

Neuropsychological tests of the future: How do we get there from here?

Robert M. Bilder & Steven P. Reise

To cite this article: Robert M. Bilder & Steven P. Reise (2019) Neuropsychological tests of the future: How do we get there from here?, *The Clinical Neuropsychologist*, 33:2, 220-245, DOI: [10.1080/13854046.2018.1521993](https://doi.org/10.1080/13854046.2018.1521993)

To link to this article: <https://doi.org/10.1080/13854046.2018.1521993>



Published online: 13 Nov 2018.



Submit your article to this journal [↗](#)



Article views: 198



View Crossmark data [↗](#)



Citing articles: 3 View citing articles [↗](#)



Neuropsychological tests of the future: How do we get there from here?

Robert M. Bilder^{a,b} and Steven P. Reise^b

^aDepartments of Psychiatry & Biobehavioral Science, Jane & Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, California, USA; ^bDepartment of Psychiatry & Biobehavioral Science, Los Angeles, California, USA

ABSTRACT

Objective: This article reviews current approaches to neuropsychological assessment, identifies opportunities for development of new methods using modern psychometric theory and advances in technology, and suggests a transition path that promotes application of novel methods without sacrificing validity.

Methods: Theoretical/state-of-the-art review.

Conclusions: Clinical neuropsychological assessment today does not reflect advances in neuroscience, modern psychometrics, or technology. Major opportunities for improving practice include both psychometric and technological strategies. Modern psychometric approaches including item response theory (IRT) enable linking procedures that can place different measures on common scales; adaptive testing algorithms that can dramatically increase efficiency of assessment; examination of differential item functioning (DIF) to detect measures that behave differently in different groups; and person fit statistics to detect aberrant patterns of responding of high value for performance validity testing. Opportunities to introduce novel technologies include computerized adaptive testing, Web-based assessment, healthcare- and bio-informatics strategies, mobile platforms, wearables, and the ‘internet-of-things’. To overcome inertia in current practices, new methods must satisfy requirements for back-compatibility with legacy instrumentation, enabling us to leverage the wealth of validity data already accrued for classic procedures. A path to achieve these goals involves creation of a global network to aggregate item-level data into a shared repository that will enable modern psychometric analyses to refine existing methods, and serve as a platform to evolve novel assessment strategies, which over time can revolutionize neuropsychological assessment practices world-wide.

ARTICLE HISTORY

Received 13 April 2018



Accepted 5 September 2018

KEYWORDS

Neuropsychology; psychometrics; psychological tests; clinical decision-making; diagnostic techniques and procedures; information science

Introduction

The most widely used clinical neuropsychological (NP) methods are based on procedures developed 50 to 150 years ago, and transformational advances in cognitive

CONTACT Robert M. Bilder  rbilder@mednet.ucla.edu  Departments of Psychiatry & Biobehavioral Science, Jane & Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Room C8-849, 740 Westwood Plaza, Los Angeles, CA 90024, USA

© 2019 Informa UK Limited, trading as Taylor & Francis Group

neuroscience, psychometrics, and technology have not yet been translated effectively to the clinic for the benefit of the public (Collins & Riley, 2016). This poses an enormous problem because the current methods are cumbersome and costly, limiting access to accurate diagnosis and treatment. The problem is particularly acute given the effects of aging on our population and increasing pressures to reduce health care spending. Further, most NP testing platforms do not support integration with either research databases or electronic health records that are emerging as critical components of modern health care decision-making. A revolution in assessment is necessary to enhance brain health globally, and this goal is achievable with current knowledge and technology, but it will demand a major shift in current practices. This article summarizes the current state of NP assessment methods, points to directions for future development, and outlines a path to accelerate the transformation of clinical neuropsychology.

Historical roots persist in current NP assessment methods

Clinical NP assessment today relies primarily on procedures with origins in the late nineteenth to mid-twentieth century that do not reflect major advances in understanding of brain-behavior relations, psychometric theory, and technology. A recent survey of NP assessment practices in the United States (Rabin, Paolillo, & Barr, 2016) revealed that the most commonly used tests were those from the 'Wechsler' family [for adults, the Wechsler Adult Intelligence Scale (WAIS); and the Wechsler Memory Scale (WMS); and for children, the Wechsler Intelligence Scale for Children]. Rounding out the top five tests were the Trail Making Test(s) and the California Verbal Learning Test. The most frequently used tests remained identical to those prioritized by neuropsychologists a decade earlier (Rabin, Barr, & Burton, 2005) and these mostly have roots in the nineteenth century and were developed into their currently used forms before the end of World War II (Bilder et al., 2009a).

