychometric features, and iteratively select stimuli that efficiently generate the most precise, reliable measures. However, related efforts suggest that validated fMRI stimuli with known measurement properties can yield large benefits including the ability to create more complete models of how individual brains process sensory, memory, and linguistic information [88–90]. Furthermore, initial evidence suggests that a small subset of fMRI timepoints disproportionately drives reliable individual differences in functional connectivity [91]. As these high-reliability timepoints tend to be elicited by the same movie segments across individuals [91], large efficiency gains may be possible by selecting stimuli that most efficiently evoke individual differences in brain function.

Naturalistic stimuli, such as movies, speeches, and complex social scenarios, may have particular benefits in this regard because they keep participants engaged, awake, and relatively still, thereby minimizing artifacts due to head motion, attention, and wakefulness [92–94]. Naturalistic stimuli also tend to have higher **ecological validity** than traditional tasks and can be easily tailored to a wide variety of content including visual, emotional, and social features that target psychological constructs of interest [88,95,96]. Relatedly, measures generated from movie watching, as well as the combination of multiple tasks with resting-state data, can yield more reliable estimates of brain function with better predictive utility than single tasks or resting-state data alone, further suggesting that reliable individual differences may be best elicited from complex, varied stimuli [5,97,98]. However, these benefits also come with tradeoffs. For example, the complexity of naturalistic stimuli typically cannot be easily controlled (e.g., color composition, spatial frequency) to levels typical of traditional cognitive neuroscience stimuli (but see [99]).

Concluding remarks

In this review, we have described the origins, challenges, and frontiers of current efforts to generate reliable fMRI measures for translational neuroscience. Many of the most commonly used fMRI measures are not yet sufficiently reliable for use as clinical biomarkers. In retrospect, this may not be altogether surprising; the majority of the fMRI measures used today were not designed to identify precise between-subjects variance but rather to reveal within-subject cognitive neuroscience effects through experimental control and group averaging. By considering these legacy constraints, fMRI researchers are now challenged to create new paradigms for the reliable measurement of individual differences in brain function. pfMRI has revealed that deep individuality in the functional organization of the brain is measurable if stable variance is systematically isolated by the collection of large amounts of data in each individual. Furthermore, emerging methods in reliability modeling, ME-fMRI, and study design suggest that reliable individual-specific fMRI measures can be more efficiently generated if protocols are optimized to isolate stable sources of between-subjects variability. Importantly, these methods could be implemented simultaneously and thus may yield complementary returns for precision and reliability. Preliminary efforts that have integrated ME-fMRI with naturalistic stimuli and item-level modeling with data aggregation suggest that the synthesis of these methods may offer the most powerful avenue to identify reliable measures of individual differences in brain function with high translational value and

Outstanding questions

How much data are needed to generate reliable individual-specific estimates of brain function when aggregation, reliability modeling, ME-fMRI, and stimulus design are combined? In other words, are these strategies complementary or substitutes for one another?

What are the 'fundamental units' of individual differences in brain function that can be measured with fMRI? What combination of task activation, functional connectivity during rest and tasks, and/or the way that functional connectivity changes during tasks to shape task activation can best explain individuality in human brain function?

What are the underlying mechanisms that drive reliable individual differences in brain function measurable with fMRI? While we have emerging evidence that such differences reflect local patterns of anatomy, myelination, and structural connectivity, deepening our understanding of these mechanisms can not only further advance fMRI strategies for reliable measurement but also inform translation of findings to clinical applications.

