Neuropsychopharmacology www.nature.com/npp

NEUROPSYCHOPHARMACOLOGY REVIEWS

Toward dynamic phenotypes and the scalable measurement of human behavior

Laura Germine[1,2], Roger W. Strong[1,2], Shifali Singh[1,2,3] and Martin J. Sliwinski[4]

Precision psychiatry demands the rapid, efficient, and temporally dense collection of large scale and multi-omic data across diverse samples, for better diagnosis and treatment of dynamic clinical phenomena. To achieve this, we need approaches for measuring behavior that are readily scalable, both across participants and over time. Efforts to quantify behavior at scale are impeded by the fact that our methods for measuring human behavior are typically developed and validated for single time-point assessment, in highly controlled settings, and with relatively homogeneous samples. As a result, when taken to scale, these measures often suffer from poor reliability, generalizability, and participant engagement. In this review, we attempt to bridge the gap between gold standard behavioral measurements in the lab or clinic and the large-scale, high frequency assessments needed for precision psychiatry. To do this, we introduce and integrate two frameworks for the translation and validation of behavioral measurements. First, borrowing principles from computer science, we lay out an approach for iterative task development that can optimize behavioral measures based on psychometric, accessibility, and engagement criteria. Second, we advocate for a participatory research framework (e.g., citizen science) that can accelerate task development as well as make large-scale behavioral research more equitable and feasible. Finally, we suggest opportunities enabled by scalable behavioral research to move beyond single time-point assessment and toward dynamic models of behavior that more closely match clinical phenomena.

Neuropsychopharmacology (2021) 46:209–216; https://doi.org/10.1038/s41386-020-0757-1

INTRODUCTION

In the era of computational and precision psychiatry, we have two fundamental goals: (1) first, to robustly characterize mechanisms of human behavior and brain function as they relate to human health and disease, and (2) to apply our understanding of mechanisms to individual-level prediction and individualized treatments [1]. Yet, as in any field of science that seeks to make both population and individual-level inferences, there is a tension between measuring a phenomenon comprehensively and precisely to characterize mechanisms and measuring that phenomenon across many individuals for generalizability and ultimately prediction. Indeed, the methods of behavioral science range from the comprehensive characterization of individual patients (e.g., the bilateral medial temporal lobectomy patient, H.M.) to the development of generalizable genetic prediction models of psychiatric disease based on coarse diagnostic classification (e.g., genome-wide association studies) [2]. Yet, a science of precision psychiatry will require both rich individual-level characterization and population-level scale (see Fig. 1).

The challenges of precision psychiatry will require a radical rethinking of the way we approach behavioral research, to enable the sort of data collection needed to build models for individual-level inferences. Not only must we address existing issues of power and generalizability that have been major barriers to our science [3–7], but also move toward a scale that is beyond the resources of most individual laboratories.

We use the term scalable behavioral research to refer to the application of both traditional and novel tools for measuring and quantifying behavior (e.g., surveys, sensors, cognitive assessments) at the scale necessary for population-based research, large cohort-based longitudinal studies, and high-frequency measurement designs. This includes a dramatic increase in the size and diversity of our samples, as well as the ability to characterize dynamic changes in behavior and cognition, over time.

This review focuses on challenges and a potential framework for translating methods from experimental science toward the measurement of mechanisms for cognition and behavior at scale. We outline some of the main challenges to implementing our current measures of mechanisms, adapted from experimental science, in large diverse samples. Finally, we suggest ways of reconceptualizing the development of measurement tools and the role of the participant, toward achieving a generalizable science of behavior that is rigorous, inclusive, and representative.

CHALLENGES TO SCALE

There are both financial and logistical challenges to a robust precision psychiatry, even when we constrain the problem to measurements of behavior. Below, we outline three major human, technical, and psychometric barriers to scale across the behavioral sciences. Our goal with this review is to suggest bottlenecks within behavioral research which, if addressed, would

[1]Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA, USA; [2]Department of Psychiatry, Harvard Medical School, Boston, MA, USA; [3]Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, IL, USA and [4]Center for Healthy Aging, Pennsylvania State University, State College, PA, USA
Correspondence: Laura Germine (lgermine@mclean.harvard.edu)

Received: 1 May 2020 Revised: 18 June 2020 Accepted: 25 June 2020
Published online: 6 July 2020

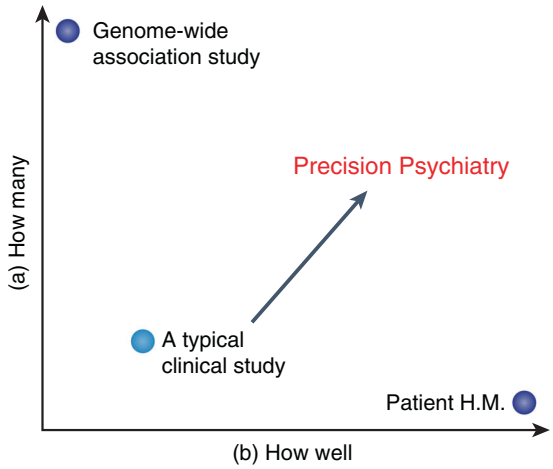

Fig. 1 The need for scale in behavioral research. Studies that achieve scale tend to do so by either measuring some characteristic across many participants, but coarsely (e.g., genome-wide association studies), or through comprehensive measurement of a limited number of participants (e.g., case studies of rare neurological phenomena, such as bilateral medial temporal lobectomy patient H. M.). Both types of scale are needed for precision psychiatry.

provide a substantial leap forward in our ability to develop a precision psychiatry.

