Review

Biological Psychiatry: CNNI

Challenges and Solutions to the Measurement of Neurocognitive Mechanisms in Developmental Settings

Patrizia Pezzoli, Sam Parsons, Rogier A. Kievit, Duncan E. Astle, Quentin J.M. Huys, Nikolaus Steinbeis, and Essi Viding

ABSTRACT

Identifying early neurocognitive mechanisms that confer risk for mental health problems is one important avenue as we seek to develop successful early interventions. Currently, however, we have limited understanding of the neurocognitive mechanisms involved in shaping mental health trajectories from childhood through young adulthood, and this constrains our ability to develop effective clinical interventions. In particular, there is an urgent need to develop more sensitive, reliable, and scalable measures of individual differences for use in developmental settings.

In this review, we outline methodological shortcomings that explain why widely used task-based measures of neurocognition currently tell us little about mental health risk. We discuss specific challenges that arise when studying neurocognitive mechanisms in developmental settings, and we share suggestions for overcoming them. We also propose a novel experimental approach—which we refer to as "cognitive microscopy"—that involves adaptive design optimization, temporally sensitive task administration, and multilevel modeling. This approach addresses some of the methodological shortcomings outlined above and provides measures of stability, variability, and developmental change in neurocognitive mechanisms within a multivariate framework.

https://doi.org/10.1016/j.bpsc.2023.03.011

Multiple developmental theories implicate the role of specific neurocognitive mechanisms in the onset and maintenance of mental health symptoms (1,2). Research has indicated that neurocognitive difficulties in childhood and adolescence, such as poor self-control, may represent transdiagnostic risk factors for psychopathology (3,4). Therefore, a better understanding of the relationships between neurocognitive mechanisms and mental health across development is needed to develop effective preventative interventions that can reduce that risk for affected individuals, families, and societies.

Accurate measurement of neurocognitive mechanisms entails the administration of task-based measures wherein participants respond to stimuli that putatively engage the mechanisms of interest. For example, in a Go/NoGo task, participants must suppress responses to NoGo stimuli, and making fewer errors on NoGo trials indicates better self-control (5). Task-based measures designed to tap neurocognitive mechanisms are widely used to study clinical and at risk groups, but recent concerns regarding the psychometric properties of these task-based measures and their sensitivity to individual differences (6–8) have forced the field to take stock of its methods. However, relatively little attention has been devoted to the implications of these concerns for developmental research. If we cannot measure neurocognitive mechanisms reliably and sensitively during development, then we are limited in our ability to discern the nature of mental health vulnerability and to develop early interventions that may prevent mental health symptoms from emerging or escalating.

In this article, we review the challenges associated with measuring neurocognitive mechanisms using task-based measures and their implications for developmental research. In addition to pointing out challenges, we suggest some potentially fruitful avenues to advance the field. More specifically, we propose triangulating methods that have been proven successful individually or in research with adult populations. We also explain how this approach may yield psychometrically valid measurements of individual differences in neurocognitive mechanisms relevant to mental health vulnerability. Here, we focus on behavioral measurements because the proposed approach readily lends itself to behavioral experimentation, including within large-scale data collection. Nevertheless, the proposed approach may also find application in neuroimaging research.

DIFFICULTIES IN USING CURRENT TASK-BASED MEASURES TO STUDY INDIVIDUAL DIFFERENCES IN NEUROCOGNITIVE MECHANISMS UNDERLYING MENTAL HEALTH RISK

Improving our understanding of individual differences in neurocognitive mechanisms during development is essential for individual-level prediction of mental health outcomes in clinical and evidentiary applied settings (9). For example, mental health

ISSN: 2451-9022

professionals routinely develop personal treatment plans in the absence of objective markers and models to aid them in predicting mental health trajectories and clinical outcomes (10,11). Improved understanding of individual variability can also help shift views of what it means to be at risk for mental health problems—from a stable feature, or even a label, to a processing style that characterizes some young people more often than others, also depending on the context. Several collaborative efforts have been made to improve our understanding of neurocognitive development and its relation to mental health risk, such as the Adolescent Brain Cognitive Development (ABCD) Study (12) and the Healthy Brain and Child Development Study (13). These efforts have great potential to provide information about normative neurodevelopment as well as biological and environmental pathways to mental ill health (14). Nonetheless, existing large-scale datasets include task-based measures that vary in their psychometric qualities (15), often administered months or years apart. Therefore, the groundwork to develop task-based measures that are sensitive to individual differences in neurocognitive development is urgently needed and can help to improve mental health diagnosis, treatment tailoring, and outcome prediction (16).