Fundamental questions about the goals of testing have remained constant over more than a century. For example, there is a long-standing tension between using more 'specific' tests of basic perceptual and motor abilities (as measured by reaction time and related procedures developed in the nineteenth century laboratories of Wilhelm Wundt in Germany and Sir Francis Galton in England), and more 'general', complex and broad-based procedures favored by the early twentieth century IQ test developers because they found these correlated more strongly with real-world functioning. Clark Wissler is credited with a 1901 publication showing that the basic 'Galtonian' tests of RT, sensory, and perceptual abilities did a poor job of predicting academic performance in Columbia and Barnard students (Gregory, 2016). The Galtonian measures were subsequently dropped from the development path that led to the most popular measures today. This tension is being echoed more than a century later, by findings that more 'basic' cognitive neuroscience procedures (such as those developed to tap the functions of discrete neural circuits and cross-validated with respect to functional MRI procedures) were less correlated with everyday functioning measures, relative to more complex NP tests (Gold et al., 2012). A critical goal for neuropsychology is to determine how test development should proceed,

based on the desired purposes of assessment, towards physiological validity or ecological validity. The current generation of NP assessment tools focuses almost exclusively on more complex measures that may be more valid with respect to real-world outcomes but fail to reflect current knowledge and theories of cognitive neuroscience. Ironically, a lack of ecological validity is perceived as one of the greatest challenges faced by clinical neuropsychologists in selecting NP tests (Rabin et al., 2016). Perhaps, current NP tests could be better targeted to ecological targets by embedding NP-sensitive technologies into the built environment (see in the section 'Leveraging technology'). On the other pole, while certain 'experimental' and basic sensory-motor procedures continue to enjoy popularity (e.g. RT measures on continuous performance tests; hand grip strength testing), technology development may soon displace the role of behavioral methods by acquiring neurophysiological information directly (e.g. via wearable neuroimaging devices using EEG, fNIRS, MEG, or PET sensors; see in the section 'Leveraging technology'). Clinical neuropsychology is at a critical juncture and must determine what complement of methods will enable optimal assessments in the future.

Leveraging psychometrics

Clinical neuropsychology has a rich legacy of development from the neurological tradition, following investigators such as Luria, Teuber, and others who developed procedures largely based on a 'sign approach' to understand the impact of focal lesions to the brain. Many of these procedures helped advance understanding of brain behavior relations by defining thresholds for clinically meaningful impairment but did not consider psychometric methods to compute the significance of deviations from normative standards. Complementing this legacy are many of the most widely used NP tests that honor measurement principles considered state-of-the-art *at the time they were developed*. This tradition, however, is primarily based on *classical* rather than *modern* psychometrics (Novick, 1966). The development of tests with improved classical psychometric properties was identified as a critical step in the progress of neuropsychology as an established discipline referred to as 'Neuropsychology 2.0' (Bilder, 2011).

But psychometric theory has made major advances over the last 50 years that have not been incorporated into NP practice. The development of modern psychometric theory, particularly item response theory (IRT), has provided noteworthy advantages for clinical assessment (Reise & Waller, 2009). The fundamental assumption underlying IRT is that a *latent trait* gives rise to a given individual's responses to each item on a test, and the relations between the trait and the responses are usually modeled by a series of parameters, although non-parametric 'empirical' IRT models also have been developed.

Most IRT analyses have focused on *unidimensional* models where a single latent trait is assumed to underlie the responses. Developments of IRT, however, enable specification of *bifactor* and *multidimensional* IRT models. In the bifactor model, items are seen to load on both a general dimension and individual factors. This has clear relevance to the assessment of human NP function, given that in clinical practice we frequently aim to determine 'global' deficits as a top priority, and then consider

the integrity of more specific NP functions. In multidimensional IRT (mIRT) models, each item can provide information about multiple different traits. This may further increase the efficiency of NP assessment, given that there are often high correlations among test measures from diverse functional domains.