How can we most effectively scale-up reliable fMRI measurement for population neuroscience? To date, pfMRI measures have been limited to highly select, niche datasets. To understand human variation in brain function, especially of clinical value, we will need innovative fMRI protocols that can be readily implemented in large-scale, population-representative samples.

potential for clinical applications [45,82]. Of course, fMRI is still a nascent tool, and many opportunities exist for continued technological innovation and development (see Outstanding questions). The strategies we highlight here are not intended to be exhaustive or prescriptive. Instead, by highlighting these strategies alongside their grounding in psychometric principles, we hope to promote the design of fMRI studies that are better positioned to generate reliable measures of individual differences in brain function. Translational neuroscience with fMRI cannot be a secondary goal of experimental cognitive neuroscience and instead demands iterative, explicit development to optimize the measurement of reliable individual-specific variability. Despite recent setbacks, we see a bright future for a cumulative translational neuroscience of individual differences given that now, more than ever, we understand the limitations of our current fMRI measures and have emerging strategies to build more precise, reliable measures.

Acknowledgments

This work was supported by the National Science Foundation Graduate Research Fellowship (no. NSF DGE-1644868) and the National Institute of Aging grant no. NIA F99 AG068432-01 to M.L.E, as well as grant no. R01AG049789 to A.R.H. We thank Avshalom Caspi, Terrie Moffitt, Tracy d'Arbeloff, Line Rasmussen, Alex Winn, and Ethan Whitman for thoughtful comments and feedback on earlier drafts of the manuscript.

Declaration of interests

No interests are declared.

References

1. Cuthbert, B.N. (2014) The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry* 13, 28–35
2. Elliott, M.L. *et al.* (2020) What is the test–retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychol. Sci.* 31, 792–806
3. Noble, S. *et al.* (2017) Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cereb. Cortex* 27, 5415–5429
4. Birn, R.M. *et al.* (2013) The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *Neuroimage* 83, 550–558
5. Elliott, M.L. *et al.* (2019) General functional connectivity: shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *Neuroimage* 189, 516–532
6. Kwong, K.K. *et al.* (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. U. S. A.* 89, 5675–5679
7. Ogawa, S. *et al.* (1992) Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. U. S. A.* 89, 5951–5955
8. Bandettini, P.A. (2012) Twenty years of functional MRI: the science and the stories. *Neuroimage* 62, 575–588
9. Yarkoni, T. and Braver, T.S. (2010) Cognitive neuroscience approaches to individual differences in working memory and executive control: conceptual and methodological issues. In *Handbook of Individual Differences in Cognition*, pp. 87–107, Springer
10. Hariri, A.R. *et al.* (2000) Modulating emotional responses: effects of a neocortical network on the limbic system. *Neuroreport* 11, 43–48
11. Breiter, H.C. *et al.* (1996) Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* 17, 875–887
12. Braver, T.S. *et al.* (2010) *Vive les differences!* Individual variation in neural mechanisms of executive control. *Curr. Opin. Neurobiol.* 20, 242–250
13. Hariri, A.R. (2009) The neurobiology of individual differences in complex behavioral traits. *Annu. Rev. Neurosci.* 32, 225–247
14. Barch, D.M. *et al.* (2013) Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* 80, 169–189
15. Miller, K.L. *et al.* (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523
16. Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLoS Med.* 2, e124
17. Aarts, A.A. *et al.* (2015) Estimating the reproducibility of psychological science. *Science* 349, aac4716
18. Vul, E. *et al.* (2009) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290
19. Eklund, A. *et al.* (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.* 113, 7900–7905
20. Yarkoni, T. (2019) The generalizability crisis. *PsyArXiv* Published online November 22, 2019. https://doi.org/10.31234/osf.io/jqw35
21. Chen, G. *et al.* (2021) To pool or not to pool: can we ignore cross-trial variability in FMRI? *Neuroimage* 225, 117496
22. Yarkoni, T. (2009) Big correlations in little studies: inflated fMRI correlations reflect low statistical power – commentary on Vul *et al.* (2009). *Perspect. Psychol. Sci.* 4, 294–298
23. Marek, A.S. *et al.* (2020) Towards reproducible brain-wide association studies. *bioRxiv* Published online August 22, 2020. https://doi.org/10.1101/2020.08.21.257758
24. Power, J.D. *et al.* (2012) Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59, 2142–2154
25. Siegel, J.S. *et al.* (2014) Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Hum. Brain Mapp.* 35, 1981–1996
26. Bennett, C.M. and Miller, M.B. (2010) How reliable are the results from functional magnetic resonance imaging? *Ann. N. Y. Acad. Sci.* 1191, 133–155
27. Cox, R.W. *et al.* (2016) *AFNI and Clustering: False Positive Rates Redux*, Cold Spring Harbor Lab Press
28. Gratton, C. *et al.* (2020) Removal of high frequency contamination from motion estimates in single-band fMRI saves data without biasing functional connectivity. *Neuroimage* 217, 116866