Questionnaire measures (completed by parents, teachers, or children themselves) typically outperform tasks in predicting real-world outcomes (17,18). However, questionnaires are not designed to—and hence are not able to—discern potentially different underlying cognitive mechanisms that may lead to similar behavioral profiles but may require different interventions. For example, a questionnaire measuring conduct disorder symptoms cannot be used to discern whether the child displays aggression as an exaggerated response to perceived threat, as a result of low tolerance for frustration, or because the child does not respond to other people's expressed distress and is thus able to act aggressively to get what they want. Knowing what information-processing differences underlie mental health symptoms is relevant for locating the source of the child's difficulty and formulating personalized intervention targets (19,20). There are several reasons why questionnaires outperform tasks in individual-level prediction. Questionnaires are completed based on "priors" that stem from accumulated data relating to each questionnaire item (e.g., having "trouble relaxing") over a particular period of time (e.g., over the last 2 weeks) (21). This averaging over time, and the implied emphasis on traits that are stable in the face of diurnal, stress-induced, and other sources of variation, may partly explain the ability of questionnaires to capture individual differences reliably and sensitively (22,23). Moreover, the long tradition of careful psychometric development of questionnaire measures of mental health vulnerability has not been accompanied by comparable work on how to extract information on individual differences from task-based measures (24). This is understandable given that task-based measures have been developed precisely to minimize between-subject variability and capture aspects of cognitive function that are consistent across individuals (25). While task-based measures perform well when examining experimental (within-subject) and between-group differences and have provided critical insight into the general principles of human brain and cognitive function, they are seldom optimized to sensitively discern individual differences in continuous trait analyses (26,27). For example, titrating tasks (e.g., the stop-signal delay) is helpful to detect group-level effects (e.g., in response inhibition), but it hampers our ability to measure individual variation, which is often larger than the variance between groups (28). Task-based measures can provide information about individual differences when additional analytical steps are implemented. One approach that is frequently used is correlating task performance with variance in questionnaire measures, but these correlations tend to be modest (29–31). Furthermore, questionnaires and task-based measures of the same putative underlying cognitive mechanisms may, in fact, assess distinct constructs (26). This problem, which is widespread in cognitive research, is referred to as the jingle-jangle fallacy (30), namely, measures with the same name tapping different constructs (jingle fallacy) and measures with different names tapping the same construct (jangle fallacy). Low overlap between the same putative constructs across measurement types limits our ability to examine associations between neurocognitive mechanisms and observed behavior (32,33).

CHALLENGES AND WAYS FORWARD FOR RELIABLE TASK-BASED RESEARCH IN DEVELOPMENTAL SETTINGS

We argue that some of the main concerns with the use of current task-based measures for individual differences research represent opportunities for research into neurocognitive development.

First, a number of well-established task-based measures have been found to display suboptimal psychometric properties, including poor test-retest reliability (6,7,25) and low internal consistency (7,34). For example, suboptimal test-retest reliability has been observed in child and adolescent longitudinal functional magnetic resonance imaging studies using attentional, emotional, and cognitive control tasks, with lower reliability of blood oxygen level–dependent signal in brain regions subject to greater developmental change (15,35,36). The problem of poor psychometric properties represents an opportunity to conduct targeted research to establish when low reliability and internal consistency estimates reflect task properties versus change in the underlying cognitive mechanisms—including their state-dependent or dynamic nature. The reliability and stability of measures over time is most relevant in developmental settings because they may reflect important dynamics of neurocognitive development. For example, poor test-retest reliability of neuroimaging tasks administered during developmentally sensitive periods may result from brain activation patterns being less stable with increasing interscan intervals, when individual differences in brain development should be expected (37). Temporally sensitive methods, like the one proposed below, offer ways to disentangle different possible contributors to low reliability and may thus lead to novel insights into neurocognitive mechanisms across development.

Task-based measures also require that measurement noise is accounted for to display reasonable psychometric properties, but this is seldom the case (10,12–14). Sources of measurement error that can affect task reliability include, for example, habituation and fatigue (38). In developmental

settings, sources of measurement error also need to be distinguished from change in the neurocognitive mechanisms of interest. Indeed, change should be considered as intrinsic and central to the scientific inquiry, rather than treated as a nuisance parameter to control for, thereby removing its effect on individual and group averages. Researchers measuring the same neurocognitive mechanism across different time points should adopt approaches that can quantify variability and change, as well as intra- and interindividual processes and mechanisms that affect the rate of change and variability (e.g., language skills, distractibility, mood, motivation). When studying neurocognitive mechanisms across multiple developmental stages, it should also be noted that the same task might tap different processes at different time points (39) and that inferring the process from the measure is a challenge in itself (40). For example, as children mature, they rely increasingly on sophisticated, goal-directed, model-based decision-making strategies rather than on habitual and computationally less demanding model-free strategies (41–43). Therefore, task-based measures of decision making administered at different time points may capture different processes. Using designs that maximize the signal-to-noise ratio, such as Bayesian optimization, and statistical methods that account for different sources of variance, such as latent variable modeling and multilevel modeling, meaningful variability in task performance can be distinguished from measurement error (25,44).