Participant engagement

Lack of participant engagement is one of the most significant barriers to feasibility for large scale behavioral research studies. Humans research participants have resource and time limitations—they are focused on myriad concerns related to family, work, and personal needs that prevent widespread engagement in research studies [8–12]. Moreover, the attention of humans is limited. Attention is captured by information that is compelling and/or goal-related, and fatigued by information that is not [13]. Yet despite these known barriers (and most researchers being human persons themselves), we tend to design our studies primarily to meet the needs of researchers and their science. The formidable issue of participant engagement has been well-described when it comes to use of digital health apps [14, 15] as well as the limitation sections of many research studies. Anticipated burden reduces enrollment and enthusiasm for participating in research [8, 16, 17], whereas actual and perceived burden increases attrition [18, 19], reduces data quality [20], and is among the largest cost drivers for longitudinal research [21, 22]. Moreover, because burden is not evenly distributed across the population or over time [8, 17], differential recruitment and attrition by sociodemographic, diagnostic, or contextual factors threatens generalizability [8, 18, 19, 23]. Participants from already disadvantaged populations (for whom research is already potentially a prohibitive burden) and in poorer health are least likely to enroll and most likely to attrite from research studies [8–12, 19, 23, 24], making addressing participant burden a major concern for any population-level behavioral science. All else being equal, where participant burden goes up, feasibility and affordability of research goes down. To build a scalable science, attention needs to be paid to the goals and needs of participants themselves. As both the source of our science and the "end user" of our discoveries, participants matter and as a field we need to build their needs into research study design [8–10, 18, 19].

Accessibility: humans and devices

The feasibility of large-scale studies critically depends not only on whether participants are willing to engage in research, but also whether they are able to engage. Accessibility is therefore another important consideration for scalable assessment. We use the term accessibility to refer to both an individual's ability to access a measurement tool—which might be impeded by physical, logistical, linguistic, or health-related barriers—as well as how easy it is for an individual to interact with that tool [17, 25, 26].

Measurement tools are often developed with a particular participant group or scientific question in mind. When those instruments are later adapted for large-scale studies, beyond their original purpose, accessibility tends to be considered on an ad hoc basis—e.g., in a study of aging, for example, one should ensure that fonts are large enough to be readable by older adults [27]. However, if a measurement tool is to be used at scale, across large and diverse populations, then that tool must be as universally accessible as possible [28, 29]. That means considering accessibility for individuals who vary in sociodemographic factors, health status, age, education, and motivation.

What factors contribute to accessibility? Is it possible to take a "design for all" approach [30]? In addition to accommodations for individuals with different sensory or motor capabilities, differences in language or language fluency, and variations in technical skills or experience, a truly accessible tool will adhere to universal design principles that have been well articulated in the literature on human factors and user interface design [30–33]. These are principles developed to improve the operability, understandability, and perceivability of a tool across individuals [29] and emphasize clarity, simplicity, and consistency [28, 29, 34, 35]. When a research tool fails along these dimensions, it imposes a barrier not just for populations with specific sensory or motor impairments, but people with general cognitive difficulties [28], including individuals with mental disorders [29]. As with the case of participant engagement, considerations of accessibility limit how well we can reach participants with diverse needs and experiences. Moreover, the same principles that make a particular instrument more accessible will also tend to make it more engaging [36], more trustworthy [37], and improve the quality of data collected [38]. Accessibility of research tools is thus a critical component of a generalizable behavioral science.

From mechanisms to individual differences

The third major barrier to a scalable behavioral science is the measurement gap between basic and applied sciences [39]. In addition to issues of power and reproducibility discussed earlier, behavioral science is currently in the midst of a crisis of measurement that we have only begun to recognize and understand [40–44]. This arises primarily out of the drive to take measures that were developed in basic science laboratories and apply them to the study of individual differences [45].

Many of the most robust experimental measures of human cognition are poor and unreliable measures of individual differences [41]. Take, for example, the well-known and well-characterized Stroop interference effect, whereby participants are slower to name the color of word when the word text and color are incongruent (e.g., the word blue written in red) than when they are congruent (e.g., the word blue written in blue). Although a robust and replicable effect, brief measures of Stroop interference often have poor reliability [41]. Reliability, in the psychometric sense we use here, refers to the consistency of results from a particular measure. An entirely unreliable or inconsistent measure will produce different results each time, whereas a perfectly reliable measure will produce the same result every time. Reliability can be further divided into test–retest reliability vs. internal reliability. Test–retest reliability refers to the consistency of a measure over longer time periods (beyond a single measurement or test session) whereas internal reliability refers to the consistency of a measure over the time period that measure is delivered or administered. While a measure may have poor test–retest reliability and still be valid (e.g., if the underlying process or behavior being measured is unstable), measures with

poor internal reliability are either measuring distinct constructs across test items or, in some cases, not measuring anything at all. Reliability, in some form, is a prerequisite for validity. Returning to our Stroop example: it is not sufficient to confirm that participants have slower response times for incongruent than congruent trials. Rather, the magnitude of an individual's response time slowing on incongruent trials should be consistent within a test session or between test sessions if a particular measure of Stroop interference is to be considered reliable.

A growing literature in affective, social, and cognitive sciences have identified foundational reliability issues with some of the most widely used and richly characterized measures [40–48]. At best, unreliable measurement reduces power and reproducibility of research studies. At worst, unreliable measurement means that ostensible variations in behavior may reflect random variations between people or over time, and are fundamentally uninterpretable [49].

Why does the translation gap for behavioral measurement exist? Part of the problem is that many areas of behavioral science do not have a tradition of reporting reliability statistics [43]. There is, however, a more fundamental issue that is related to the types of variance that are the focus of the basic sciences, including neuroscience and experimental psychology [45].