29. Fair, D.A. *et al.* (2020) Correction of respiratory artifacts in MRI head motion estimates. *Neuroimage* 208, 116400
30. Cole, M.W. *et al.* (2014) Intrinsic and task-evoked network architectures of the human brain. *Neuron* 83, 238–251
31. Poldrack, R.A. *et al.* (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* 18, 115–126
32. Winkler, A.M. *et al.* (2016) Faster permutation inference in brain imaging. *Neuroimage* 141, 502–516
33. Dubois, J. and Adolphs, R. (2016) Building a science of individual differences from fMRI. *Trends Cogn. Sci.* 20, 425–443
34. Woo, C.W. *et al.* (2017) Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* 20, 365–377
35. Liu, T.T. (2016) Noise contributions to the fMRI signal: an overview. *Neuroimage* 143, 141–151
36. Gratton, C. *et al.* (2020) Defining individual-specific functional neuroanatomy for precision psychiatry. *Biol. Psychiatry* 88, 28–39
37. Gratton, C. *et al.* (2018) Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. *Neuron* 98, 439–452.e5
38. Teeuw, J. *et al.* (2021) Reliability modelling of resting-state functional connectivity. *Neuroimage* 231, 117842
39. Cho, J.W. *et al.* (2021) Impact of concatenating fMRI data on reliability for functional connectomics. *Neuroimage* 226, 117549
40. Raut, R.V. *et al.* (2020) Hierarchical dynamics as a macroscopic organizing principle of the human brain. *Proc. Natl. Acad. Sci. U. S. A.* 117, 20890–20897
41. Gordon, E.M. *et al.* (2017) Precision functional mapping of individual human brains. *Neuron* 95, 791–807.e7
42. Seitzman, B.A. *et al.* (2019) Trait-like variants in human functional brain networks. *Proc. Natl. Acad. Sci. U. S. A.* 116, 22851–22861
43. Braga, R.M. and Buckner, R.L. (2017) Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. *Neuron* 95, 457–471.e5
44. Buckner, R.L. and DiNicola, L.M. (2019) The brain's default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* 20, 593–608
45. DiNicola, L.M. *et al.* (2020) Parallel distributed networks dissociate episodic and social functions within the individual. *J. Neurophysiol.* 123, 1144–1179
46. Cash, R.F.H. *et al.* (2021) Personalized connectivity-guided DLPFC-TMS for depression: advancing computational feasibility, precision and reproducibility. *Hum. Brain Mapp.* Published online February 5, 2020. https://doi.org/10.1002/hbm.25330
47. Cash, R.F.H. *et al.* (2020) Functional magnetic resonance imaging-guided personalization of transcranial magnetic stimulation treatment for depression. *JAMA Psychiatry* 78, 337–339
48. Gordon, E.M. *et al.* (2018) High-fidelity measures of whole-brain functional connectivity and white matter integrity mediate relationships between traumatic brain injury and post-traumatic stress disorder symptoms. *J. Neurotrauma* 35, 767–779
49. Newbold, D.J. *et al.* (2020) Plasticity and spontaneous activity pulses in disused human brain circuits. *Neuron* 107, 580–589.e6
50. Laumann, T.O. *et al.* (2021) Brain network reorganisation in an adolescent after bilateral perinatal strokes. *Lancet Neurol.* 20, 255–256
51. Lynch, C.J. *et al.* (2020) Rapid precision functional mapping of individuals using multi-echo fMRI. *Cell Rep.* 33, 108540
52. Kundu, P. *et al.* (2017) Multi-echo fMRI: a review of applications in fMRI denoising and analysis of BOLD signals. *Neuroimage* 154, 59–80
53. Glasser, M.F. *et al.* (2013) The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* 80, 105–124
54. Marek, S. *et al.* (2018) Spatial and temporal organization of the individual human cerebellum. *Neuron* 100, 977–993.e7
55. Casey, B.J. *et al.* (2018) The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* 32, 43–54
56. Figee, M. and Mayberg, H. (2021) The future of personalized brain stimulation. *Nat. Med.* 27, 196–197