In addition, the emergence of mental health symptoms reflects a multitude of genetic and environmental risk factors, each of which affects multiple neurocognitive systems and contributes a small proportion of variance in total mental health risk (45). Multivariate tools capable of capturing contributing pathways that may contain the shared variance of multiple risk factors can provide predictive leverage (46). Multivariate techniques can be used to build predictive models of mental health risk, for example, by identifying groupings of youths who are experiencing mental health problems over time based on their performance on relevant task-based measures. These data-driven groupings may align better with underlying mechanisms than traditional diagnostic categories (47,48). Neurocognitive mechanisms also do not develop in isolation; they emerge in the context of other processes, some of which may act as gatekeepers. For example, phonological awareness may act as a gatekeeper to working memory during early development. This means that the construct validity of working memory tasks could be affected by phonological awareness during early childhood, but less so at later stages (49). Multivariate approaches also allow examining whether and why variability in task performance tends to decrease across development, while mean performance improves (50). One possibility is that neurocognitive mechanisms are more variable in and of themselves during early relative to late development. Alternatively, decreased task variability over time could be due to decreased gatekeeping by neurocognitive mechanisms other than the one under study. For example, verbal ability may contribute to variability in task performance—as well as with a child's ability to comprehend task instructions—during early childhood more than during late childhood. Therefore, researchers who are interested in complex cognitive processes may need to measure and model multiple different processes to capture their codevelopment reliably and sensitively over time.

Neurocognitive development is also intrinsically interactive and situated in a particular context. Traditional experimental studies are rarely designed to account for contextual factors such as circadian rhythms, hormonal fluctuations, or changes in the social environment. Contextual factors are often controlled for, but they could instead be examined as factors contributing to variability in the neurocognitive mechanism of interest. This can be achieved by combining cognitive measurements with measures of biological or social environmental factors, particularly measures addressing social risk and protective factors that are critical in shaping development and mental health and that themselves evolve during development. For example, behavioral genetics research has indicated that the rearing environment influences individual differences in cognitive ability during early childhood, but does so to a lesser extent during adolescence, when nonshared environmental exposures (e.g., different peer groups) become relatively more influential (51). In addition, people are active cocreators of their environments, which partly explains why social risk factors are not distributed at random in the population (52,53). Consequently, studying the covariation between neurocognitive function and contextual factors in a temporally sensitive way can help to explain how individual differences in neurocognitive mechanisms relate to the generation and maintenance of social risk.

COGNITIVE MICROSCOPY: ADAPTIVE DESIGN OPTIMIZATION, TEMPORALLY SENSITIVE TASK ADMINISTRATION, AND MULTILEVEL MODELING

We propose one new approach for overcoming some of the methodological shortcomings outlined above, which we refer to as cognitive microscopy (Figure 1). This approach integrates 3 main methodologies—adaptive design optimization, temporally sensitive task administration, and multilevel modeling—to address the challenges of extracting metrics of variability and change within a multivariate framework.

Adaptive design optimization involves sampling parameters strategically to obtain a sensitive assessment of performance thresholds (54,55). This approach involves 2 main steps. A task is first developed and characterized in terms of a generative model, namely, a model that relates parameter variability to variability in task performance (56). This model is then used during administration of the adaptive optimized task version. On each task trial, a parameter estimate based on the data obtained thus far is inferred, and the following trial is chosen to maximize the amount of information gained. Because this approach maximizes the informative value of each trial, it can be especially beneficial when lengthy cognitive assessments are impractical or costly, such as in large-scale data collections or in functional magnetic resonance imaging research (54). For example, in a decision-making task, each participant would be shown choice options that are tailored to their response patterns rather than all possible options (55). Bayesian adaptive methods are the state-of-the-art for efficient adaptive design and allow an optimal tradeoff between minimal task length and efficient parameter estimation from task performance (34). Bayesian adaptive methods are particularly