When we seek to characterize variations in mechanisms across individuals, we draw from a rich and diverse basic science literature whose goal is to characterize those mechanisms. The Stroop interference effect, for example, has helped us understand processes related to automaticity, selective attention, and response inhibition. In translational and clinical science, we want to be able to take mechanisms—such as response inhibition—and look at how variations in those mechanisms might contribute to differences in disease risk and selection of appropriate treatments [1]. Yet, measurement approaches that are the most sensitive to differences between conditions often have the least variability between persons. The Stroop effect is so well-characterized in experimental psychology precisely because nearly all individuals show the expected pattern of response times to incongruent vs. congruent trials. Rather than being sensitive to individual differences, the optimal scenario for experimental validation is if a mechanism (or its measurement) is as invariant as possible across individuals [45]. However, as sensitivity to between-person individual differences is a prerequisite for understanding how variations in mechanisms contribute to human disease, the assumption of reliable between-person variability must be tested [41, 43].

In summary, to understand variations in mechanisms, we need to take the challenge of translation of measurement tools from basic science far more seriously. Such considerations are a foundational part of a scalable behavioral science, and should be central to our study design, interpretation of results, and overall scientific priorities.

FRAMEWORKS FOR SCALING BEHAVIORAL MEASUREMENT

The limitations articulated above are daunting and, when considered together, paint a negative picture of the feasibility of a broadly scalable behavioral science to drive progress in psychiatry. Yet, we note that these challenges are not restricted to large-scale behavioral research, but exist, in some form, across human individual differences and clinical research. The drive toward larger-scale studies, diagnostics, and interventions acts as a lens—magnifying and bringing into focus the many barriers and limitations that already existed within the silos of our laboratories, institutions, or subfields. The goal in addressing these issues is to build better models of human behavior and disease, advance the progress of science, and ultimately develop better treatments.

In this section, we provide two approaches to behavioral research we believe will help address the barriers described in the previous section. These are approaches used by the authors in their own work, but are not the only potential solutions. Rather, our goal is to spark a conversation about ways that we might reconceptualize the research process and research laboratory in behavioral science, to make scalable research more feasible.

Iterative task development

Our current approach to the development of research methods and studies is approximately linear. Once a basic mechanism has been identified, research measurement tools (or tasks) initially developed to characterize that mechanism are adapted for an applied or clinical context. These tasks may then be piloted to assess feasibility and basic aspects of validity in a small sample drawn from a target patient population or among healthy controls. In this initial piloting phase, perhaps it is discovered that a particular task or condition produces "better" data than another (based on a diverse and heterogeneous set of criteria). This then informs selection of tasks and measures for a larger study. As noted above, the reporting of reliability metrics at this stage is inconsistent and, in some subfields, regularly omitted. If all goes well, a larger study is eventually conducted, leading to a mixture of negative and positive results that then enter the research literature and contribute to progress (or not) in a particular subfield.

We would argue that this process often goes awry at the earliest stage of task development—the translation of tasks from basic science to clinical research (or the study of individual differences). In addition to unknown reliability, many tasks are never evaluated for their participant burden characteristics or accessibility across populations, devices, and contexts. Usually it is only after the task has been used in a large study or several studies that it is recognized that the task falls short along one of these dimensions. A reasonable approach for addressing participant engagement, accessibility, and psychometric generalizability is necessary for macroscale behavioral research.

We look to computer science (a field of engineering) for potential solutions. Consumer-oriented software development, in particular, has evolved best practices for the development of applications geared toward addressing many of the same human and logistical barriers articulated above for measurement tools. Such software applications will fail if they are not useable, engaging, accessible, and scalable. Importantly, and like in the case of our research tools, it is often not clear what precise parameters or characteristics will lead to maximum useability, engagement, accessibility, and scalability.

The model used throughout software development—and in other areas of engineering and design—relies on iterative refinement and randomization (also known as A/B testing) [50–52]. That is, it is not enough to build an application and assume (based on first principles) that it will work as intended. Rather, a part of the application development process is the successive validation and refinement of the application along multiple simultaneous criteria. And, as in behavioral science, randomization of users to different test conditions (A/B testing) permits the selection of parameters, features, and user interface characteristics in a data-driven and unbiased manner [52]. Here, we describe the application of such an iterative A/B testing framework to the development of measurement tools focused on cognition (see Fig. 2).

Iterative task development begins with the selection of parameters for a particular task (overall task procedure, items, length, instructions, formatting, etc). This defines an initial prototype for further development. The prototype is based on a best guess of what has worked previously, either based on existing research in similar populations or based on measures for which there is a strong foundation of experimental science with well-characterized mechanisms. Next, additional parameters are selected that might be expected to change behavior: for instance,

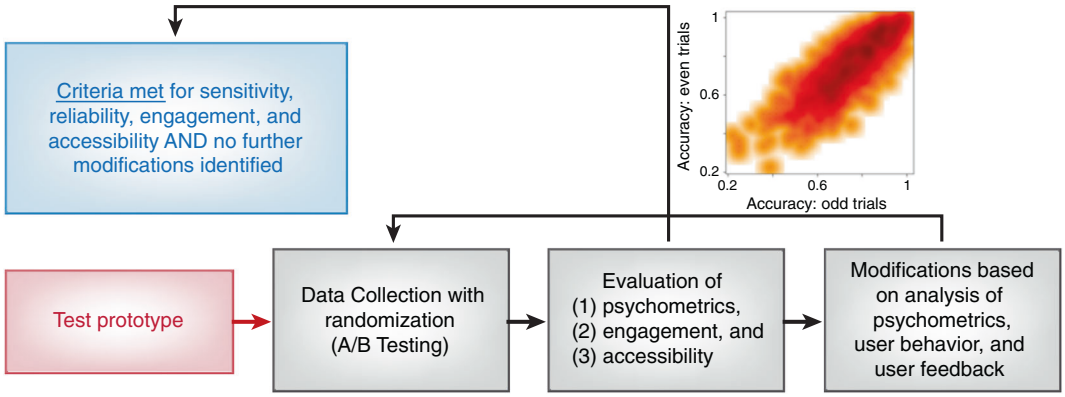

Fig. 2 Iterative task development. Shown is a schematic of a basic iterative task development procedure. The inset graph shows an example visualization of reliability for an accuracy-based cognitive task. Measurement reliability is an often-neglected characteristic of assessments adapted from experimental/basic science.