57. Falk, E.B. *et al.* (2013) What is a representative brain? Neuroscience meets population science. *Proc. Natl Acad. Sci. U. S. A.* 110, 17615–17622
58. Rushton, J.P. *et al.* (1983) Behavioral development and construct validity: the principle of aggregation. *Psychol. Bull.* 94, 18–38
59. Hedge, C. *et al.* (2018) The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186
60. Meyer, A. *et al.* (2017) Considering ERP difference scores as individual difference measures: issues with subtraction and alternative approaches. *Psychophysiology* 54, 114–122
61. Chen, G. *et al.* (2021) Beyond the intraclass correlation: a hierarchical modeling approach to test–retest assessment. *bioRxiv* Published online January 5, 2021. https://doi.org/10.1101/2021.01.04.425305
62. Brandmaier, A.M. *et al.* (2018) Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). *Elife* 7, e35718
63. Kong, R. *et al.* (2019) Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cereb. Cortex* 29, 2533–2551
64. Cooper, S.R. *et al.* (2019) Neuroimaging of individual differences: a latent variable modeling perspective. *Neurosci. Biobehav. Rev.* 98, 29–46
65. Betzel, R.F. *et al.* (2019) High-amplitude co-fluctuations in cortical activity drive resting-state functional connectivity. *bioRxiv* Published online October 13, 2019. https://doi.org/10.1101/800045
66. Teeuw, J. *et al.* (2019) Genetic and environmental influences on functional connectivity within and between canonical cortical resting-state networks throughout adolescent development in boys and girls. *Neuroimage* 202, 116073
67. Anderson, K.M. *et al.* (2021) Heritability of individualized cortical network topography. *Proc. Natl Acad. Sci. U.S. A.* 118, e2016271118
68. McCormick, E.M. *et al.* (2021) Latent functional connectivity underlying multiple brain states. *bioRxiv* Published online April 6, 2021. https://doi.org/10.1101/2021.04.05.438534
69. Nielsen, J.A. *et al.* (2019) Precision brain morphometry: feasibility and opportunities of extreme rapid scans. *bioRxiv* Published online January 26, 2019. https://doi.org/10.1101/530436
70. van Dijk, K.R.A. *et al.* (2012) The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage* 59, 431–438
71. Hodgson, K. *et al.* (2017) Shared genetic factors influence head motion during MRI and body mass index. *Cereb. Cortex* 27, 5539–5546
72. Siegel, J.S. *et al.* (2017) Data quality influences observed links between functional connectivity and behavior. *Cereb. Cortex* 27, 4492–4502
73. Power, J.D. *et al.* (2020) A critical, event-related appraisal of denoising in resting-state fMRI studies. *Cereb. Cortex* 30, 5544–5559
74. Power, J.D. *et al.* (2014) Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* 84, 320–341
75. Power, J.D. *et al.* (2018) Ridding fMRI data of motion-related influences: removal of signals with distinct spatial and physical bases in multiecho data. *Proc. Natl Acad. Sci. U.S. A.* 115, 201720985
76. Caballero-Gaudes, C. and Reynolds, R.C. (2017) Methods for cleaning the BOLD fMRI signal. *Neuroimage* 154, 128–149
77. Kundu, P. *et al.* (2012) Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *Neuroimage* 60, 1759–1770
78. Lombardo, M.V. *et al.* (2016) Improving effect size estimation and statistical power with multi-echo fMRI and its impact on understanding the neural systems supporting mentalizing. *Neuroimage* 142, 55–66
79. Kundu, P. *et al.* (2013) Integrated strategy for improving functional connectivity mapping using multiecho fMRI. *Proc. Natl. Acad. Sci. U. S. A.* 110, 16187–16192
80. Olafsson, V. *et al.* (2015) Enhanced identification of BOLD-like components with multi-echo simultaneous multi-slice (MESMS) fMRI and multi-echo ICA. *Neuroimage* 112, 43–51