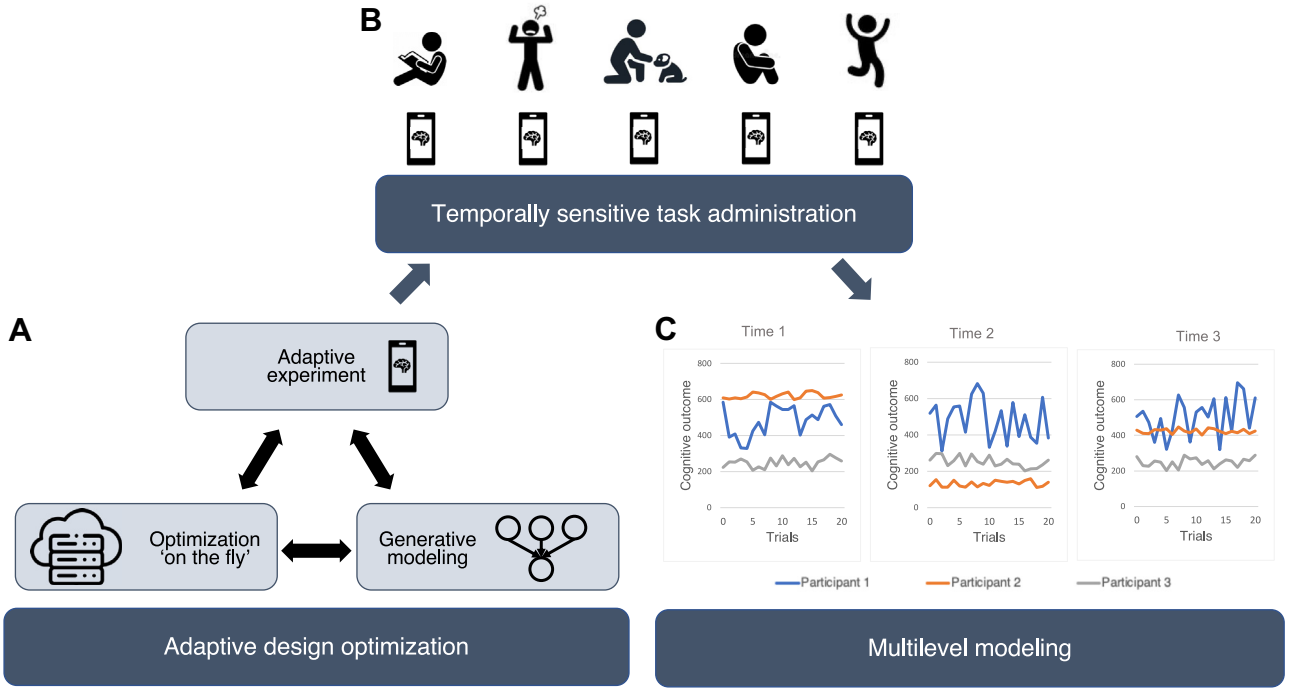

Figure 1. Visual representation of the proposed cognitive microscopy approach, consisting of 3 steps. **(A)** Adaptive design optimization: designing a task that dynamically modifies some aspects (e.g., task difficulty) based on participant performance, i.e., using Bayesian adaptive methods to estimate parameter values best describing the data as they accumulate (after each trial or a number of trials) and optimizing trials accordingly. **(B)** Temporally sensitive task administration: administering task-based assessments in a way that accounts for temporal variability in the cognitive function of interest, e.g., by sampling it at different times of the day for short intervals through mobile technology. **(C)** Multilevel modeling: using statistical techniques to analyze average performance (stable between-person variance) but also individual variability (intertrial and interassessment variance) and developmental change (interassessment variance), net of measurement error. In the example, participant 1 has stable average but high variation in performance, participant 2 has low within-session variability but high between-session variability, and participant 3 is relatively stable over time.

efficient because they minimize the number of steps required to identify the underlying cognitive model and its parameter values (57). Therefore, such methods can obviate the need to collect data from large samples or to administer long and demanding assessments, which is especially difficult in developmental settings (58).

Temporally sensitive task administration means flexible sampling that can be tailored to the neurocognitive mechanism of interest to detect temporal variability within a given time frame—from minutes to years, depending on the research question. One example of this approach is repeated short task administration, which allows capturing snapshots of cognitive function and extracting within-person metrics without introducing some of the measurement artifacts associated with traditional single-shot full-length task administration in laboratory settings (10,59,60). This approach can detect developmental change even when considering a relatively small number of repeated task administrations in a limited time window. When increasing sample size is precluded, repeated task administration is also an alternative method to increase statistical power (61). Although traditional one-occasion snapshot measurements can capture the average performance in a given setting, they do not allow the investigation of stability and variability over time. Repeated short tasks could be delivered noninvasively in naturalistic settings using portable or wearable devices (62). In this modality, repeated short tasks may also be readily integrated into large-scale data collection, thereby overcoming power limitations of small-sample studies (63,64). This approach shares features with methods such as experience sampling, ambulatory assessment, ecological momentary assessment, and intensive longitudinal data collection (65). These methods have been increasingly employed to assess affect and mood but less so for neurocognitive mechanisms (66,67), likely because of the length of traditional task-based measures and concerns about their reliability. Bayesian adaptive optimization methods have recently been used to obtain metrics of stability and variability based on brief and internally valid individual task assessments (34,68). However, only limited psychometric work has been conducted on extracting stable and variable properties of cognitive function from short tasks despite the possibility that these would offer something conceptually comparable with questionnaire ratings, which are based on a number of exemplars of a particular trait or behavior. In other words, repeated short task administration could enable extracting performance averages that represent stable information-processing characteristics comparable with the stable characteristics that are sampled by questionnaire ratings. Furthermore, variability in task performance over time has rarely been the object of study in and of itself (69), with some exceptions such as the study of reaction time variability in attention-deficit/hyperactivity disorder (70–72). This is an area

that deserves more attention because the consistency of a particular information-processing style may itself indicate greater or lesser mental health risk. Advances have also been made in delivering brief task assessments at scale (73,74). However, the studies that have been conducted to date have typically been cross-sectional, with no explicit focus on the longitudinal characterization of neurocognitive mechanisms, which we argue is particularly beneficial when studying developmental samples. Metrics of stability and variability in cognitive function may represent markers of mental health risk and also, therefore, offer important clues regarding mechanisms that should be targeted in interventions.