differences in instructions, methods of delivery (auditory vs. visual), methods of eliciting response (e.g., likert vs. T/F for questionnaires), and incentives for participation (e.g., return of research results, payment schedules or triggers, lottery). Criteria should be set in advance for determining whether a task meets some minimally acceptable standard of reliability, sensitivity, validity, generalizability, accessibility, and engagement across participants. Participants are then randomized to different versions of the task. The impacts of different task parameters are then evaluated and decisions are made about which parameters led to improvements and which did not. At that point, these parameters can be refined for subsequent rounds of A/B testing and task development. Notably, in this model, some measures would never exit the development stage as no combination of parameters tested yield versions of the test that meet criteria for minimum acceptability. Based on a priori criteria, these tests would not be considered appropriate for wide scale deployment in research studies designed to produce generalizable knowledge. Task development is complete when no further improvements are identified.

Approaches that rely on randomization to different items or parameters for task development and validation are not novel. Similar general models are used in the development of measures that rely on item banks—a large number of potential test or survey items are generated and then tested (using random assignment) to estimate psychometric characteristics of each item, allowing items with better psychometric characteristics to be identified and used in the development of custom applications (e.g., the SAPA Project) [53, 54]. The metastudy approach [55] similarly relies on randomization of participants to many possible variations of a task or experiment. The purpose of the metastudy, however, is to determine whether a particular effect or outcome is robust to variability across a range of nuisance parameters [55], rather than task development. Our iterative task development framework extends the logic of these item and parameter randomization approaches to include multiple successive (iterative) phases of task optimization and metrics that capture human factors such as accessibility and engagment.

When moving beyond psychometric criteria for task development, it is expected that better optimization along one dimension can lead to poorer optimization along another dimension. For example, one way to limit variability in scores due to non-human sources (e.g., for a measure of simple reaction time) is to limit the types of devices and contexts where a measure can be completed. iOS devices, for example, tend to have shorter response time latencies than Android devices (due to the latter's variation in hardware) making them better suited as a class for measuring response times. At the same time, however, the lower cost of many Android smartphones means that the average education and socioeconomic status of iOS years tends to be higher than Android users [56]—factors that have robust and replicable associations with cognition and mental health. Thus, limiting a study to iOS devices will improve precision of measurement (and reduce the influence of a potential confound), but also exclude the majority of smartphone users [56].

Another example is the tension between task length and task reliability. The most robust and generalizable method for increasing the reliability of a measure is to increase its length. Unfortunately, increased test length or administration time contributes to participant burden—which reduces enrollment, increases attrition, and can threaten generalizability [16, 21]. While this concern is potentially less applicable for passive data collection (e.g., gps, actigraphy, or other sensor based modalities), more dense or frequent measurement can interfere with device processing speed or battery life, which can interfere with the participant's use of a device and increase burden [57]. Across modalities, more precise or more comprehensive measurement is usually more burdensome.

One way to address the trade-off between psychometric and useability considerations is to create joint optimization metrics. For instance, one can use a metric that captures both task reliability and participant burden by looking at the minimum duration of a task needed for acceptable reliability across tasks or versions of a task (or minDAR) [44]. Based on an analysis of 25 cognitive tests, Passell et al. [44] reported dramatic variation in the task duration needed to produce acceptably reliable scores (defined as an internal reliability of at least $r = 0.7$). Some tasks had minDARs of only 30 s. For other tasks, the minDAR exceeded 10 min [44]. One might posit similar such joint optimization metrics for looking at the minimum duration for acceptable validity (based on associations with some predefined criterion) that allows comparison across task parameters.

There are two potential critiques of the iterative task development approach described here. First, optimization of certain parameters might threaten the validity and generalizability of the original task. That is, there is a risk that the more a task is modified, the less likely it is that the existing literature and validation for that task (including literature on basic mechanisms) can be applied. This is a valid concern, and appropriate checks should be included in the task development process to track validity. One might evaluate, for example, whether modifications that improve accessibility and reduce burden reduce the magnitude of important between condition effects. A brief and reliable measure of Stroop interference (if one exists [41]) should still have longer reaction times for incongruent trials than congruent trials. Otherwise, it is not a measure of Stroop interference. As limitations

related to participant engagement, accessibility, and measurement already pose a threat to validity and generalizability, we believe that systematic efforts to address these barriers will tend to improve capacity for scalable assessment that advances behavioral science.

The second critique is that the sample sizes needed for an iterative task development approach are prohibitive for most studies. This is also a valid concern. We focus in this manuscript specifically on the contexts where applications of behavioral or cognitive measures at macroscale are both desirable and potentially achievable. In these cases, the investment in a robust iterative task development phase for translating such measures toward scalable contexts can save both human and financial resources in the long run. We also note that there are now many low cost and high throughput methods for large sample participant recruitment that do not require the same level of resource as a large traditional research study. If it is not possible to iteratively develop a measurement meant for schizophrenia research using a large sample of schizophrenia patients, a reasonably good first approximation of task psychometrics, accessibility, and burden can be made using large, diverse samples of mostly healthy participants. There are now numerous platforms for recruiting large numbers of participants to complete research assessments, including Amazon's Mechanical Turk [58], Prolific [59], and Crowdflower [60]. These platforms can be a rapid and inexpensive source of participants, but researchers should be aware of their challenges and limitations [58].

Yet another approach for engaging participants is a citizen science model of recruitment, which treats the participant as a partner in the scientific discovery process. This approach is described in the next section.

Citizen science: participant as collaborator

Patient-centered, participant-centered, and/or participatory research frameworks have received a lot of attention over the past decade, and with good reason: the integration of the patient or participant perspective into research at all stages makes both practical and ethical sense [61–63]. Such an approach can help identify new opportunities [64] or fundamental design flaws [10] early in the research process. It also makes the identification and selection of incentives for participation both more comprehensive and clear [9, 10, 18, 24, 65].