81. Evans, J.W. *et al.* (2015) Separating slow BOLD from non-BOLD baseline drifts using multi-echo fMRI. *Neuroimage* 105, 189–197
82. Caballero-Gaudes, C. *et al.* (2019) A deconvolution algorithm for multi-echo functional MRI: multi-echo sparse paradigm free mapping. *Neuroimage* 202, 116081
83. Gonzalez-Castillo, J. *et al.* (2016) Evaluation of multi-echo ICA denoising for task based fMRI studies: block designs, rapid event-related designs, and cardiac-gated fMRI. *Neuroimage* 141, 452–468
84. Hajcak, G. *et al.* (2017) Psychometrics and the neuroscience of individual differences: internal consistency limits between-subjects effects. *J. Abnorm. Psychol.* 126, 823–834
85. Tholen, M.G. *et al.* (2020) Functional magnetic resonance imaging (fMRI) item analysis of empathy and theory of mind. *Hum. Brain Mapp.* 41, 2611–2628
86. Dodell-Feder, D. *et al.* (2011) FMRI item analysis in a theory of mind task. *Neuroimage* 55, 705–712
87. Wilson, K.A. *et al.* (2021) Using item response theory to select emotional pictures for psychophysiological experiments. *Int. J. Psychophysiol.* 162, 116–179
88. Naselaris, T. *et al.* (2021) Extensive sampling for complete models of individual brains. *Curr. Opin. Behav. Sci.* 40, 45–51
89. Hamilton, L.S. and Huth, A.G. (2020) The revolution will not be controlled: natural stimuli in speech neuroscience. *Lang. Cogn. Neurosci.* 35, 573–582
90. Allen, E.J. *et al.* (2021) A massive 7T fMRI dataset to bridge cognitive and computational neuroscience. *bioRxiv* Published online February 22, 2021. https://doi.org/10.1101/2021.02.22.432340
91. Esfahlani, F.Z. *et al.* (2020) High-amplitude cofluctuations in cortical activity drive functional connectivity. *Proc. Natl. Acad. Sci. U. S. A.* 117, 28393–28401
92. Tagliazucchi, E. and Laufs, H. (2014) Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. *Neuron* 82, 695–708
93. Eickhoff, S.B. *et al.* (2020) Towards clinical applications of movie fMRI. *Neuroimage* 217, 116860
94. Vanderwal, T. *et al.* (2019) Movies in the magnet: naturalistic paradigms in developmental functional neuroimaging. *Dev. Cogn. Neurosci.* 36, 100600
95. Mehrer, J. *et al.* (2021) An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc. Natl Acad. Sci. U. S. A.* 118, e2011417118
96. Hasson, U. *et al.* (2010) Reliability of cortical activity during natural stimulation. *Trends Cogn. Sci.* 14, 40–48
97. Finn, E.S. *et al.* (2017) Can brain state be manipulated to emphasize individual differences in functional connectivity? *Neuroimage* 160, 140–151
98. Finn, E.S. and Bandettini, P.A. (2021) Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *Neuroimage* 235, 117963
99. Slivkoff, S. and Gallant, J.L. (2021) Design of complex neuroscience experiments using mixed-integer linear programming. *Neuron* 109, 1433–1448
100. Button, K.S. *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376