Multilevel models (or hierarchical linear models) can be especially fruitful when analyzing data from repeated adaptive task administrations in developmental settings. Multilevel models are a family of statistical techniques that can be used to detect sources of variability in the presence of multiple sampling dimensions, such as clusters of participants within groups (e.g., students in the same class) or assessment visits in longitudinal designs (75,76). Because they separate variability within participants, differences between them, and measurement error, these models are better suited to studying development as a continuous process than traditional statistical methods (27,77–79). In this context, multilevel modeling strategies have been used to study the development of cognitive functions (80) and brain circuits supporting them (81). Although multilevel modeling requires considerable quantitative expertise, the benefits offered by this approach may motivate researchers to develop expertise in this area. A wide range of open-source, hands-on tutorials (e.g., R Boot camp: Introduction to Multilevel Model and Interactions, offered by Penn State University) and tools for classical (82) and Bayesian (83) estimation are freely available online. Moreover, large-scale collaborative efforts and consortia have created opportunities to aggregate datasets and increase power to conduct such sophisticated statistical analyses. A case could also be made for academic institutions to provide researchers, especially early-career researchers, with the time and training infrastructure needed to acquire expertise in relevant analytical tools. Ideally, institutions would invest in growing and retaining methodological expertise in this area by offering job security and career progression pathways to researchers who focus specifically on developing and applying such analytical tools.

CONCLUSIONS

Efforts to develop more sensitive, reliable, and scalable neurocognitive measurements are required to identify developmental mechanisms of mental health problems. Our progress toward effective intervention will undoubtedly rely on our ability to build a proper mechanistic understanding of mental health conditions, for which improved cognitive phenotyping is vital. In developmental settings, this painstaking work needs to account for the stability, variability, and change in neurocognitive mechanisms that occur across development. Methodological approaches that aim to do so, such as adaptive design optimization, temporally sensitive task administration, and multilevel modeling, have the potential to improve our understanding of the neurocognitive mechanisms that underly mental health risk across developmental stages.

ACKNOWLEDGMENTS AND DISCLOSURES

SP is currently funded by a Radboud Excellence fellowship from Radboud University/UMC in Nijmegen, Netherlands. DEA is supported by Medical Research Council Programme Grant (MC-A0606-5PQ41), an opportunity award from The James S. McDonnell Foundation, by the Gnodde Goldman Sachs endowed Professorship, and acknowledges the support of the University of Cambridge National Institute for Health and Care Research Biomedical Research Centre (NIHR BRC). QJMH is supported by the Wellcome Trust (221826/Z/20/Z), Carigest S.A., Koa Health, the University College London Hospitals NHS Foundation Trust National Institute for Health and Care Research Biomedical Research Centre (UCLH NIHR BRC). EV is supported by Medical Research Council (MR/V033905/1).

QJMH has received consultancy fees and holds equity options in Aya Technologies Ltd. and Alto Neuroscience Ltd. QJMH is in receipt of a research grant from Koa Health. All other authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Division of Psychology and Language Sciences, University College London, London, United Kingdom (PP, NS, EV); Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Center, Nijmegen, the Netherlands (SP, RAK); Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom (DEA); Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom (DEA); and Applied Computational Psychiatry Laboratory, Mental Health Neuroscience Department, Division of Psychiatry and Max Planck Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology, University College London, London, United Kingdom (QJMH).

Address correspondence to Patrizia Pezzoli, Ph.D., at p.pezzoli@ucl.uc.uk, or Essi Viding, Ph.D., at e.viding@ucl.uc.uk.

Received Aug 31, 2022; revised Mar 15, 2023; accepted Mar 20, 2023.