Here, we focus on a citizen science framework for participatory research in behavioral science [65, 66]. In this framework, participation in research is incentivized by the desire to contribute to science, insight into the research question being studied, as well as return of study data and individual research results [65–68]. Participants use structured research tools to answer their own research questions or contribute to an overall research program by collecting their own data [69]. In behavioral science, that data collection involves completing surveys, behavioral measures, and cognitive tasks that provide individual-level feedback about major outcome variables or performance. The benefit of data collection using this model is threefold. First, the incentives are aligned for participant and researcher: they both want to understand the participant's capabilities [65]. Participants will tend to exert effort toward better performance in a way that fits the assumptions of our research studies and can produce higher quality data than financial incentives [70]. Second, participants can recruit other participants in a way that leads to large scale participation. TestMyBrain.org, for example, receives about 500–1000 participants per day, of whom about 2/3 are new to research participation [44, 65, 67]. Third, it invites and encourages participants to provide feedback on research methods that can help generate insights about potential technical problems, issues with instructions, user interface improvements, or accessibility barriers that would otherwise be difficult to identify.

Many of the major reservations that researchers have about this approach to data collection are around the ethics of return of research results—specifically, where each participant is provided with their data or some individual-level metric derived from their data. How will a participant interpret the data [71]? Will they know what to do with it [72]? What if results cause distress or lead to decisions about treatment-seeking or care that ultimately have a negative impact [73]? One might flip this question, however, by asking: who has the most fundamental right to a participant's data? If data are generated using the body and behavior of an individual, should researchers have the right to limit that individual's access to that data? Rather than asking whether data should be shared with the participant, we should perhaps be asking how to best share data with participants [74, 75]. These are important considerations that are currently being deeply considered elsewhere as part of national and international initiatives [71–77], including the US Precision Medicine Initiative (All of Us research program) [76].

In the case of low risk measures where nonclinical interpretations of scores can be made interesting and understandable, we and others have found the return of research results to be a positive incentive for community education [78], engagement in research [65, 67], and developing relationships with participant communities that enhance research and improve public understanding of science [64]. In addition to TestMyBrain.org [65], initiatives that have had similar (or greater!) success at recruiting citizen science participants for studies of human cognition and behavior through return of research results include LabintheWild [79], Games with Words [67], Project Implicit [80], My Social Brain [81], and the SAPA Project [53]. While not suitable for all test development modalities or applications, the combination of crowdsourcing and/or citizen science approaches allow rapid evaluation of task characteristics like participant burden, reliability, and accessibility at relatively low cost. It remains to be seen whether digital citizen approaches can address engagement barriers in longitudinal research, where personal relationships (e.g., with research personnel) can also serve as engagement incentives. As with any method that relies on digital technology, targeted outreach will also be needed to ensure participation among communities with reduced access to smartphone technologies, including rural and underserved communities.

FUTURE RESEARCH DIRECTIONS

We have argued that, as we move toward precision psychiatry, there is a pressing need for broadly scalable behavioral research approaches—the development and validation of reliable, accessible, engaging, and generalizable methods is the only way to achieve a precision psychiatry that can precisely and dynamically characterize the behavior of many individuals, over time (Fig. 1). Below, we outline an emerging new field of behavioral science that focuses on temporal dynamics of cognition and behavior, enabled by new technologies and approaches to assessment, that could transform psychiatry and our understanding of the human mind.

Behavior, over time

Cross-sectional psychiatric research often implicitly assumes that variations in mechanisms that are associated with psychopathology between persons can generalize to variations in symptoms or psychopathology within a person over time. Yet, many psychological processes and mechanisms violate this ergodicity assumption [82].

One of the most exciting innovations that is enabled by digital technology and the shift toward larger-scale behavioral data collection is the ability to measure and monitor change over time in behavior and cognition. Typical approaches to measuring change rely on longitudinal "single-shot" designs [83], in which

single time-point assessments are repeated across widely spaced intervals (e.g., annual assessments). The single-shot approach assumes that differences in cognition and behavior are relatively stable over brief time scales, and that meaningful change occurs relatively slowly. Ecological momentary assessment and measurement burst designs have revealed significant variability in cognition and behavior over hours and days, however, where distinct patterns of variability are associated with different behaviors [84], psychiatric risk [85], and the effectiveness of interventions [86]. Cognition, in particular, demonstrates significant within-person variability that accounts for as much as 40–60% of the total variability in performance [87].

Reliance on single-shot assessments to measure what are likely dynamic processes has two important consequences. First, it introduces temporal sampling error—or differences in measurement that reflect time-of-testing effects that can differ substantially from a person's average [83]. Second, single time-point assessments assume that variability in performance is not meaningful for characterizing phenotypes. This is a relatively unsupported assumption, and one that is at odds with recent conceptualizations of dynamic phenotypes, a term originally coined to refer to time-dependent observable characteristics of single cells [88]. Because human behavior and performance are time-dependent, and display meaningful variability at a relatively fast time-scale (e.g., moments, hours, days), precisely characterizing important phenotypes require tools that can capture behavior as it unfolds in as near to real-time as possible [89].