REFERENCES

1. McGorry PD, Mei C (2021): Clinical staging for youth mental disorders: Progress in reforming diagnosis and clinical care. Annu Rev Dev Psychol 3:15–39.
2. Astle DE, Holmes J, Kievit R, Gathercole SE (2022): Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. J Child Psychol Psychiatry 63:397–417.
3. Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, *et al.* (2014): The p factor: One general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci 2: 119–137.
4. Moffitt TE, Arseneault L, Belsky D, Dickson N, Hancox RJ, Harrington HL, *et al.* (2011): A gradient of childhood self-control predicts health, wealth, and public safety. Proc Natl Acad Sci USA 108:2693–2698.
5. Simmonds DJ, Pekar JJ, Mostofsky SH (2008): Meta-analysis of go/no-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. Neuropsychologia 46:224–232.
6. Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, *et al.* (2020): What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol Sci 31:792–806.
7. Parsons S, Kruijt A-W, Fox E (2019): Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. Adv Methods Pract Psychol Sci 2:378–395.
8. Nour MM, Liu Y, Dolan RJ (2022): Functional neuroimaging in psychiatry and the case for failing better. Neuron 110:2524–2544.

9. Bennett CM, Miller MB (2010): How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci 1191:133–155.
10. Blair RJR, Mathur A, Haines N, Bajaj S (2022): Future directions for cognitive neuroscience in psychiatry: Recommendations for biomarker design based on recent test re-test reliability work. Curr Opin Behav Sci 44:101102.
11. Nord CL, Gray A, Charpentier CJ, Robinson OJ, Roiser JP (2017): Unreliability of putative fMRI biomarkers during emotional face processing. Neuroimage 156:119–127.
12. Karcher NR, Barch DM (2021): The ABCD study: Understanding the development of risk for mental and physical health outcomes. Neuropsychopharmacology 46:131–142.
13. Morris AS, Wakschlag L, Krogh-Jespersen S, Fox N, Planalp B, Perlman SB, *et al.* (2020): Principles for guiding the selection of early childhood neurodevelopmental risk and resilience measures: HEALthy Brain and Child Development Study as an exemplar. Advers Resil Sci 1:247–267.
14. Feldstein Ewing SW, Bjork JM, Luciana M (2018): Implications of the ABCD study for developmental neuroscience. Dev Cogn Neurosci 32:161–164.
15. Kennedy JT, Harms MP, Korucuoglu O, Astafiev SV, Barch DM, Thompson WK, *et al.* (2022): Reliability and stability challenges in ABCD task fMRI data. Neuroimage 252:119046.
16. Huys QJM, Maia TV, Frank MJ (2016): Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci 19:404–413.
17. Boon-Falleur M, Bouguen A, Charpentier A, Algan Y, Huillery É, Chevallier C (2022): Simple questionnaires outperform behavioral tasks to measure socio-emotional skills in students. Sci Rep 12:442.
18. Frey R, Pedroni A, Mata R, Rieskamp J, Hertwig R (2017): Risk preference shares the psychometric structure of major psychological traits. Sci Adv 3:e1701381.
19. Blair RJ, Viding E (2009): Psychopathy. In: Rutter M, Bishop DVM, Pine DS, Scott S, Stevenson J, Taylor E, Thapar A, editors. Rutter's Child and Adolescent Psychiatry, 5th ed. Oxford: Blackwell Publishing Ltd, 852–863.
20. Frick PJ, Viding E (2009): Antisocial behavior from a developmental psychopathology perspective. Dev Psychopathol 21:1111–1131.
21. Spitzer RL, Kroenke K, Williams JBW, Löwe B (2006): A brief measure for assessing generalized anxiety disorder: The GAD-7. Arch Intern Med 166:1092–1097.
22. Borsboom D, Kievit RA, Cervone D, Hood SB (2009): The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In: Valsiner J, Molenaar P, Lyra M, Chaudhary N, editors. Dynamic Process Methodology in the Social and Developmental Sciences. New York, NY: Springer, 67–97.
23. Palminteri S, Chevallier C (2018): Can we infer inter-individual differences in risk-taking from behavioral tasks? Front Psychol 9:2307.
24. Huys QJM (2018): Bayesian approaches to learning and decision-making. In: Anticevic A, Murray J, editors. Computational Psychiatry: Mathematical Modeling of Mental Illness. San Diego, CA: Elsevier, 247–271.
25. Hedge C, Powell G, Sumner P (2018): The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behav Res Methods 50:1166–1186.