Behavior, in context

Finally, new technologies for measuring physiology, mood, and environmental variables can now provide richness and context for more traditional behavioral assessments—fully removing the laboratory from the confines of brick and mortar and into people's everyday environments. Digital sensors embedded in everyday wearables and personal digital devices can measure movement, sleep, vocal patterns, and even physiological signals that are related to mood, arousal, and health [90–92]. While some of the most innovative applications involve extracting signals from dense multimodal datasets that combine sensors using machine learning for prediction and diagnosis [92], there are also more immediate applications that make traditional tools both more powerful and more interpretable [93]. Processing of speech from voice and text can be used to rapidly extract information from sources that are otherwise hard to process [93, 94], as well as provide indicators of emotion and psychological status [90, 95]. Sensors embedded in smartphones, together with active measures of behavior such as surveys or cognitive tests, can give information about the context in which a behavior, experience, or cognitive process occurs [96]. Computer vision algorithms can take slices of data from human video and images to understand things like emotional experiences, social behavior, and attention [97, 98]. Actigraphy or sleep data can provide information about circadian rhythms that might track fluctuations in behavior or cognitive performance that provide meaningful signals related to brain and cognitive health [99]. Such applications are being widely tested by researchers in the field and time will tell which provide the most promising signals related to human cognition and behavior.

As human beings moving through the world, our behavior is both dynamic and exquisitely responsive to social and environmental contexts. Methods that allow us to access that dynamic and context-rich view of human cognition and behavior will open up new areas of investigation and potentially provide better models for understanding psychopathology. The coming years may reveal new architectures of human cognition and behavior based on temporal variation or state-related change, which can yield insights into the pathophysiology of mental disorders that were previously inaccessible due to methods limitations.

CONCLUSION

The scaling of methods for the assessment of health-related characteristics is happening throughout science and medicine, owing to the explosion of new technologies, new analytic approaches, and the unprecedented connectedness of human societies. We are now able to conduct research at a scale that was previously unimaginable.

In this review, we have attempted to lay out our view of some of the major considerations for scaling the science of behavior across individuals and over time, as well as potential approaches that emphasize iterative design of reliable, engaging, and accessible measures, together with thoughtful integration of participants in the research process. In no way, however, do we imply that the solutions suggested are the only path forward. Indeed, one of the most exciting things about the shift from individual investigators to communities of scientists and participants working together on ambitious projects is the potential to rethink our assumptions about the research process, where it is centered, and how best to drive scientific progress.

FUNDING AND DISCLOSURE

Funding was provided by a National Institutes of Health grant to LG (NIMH R01MH121617) and National Institutes of Health grant to MJS (NIA U02AG060408). LG has received compensation as a member of the scientific advisory board of Sage Bionetworks, a 501c3 nonprofit organization. LG is on the Board of Directors of the Many Brains Project, a 501c3 nonprofit organization. The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

LG drafted the manuscript based on input from RS, SS, and MJS. All authors read and approved the final manuscript.

REFERENCES

1. Redish AD, Gordon JA, editors. Computational psychiatry: new perspectives on mental illness. Cambridge, MA: MIT Press; 2016.
2. Yehia L, Eng C. Largescale population genomics versus deep phenotyping: brute force or elegant pragmatism towards precision medicine. NPJ Genome Med. 2019;4:1–2.
3. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349:aac4716.
4. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14:365–76.
5. Szucs D, Ioannidis JP. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 2017; 15:1–18.
6. Henrich J, Heine SJ, Norenzayan A. Beyond WEIRD: towards a broad-based behavioral science. Behav Brain Sci. 2010;33:111.
7. Williams DR, Jackson PB. Social sources of racial disparities in health. Health Aff. 2005;24:325–34.
8. Brown G, Marshall M, Bower P, Woodham A, Waheed W. Barriers to recruiting ethnic minorities to mental health research: a systematic review. Int J Meth Psychiatr Res. 2014;23:36–48.
9. Arean PA, Alvidrez J, Nery R, Estes C, Linkins K. Recruitment and retention of older minorities in mental health services research. Gerontologist. 2003;43:36–44.
10. Chen H, Kramer EJ, Chen T, Chung H. Engaging Asian Americans for mental health research: challenges and solutions. J Immigr Health. 2005;7:109–18.
11. Le HN, Lara MA, Perry DF. Recruiting Latino women in the US and women in Mexico in postpartum depression prevention research. Arch Women's Ment Health. 2008;11:159–69.
12. Miranda J. Introduction to the special section on recruiting and retaining minorities in psychotherapy research. J Consult Clin Psychol. 1996;64:848.
13. Cohen RA, Sparling-Cohen YA, O'Donnell BF. The neuropsychology of attention. New York, NY: Plenum Press; 1993.
14. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health. 2018;21:116–9.

15. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv. 2019;70:538–44.
16. Apodaca R, Lea S, Edwards B. The effect of longitudinal burden on survey participation. In: Proceedings of the Survey Research Methods Section. American Statistical Association; 1998.
17. Kerr DC, Ornelas IJ, Lilly MM, Calhoun R, Meischke H. Participant engagement in and perspectives on a web-based mindfulness intervention for 9-1-1 telecommunicators: multimethod study. J Med Internet Res. 2019;21:e13449.
18. Yancey AK, Ortega AN, Kumanyika SK. Effective recruitment and retention of minority research participants. Annu Rev Public Health. 2006;27:1–28.
19. Gilliss CL, Lee KA, Gutierrez Y, Taylor D, Beyene Y, Neuhaus J, et al. Recruitment and retention of healthy minority women into community-based longitudinal research. J Wom Health Gend Base Med. 2001;10:77–85.
20. Musthag M, Raij A, Ganesan D, Kumar S, Shiffman S. Exploring micro-incentive strategies for participant compensation in high-burden studies. In: Proceedings of the 13th International Conference on Ubiquitous Computing; 2011. p. 435–44.
21. Loxton D, Young A. Longitudinal survey development and design. Int J Mult Res Approaches. 2007;1:114–25.
22. Anguera JA, Jordan JT, Castaneda D, Gazzaley A, Areán PA. Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense. BMJ Innov. 2016;2:14–21.
23. Ejiogu N, Norbeck JH, Mason MA, Cromwell BC, Zonderman AB, Evans MK. Recruitment and retention strategies for minority or poor clinical research participants: lessons from the Healthy Aging in Neighborhoods of Diversity across the Life Span study. Gerontologist. 2011;51:S33–45.
24. Loue S, Sajatovic M. Research with severely mentally ill Latinas: successful recruitment and retention strategies. J Immigr Minor Health. 2008;10:145–53.
25. Anderson ML, Riker T, Hakulin S, Meehan J, Gagne K, Higgins T, et al. Deaf ACCESS: adapting consent through community engagement and state-of-the-art simulation. J Def Stud Deaf Educ. 2020;25:115–25.
26. Deering S, Grade MM, Uppal JK, Foschini L, Juusola JL, Amdur AM, et al. Accelerating research with technology: rapid recruitment for a large-scale web-based sleep study. JMIR Res Protoc. 2019;8:e10974.
27. Zaphiris P, Kurniawan S, Ghiawadwala M. A systematic approach to the development of research-based web design guidelines for older people. Univers Access Inf Soc. 2007;6:59.
28. Friedman MG, Bryen DN. Web accessibility design recommendations for people with cognitive disabilities. Technol Disabil. 2007;19:205–12.
29. Bernard R, Sabariego C, Cieza A. Barriers and facilitation measures related to people with mental disorders when using the web: a systematic review. J Med Internet Res. 2016;18:e157.
30. Akoumianakis D, Stephanidis C. Universal design in HCI: a critical review of current research and practice. Eng Constr. 1989;754.
31. McCarthy JE, Swierenga SJ. What we know about dyslexia and web accessibility: a research review. Univers Access Inf Soc. 2010;9:147–52.
32. Nordhoff M, August T, Oliveira NA, Reinecke K. A case for design localization: diversity of website aesthetics in 44 countries. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018. p. 1–12.
33. Gajos KZ, Chauncey K. The influence of personality traits and cognitive load on the use of adaptive user interfaces. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. 2017. p. 301–6.
34. Eraslan S, Yaneva V, Yesilada Y, Harper S. Web users with autism: eye tracking evidence for differences. Behav Inf Technol. 2019;38:678–700.
35. Schwartz AE, Kramer JM, Longo AL. Patient-reported outcome measures for young people with developmental disabilities: incorporation of design features to reduce cognitive demands. Dev Med Child Neurol. 2018;60:173–84.
36. Hawthorn D. Interface design and engagement with older people. Behav Inf Technol. 2007;26:333–41.
37. Lindgaard G, Dudek C, Sen D, Sumegi L, Noonan P. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. ACM Trans Comput Hum Interact. 2011;18:1–30.
38. Finnerty A, Kucherbaev P, Tranquillini S, Convertino G. Keep it simple: reward and task design in crowdsourcing. In: Proceedings of the biannual conference of the Italian chapter of SIGCHI. New York, NY: Association for Computing Machinery; 2013. p.1–4.
39. Kosslyn SM, Cacioppo JT, Davidson RJ, Hugdahl K, Lovallo WR, Spiegel D, et al. Bridging psychology and biology: the analysis of individuals in groups. Am Psychol. 2002;57:341.
40. Enkavi AZ, Eisenberg IW, Bissett PG, Mazza GL, MacKinnon DP, Marsch LA, et al. Large-scale analysis of test–retest reliabilities of self-regulation measures. Proc Nat Acad Sci. 2019;116:5472–7.
41. Hedge C, Powell G, Sumner P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav Res Methods. 2018;50:1166–86.
42. McNally RJ. Attentional bias for threat: crisis or opportunity? Clin Psychol Rev. 2019;69:4–13.
43. Parsons S, Kruijt AW, Fox E. Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. Adv Methods Pract Psychol Sci. 2019;2:378–95.
44. Passell E, Dillon DG, Baker JT, Vogel SC, Scheuer LS, Mirin NL, et al. Digital cognitive assessment: results from the TestMyBrain NIMH Research Domain Criteria (RDoC) field test battery report. Psyarxiv. 2019. https://doi.org/10.31234/osf.io/dcszr.
45. Plomin R, Kosslyn SM. Genes, brain and cognition. Nat Neurosci 2001;4:1153–4.
46. Rodebaugh TL, Scullin RB, Langer JK, Dixon DJ, Huppert JD, Bernstein A, et al. Unreliability as a threat to understanding psychopathology: the cautionary tale of attentional bias. J Abnorm Psychol. 2016;125:840.
47. Kappenman ES, Farrens JL, Luck SJ, Proudfit GH. Behavioral and ERP measures of attentional bias to threat in the dot-probe task: poor reliability and lack of correlation with anxiety. Front Psychol. 2014;5:1368.
48. Waechter S, Nelson AL, Wright C, Hyatt A, Oakman J. Measuring attentional bias to threat: reliability of dot probe and eye movement indices. Cogn Ther Res. 2014;38:313–33.
49. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52:281.
50. Basil VR, Turner AJ. Iterative enhancement: a practical technique for software development. IEEE Trans Softw Eng. 1975;4:390–6.
51. Nielsen J. Iterative user-interface design. Computer. 1993;26:32–41.
52. Kohavi R, Longbotham R. Online controlled experiments and A/B testing. Encycl Mach Learn Data Min. 2017;7:922–9.
53. Condon DM, Revelle W. The international cognitive ability resource: development and initial validation of a public-domain measure. Intelligence. 2014;43:52–64.
54. Condon DM, Revelle W. Selected ICAR data from the SAPA-Project: development and initial validation of a public-domain measure. J Open Psychol Data. 2016;4.
55. Baribault B, Donkin C, Little DR, Trueblood JS, Oravecz Z, van Ravenzwaaij D, et al. Metastudies for robust tests of theory. Proc Nat Acad Sci. 2018;115:2607–12.
56. Germine L, Reinecke K, Chaytor NS. Digital neuropsychology: challenges and opportunities at the intersection of science and software. Clin Neuropsychol. 2019;33:271–86.
57. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: Longitudinal, Observational Feasibility Study. JMIR mHealth uHealth. 2020;8:e14368.
58. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality data? In: Kazdin E, editor. Methodological issues and strategies in clinical research. 2016. p. 133–9.
59. Palan S, Schitter C. Prolific. ac—a subject pool for online experiments. J Behav Exp Financ. 2018;17:22–7.
60. Van Pelt C, Sorokin A. Designing a scalable crowdsourcing platform. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. 2012. p. 765–6.
61. Cornwall A, Jewkes R. What is participatory research? Soc Sci Med. 1995;41:1667–76.
62. Minkler M, Wallerstein N, editors. Community-based participatory research for health: from process to outcomes. San Francisco, CA: John Wiley & Sons; 2011.
63. Horowitz CR, Robinson M, Seifer S. Community-based participatory research from the margin to the mainstream: are researchers prepared? Circulation. 2009;119:2633–42.
64. Duchaine B, Germine L, Nakayama K. Family resemblance: ten family members with prosopagnosia and within-class object agnosia. Cogn Neuropsychol. 2007;24:419–30.
65. Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. Psychon Bull Rev. 2012;19:847–57.
66. Oliveira N, Jun E, Reinecke K. Citizen science opportunities in volunteer-based online experiments. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017. p. 6800–12.
67. Hartshorne JK, Germine LT. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. Psychol Sci. 2015;26:433–43.
68. Jun E, Hsieh G, Reinecke K. Types of motivation affect study selection, attention, and dropouts in online experiments. Proc ACM Hum Comput Interact. 2017;1:1–5.
69. Li Q, Gajos KZ, Reinecke K. Volunteer-based online studies with older adults and people with disabilities. In: Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility. 2018. p. 229–41.
70. Ye T, Reinecke K, Robert Jr LP. Personalized feedback versus money: the effect on reliability of subjective data in online experimental platforms. In: Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. New York, NY: Association for Computing Machinery; 2017. p. 343–6.
71. Fabsitz RR, McGuire A, Sharp RR, Puggal M, Beskow LM, Biesecker LG, et al. Ethical and practical guidelines for reporting genetic research results to study

participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. Circ Cardiovasc Genet. 2010;3:574–80.
72. Wallace SE, Kent A. Population biobanks and returning individual research results: mission impossible or new directions? Hum Genet. 2011;130: 393–401.
73. Burke W, Evans BJ, Jarvik GP. Return of results: ethical and legal distinctions between research and clinical care. Am J Med Genet Part C Semin Med Genet. 2014;166C:105–11.
74. Fernandez CV, Kodish E, Weijer C. Informing study participants of research results: an ethical imperative. IRB: Ethics Hum Res. 2003;25:12–9.
75. Jarvik GP, Amendola LM, Berg JS, Brothers K, Clayton EW, Chung W, et al. Return of genomic results to research participants: the floor, the ceiling, and the choices in between. Am J Hum Genet. 2014;94:818–26.
76. Sankar PL, Parker LS. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. Genet Med. 2017;19:743–50.
77. Wong CA, Hernandez AF, Califf RM. Return of research results to study participants: uncharted and untested. JAMA. 2018;320:435–6.
78. Macdonald K, Germine L, Anderson A, Christodoulou J, McGrath LM. Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. Front Psychol. 2017;8:1314.
79. Reinecke K, Gajos KZ. LabintheWild: conducting large-scale online experiments with uncompensated samples. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. 2015. p. 1364–78.
80. Xu K, Nosek B, Greenwald A. Psychology data from the race implicit association test on the project implicit demo website. J Open Psychol Data. 2014;2.
81. Thornton MA, Tamir D. Six dimensions describe action understanding: the ACT-FASTaxonomy. PsyArxiv. 2019. https://doi.org/10.31234/osf.io/gt6bw.
82. Molenaar PC, Campbell CG. The new person-specific paradigm in psychology. Cur Dir Psychol. 2009;18:112–7.
83. Sliwinski MJ. Measurement-burst designs for social health research. Soc Pers Psychol Compass. 2008;2:245–61.
84. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol. 2008;4:1–32.
85. Russell MA, Gajos JM. Annual research review: ecological momentary assessment studies in child psychology and psychiatry. J Child Psychol Psychiatry. 2020;61: 376–94.
86. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol. 2010;15:1–39.
87. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. Assessment. 2018;25:14–30.
88. Ruderman D. The emergence of dynamic phenotyping. Cell Biol Toxicol. 2017; 33:507–9.
89. Ram N, Gerstorf D. Time-structured and net intraindividual variability: tools for examining the development of dynamic characteristics and processes. Psychol Aging. 2009;24:778.
90. Baker JT, Germine LT, Ressler KJ, Rauch SL, Carlezon WA. Digital devices and continuous telemetry: opportunities for aligning psychiatry and neuroscience. Neuropsychopharmacology. 2018;43:2499–503.
91. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology. 2016;41:1691–6.
92. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology. 2018;43:1660–6.
93. McCoy TH, Castro VM, Roberson AM, Snapper LA, Perlis RH. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatr. 2016;73:1064–71.
94. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 2015;1:15030.
95. Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatr. 2018;17:67–75.
96. Murphy E, King EA. Testing the accuracy of smartphones and sound level meter applications for measuring environmental noise. Appl Acoust. 2016;106:16–22.
97. Harati S, Crowell A, Mayberg H, Kong J, Nemati S. Discriminating clinical phases of recovery from major depressive disorder using the dynamics of facial expression. In: Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2016. p. 2254–7.
98. Campbell K, Carpenter KL, Hashemi J, Espinosa S, Marsan S, Borg JS, et al. Computer vision analysis captures atypical attention in toddlers with autism. Autism. 2019;23:619–28.
99. Jones SH, Hare DJ, Evershed K. Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder. Bipolar Disord. 2005;7:176–86.