26. Eisenberg IW, Bissett PG, Zeynep Enkavi A, Li J, MacKinnon DP, Marsch LA, Poldrack RA (2019): Uncovering the structure of self-regulation through data-driven ontology discovery. Nat Commun 10:2319.
27. Rouder JN, Haaf JM (2019): A psychometrics of individual differences in experimental tasks. Psychon Bull Rev 26:452–467.
28. Fisher AJ, Medaglia JD, Jeronimus BF (2018): Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci USA 115:E6106–E6115.
29. Clark IA, Maguire EA (2020): Do questionnaires reflect their purported cognitive functions? Cognition 195:104114.
30. Dang J, King KM, Inzlicht M (2020): Why are self-report and behavioral measures weakly correlated? Trends Cogn Sci 24:267–269.
31. Friedman NP, Banich MT (2019): Questionnaires and task-based measures assess different aspects of self-regulation: Both are needed. Proc Natl Acad Sci USA 116:24396–24397.
32. Nęcka E, Gruszka A, Orzechowski J, Nowak M, Wójcik N (2018): The (In)significance of executive functions for the trait of self-control: A psychometric study. Front Psychol 9:1139.
33. Cyders MA, Coskunpinar A (2011): Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity? Clin Psychol Rev 31:965–982.
34. Hajcak G, Meyer A, Kotov R (2017): Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects. J Abnorm Psychol 126:823–834.
35. Herting MM, Gautam P, Chen Z, Mezher A, Vetter NC (2018): Test-retest reliability of longitudinal task-based fMRI: Implications for developmental studies. Dev Cogn Neurosci 33:17–26.
36. Peters S, Van Duijvenvoorde ACK, Koolschijn PCMP, Crone EA (2016): Longitudinal development of frontoparietal activity during feedback learning: Contributions of age, performance, working memory and cortical thickness. Dev Cogn Neurosci 19:211–222.
37. Hartley CA, Lee FS (2015): Sensitive periods in affective development: Nonlinear maturation of fear learning. Neuropsychopharmacology 40:50–60.
38. Cooper SR, Gonthier C, Barch DM, Braver TS (2017): The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. Front Psychol 8:1482.
39. Chuderski A (2013): When are fluid intelligence and working memory isomorphic and when are they not? Intelligence 41:244–262.
40. Simpson-Kent IL, Fuhrmann D, Bathelt J, Achterberg J, Borgeest GS, Kievit RA, CALM Team (2020): Neurocognitive reorganization between crystallized intelligence, fluid intelligence and white matter microstructure in two age-heterogeneous developmental cohorts. Dev Cogn Neurosci 41:100743.
41. Smid CR, Kool W, Hauser TU, Steinbeis N (2023): Computational and behavioral markers of model-based decision making in childhood. Dev Sci 26:e13295.
42. Bolenz F, Eppinger B (2022): Valence bias in metacontrol of decision making in adolescents and young adults. Child Dev 93:e103–e116.
43. Drummond N, Niv Y (2020): Model-based decision making and model-free learning. Curr Biol 30:R860–R865.
44. Cooper SR, Jackson JJ, Barch DM, Braver TS (2019): Neuroimaging of individual differences: A latent variable modeling perspective. Neurosci Biobehav Rev 98:29–46.
45. Plomin R, Simpson MA (2013): The future of genomics for developmentalists. Dev Psychopathol 25:1263–1278.
46. Holmes J, Mareva S, Bennett MP, Black MJ, Guy J (2021): Higher-order dimensions of psychopathology in a neurodevelopmental transdiagnostic sample. J Abnorm Psychol 130:909–922.
47. Bathelt J, Holmes J, Astle DE, , Centre for Attention Learning and Memory (CALM) Team (2018): Data-driven subtyping of executive function-related behavioral problems in children. J Am Acad Child Adolesc Psychiatry 57:252–262.e4.
48. Bathelt J, Vignoles A, Astle DE (2021): Just a phase? Mapping the transition of behavioural problems from childhood to adolescence. Soc Psychiatry Psychiatr Epidemiol 56:821–836.
49. Astle D, Bassett DS, Viding E (2022): Capturing developmental dynamics within a transdiagnostic framework: Challenges and promises. PsyArXiv https://doi.org/10.31234/osf.io/jfpys.
50. Cañigueral R, Ganesan K, Smid CR, Thompson A, Dosenbach NUF, Steinbeis N (2022): Adaptiveness of fluctuations in intra-individual variability of performance is process-dependent in middle childhood. PsyArXiv https://doi.org/10.31234/osf.io/y7c5d.
51. Haworth CM, Wright MJ, Luciano M, Martin NG, de Geus EJ, van Beijsterveldt EJ, van Beijsterveldt CE, *et al.* (2010): The heritability of general cognitive ability increases linearly from childhood to young adulthood. Mol Psychiatry 15:1112–1120.

52. Beam CR, Pezzoli P, Mendle J, Burt SA, Neale MC, Boker SM, *et al.* (2022): How nonshared environmental factors come to correlate with heredity. Dev Psychopathol 34:321–333.
53. Bignardi G, Dalmaijer ES, Astle DE (2022): Testing the specificity of environmental risk factors for developmental outcomes. Child Dev 93:e282–e298.
54. Cavagnaro DR, Myung JI, Pitt MA, Kujala JV (2010): Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. Neural Comput 22:887–905.
55. Myung JI, Cavagnaro DR, Pitt MA (2013): A tutorial on adaptive design optimization. J Math Psychol 57:53–67.
56. Haines N, Kvam PD, Irving L, Smith CT, Beauchaine TP, Pitt MA, *et al.* (2020): Theoretically informed generative models can advance the psychological and brain sciences: lessons from the reliability paradox. PsyArXiv https://doi.org/10.31234/osf.io/xr7y3.
57. Ahn WY, Busemeyer JR (2016): Challenges and promises for translating computational tools into clinical practice. Curr Opin Behav Sci 11:1–7.
58. Cavagnaro DR, Tang Y, Myung JI, Pitt MA (2009): Better data with fewer participants and trials: Improving experiment efficiency with adaptive design optimization. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society 31:93–98.
59. Goodhew SC, Edwards M (2019): Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. Conscious Cogn 69:14–25.
60. Galeano Weber EMG, Dirk J, Schmiedek F (2018): Variability in the precision of children's spatial working memory. J Intell 6:1–19.
61. Goulet M-A, Cousineau D (2019): The power of replicated measures to increase statistical power. Adv Methods Pract Psychol Sci 2:199–213.
62. Wrzus C, Neubauer AB (2023): Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. Assessment 30:825–846.
63. Szucs D, Ioannidis JP (2020): Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. Neuroimage 221: 117164.
64. Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Dreber A, *et al.* (2022): Replicability, robustness, and reproducibility in psychological science. Annu Rev Psychol 73:719–748.
65. McNeish D, Mackinnon DP, Marsch LA, Poldrack RA (2021): Measurement in intensive longitudinal data. Struct Equ Modeling 28:807–822.
66. Russell MA, Gajos JM (2020): Annual research review: Ecological momentary assessment studies in child psychology and psychiatry. J Child Psychol Psychiatry 61:376–394.
67. Wenze SJ, Miller IW (2010): Use of ecological momentary assessment in mood disorders research. Clin Psychol Rev 30:794–804.
68. Ahn WY, Gu H, Shen Y, Haines N, Hahn HA, Teater JE, *et al.* (2020): Rapid, precise, and reliable measurement of delay discounting using a Bayesian learning algorithm. Sci Rep 10:12091.
69. Neubauer AB, Dirk J, Schmiedek F (2019): Momentary working memory performance is coupled with different dimensions of affect for different children: A mixture model analysis of ambulatory assessment data. Dev Psychol 55:754–766.
70. Kofler MJ, Rapport MD, Sarver DE, Raiker JS, Orban SA, Friedman LM, Kolomeyer EG (2013): Reaction time variability in ADHD: A meta-analytic review of 319 studies. Clin Psychol Rev 33:795–811.
71. Mccormick EM, Cambridge Centre for Ageing and Neuroscience, Kievit RA (2023): Poorer white matter microstructure predicts slower and more variable reaction time performance: Evidence for a neural noise hypothesis in a large lifespan cohort. J Neurosci 43:3557–3566.
72. MacDonald SWS, Nyberg L, Bäckman L (2006): Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. Trends Neurosci 29:474–480.
73. McNab F, Zeidman P, Rutledge RB, Smittenaar P, Brown HR, Adams RA, Dolan RJ (2015): Age-related changes in working memory and the ability to ignore distraction. Proc Natl Acad Sci USA 112:6515–6518.
74. Rutledge RB, Chekroud AM, Huys QJ (2019): Machine learning and big data in psychiatry: Toward clinical applications. Curr Opin Neurobiol 55:152–159.
75. Hoffman L, Walters RW (2022): Catching up on multilevel modeling. Annu Rev Psychol 73:659–689.
76. Aarts E, Verhage M, Veenvliet JV, Dolan CV, Van Der Sluis S (2014): A solution to dependency: Using multilevel analysis to accommodate nested data. Nat Neurosci 17:491–496.
77. Castro-Alvarez S, Tendeiro JN, Meijer RR, Bringmann LF (2022): Using structural equation modeling to study traits and states in intensive longitudinal data. Psychol Methods 27:17–43.
78. Parsons SD (2020): Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. Meta-Psychology 6.
79. Boyle MH, Willms JD (2001): Multilevel modelling of hierarchical data in developmental studies. J Child Psychol Psychiatry Allied Discip 42:141–162.
80. Fine KL, Grimm KJ (2019): Multilevel modeling and multilevel structural equation modeling in lifespan developmental analyses. In: Knight BG, editor. Oxford Research Encyclopedia of Psychology. Oxford: Oxford University Press.
81. Battista C, Evans TM, Ngoon TJ, Chen T, Chen L, Kochalka J, Menon V (2018): Mechanisms of interactive specialization and emergence of functional brain circuits supporting cognitive development in children. NPJ Sci Learn 3:1.
82. McCoach DB, Rifenbark GG, Newton SD, Li X, Kooken J, Yomtov D, *et al.* (2018): Does the package matter? A comparison of five common multilevel modeling software packages. J Educ Behav Stat 43: 594–627.
83. Bürkner PC (2017): brms: An R package for Bayesian multilevel models using Stan. J Stat Soft 80:1–28.