Following Reise and Waller (2009), we consider several key advantages of IRT over classical test theory (CTT) for clinical NP assessment including: (a) leveraging information from the nominal response model (when there are multiple responses within a given item); (b) test linking; (c) computerized adaptive testing (CAT); (d) differential item functioning; and (e) application of person-fit statistics. We also consider approaches to leveraging item-level data even when IRT may not be appropriate. Table 1 provides an overview of potential methodological advances including these features.

Nominal response model

For test items that have *polytomous* response alternatives (i.e. when there are more than two response options), the *different responses may convey unique information*. For example, in tests that use a multiple-choice format,¹ each of the different erroneous responses carries distinctive information about the test-taker's ability (Preston & Reise, 2014, 2015; Preston, Reise, Cai, & Hays, 2011). Usually, some answers are 'more wrong' than others, so selecting a poorer response option may signal a lower ability level than selecting an 'almost correct' response. This confers an additional advantage over many conventional tests where credit is usually given on a *dichotomous* or an all-or-none basis.² Classic examples of this approach showed that precision of measurement could be enhanced for the Raven Standard Progressive Matrices Test (Thissen, 1976) and Vocabulary, especially for individuals with lower levels of ability (Bock, 1972). It is surprising, given this work done almost 50 years ago, that the additional information from wrong answers is seldom used to increase measurement precision or reduce the total number of items in modern NP tests. Leveraging the nominal response model may be particularly valuable now that online testing is becoming more accessible, given that multiple-choice formats are easier to administer over the internet relative to free verbal responses.

Test linking

Linking enables items from different measures to be placed onto a common scale. For example, McHorney and Cohen (2000) used linking effectively to generate a well-calibrated item pool spanning 75 different self-report instruments about 'functional status'. Similar methods could readily be applied to selected constructs in neuropsychology. For example, to the extent that certain indicators (such as long delay free recall) from different auditory verbal list-learning procedures (e.g. Rey Auditory Verbal Learning Test, California Verbal Learning Test) are assumed to tap the same latent trait, we could use IRT-based linking methods to identify an individual's ability level using either test, with confidence that the identified ability levels are comparable. These methods were used to 'co-calibrate' the Mini-Mental State Exam (MMSE), 3MS,



Table 1. Overview of potential methodological advances in neuropsychological (NP) assessment.

Method	Current	Future	Advantage
NP trait models	Unidimensional	Bifactor models, multidimensional IRT models (mIRT)	Each item can provide information about different traits; a single item or test can help specify both general factors and domain scores
Nominal response model	Different kinds of errors are treated identically	Each wrong response has a different meaning	Each item carries more information, enabling greater precision and/or assessing different constructs
Test linking	Total scores are compared in studies that use both tests	Item banks can be drawn from existing tests and new items, and all items calibrated together	Enables direct comparison of different tests and construction of new tests that are back-compatible with the originals
Computerized adaptive testing	Paper-pencil, fixed administration order, minimal branching	Information from each item response selection and speed used to select next most informative item	Efficiency gain of 50–95% in administration time or precision of measurement
Differential item functioning (DIF)	Effects of group (diagnostic, age, sexual, racial, ethnic, cultural, etc.) determined by comparing total scores	DIF examines group effects for each item	Increased precision in specifying diagnostic and other group differences that may not be apparent in the scores of the whole test
Person fit statistics	Performance validity based on 'cutoff' scores, mostly based on accuracy	Performance validity based on the fit of item response characteristics to the examinee's overall estimated trait level	Performance validity can be examined within each test; every item response can be useful in detecting anomalies; increase sensitivity to intentional failure
Non-IRT item-level strategies	Most emphasis on summary scores not trial-by-trial analysis	Focus on sequential dependence of responses and meaning of response sequences	Increased efficiency in identifying primary constructs; identification of qualitatively distinct response patterns
Evidence-based diagnostic batteries	Batteries with limited flexibility involve redundant testing	Test selection will proceed based on positive predictive power	Testing efficiently focuses time with respect to differential diagnostic questions or recommendations
Computerized testing	Print publishing model; paper-pencil data acquisition and scoring	Computerized tests for stimulus presentation and response acquisition	Precision in timing of stimulus presentation and response collection, automatic recording, scoring and database entry of responses; and automatic updating of software to new versions; acquisition of voice, video, motion
Web-based testing	Testing done in clinic or lab	Testing done at home or wherever convenient for examinee	Scalable assessment at lower cost
Healthcare informatics and bioinformatics	Test results go to file cabinets, report text goes on medical record	Data elements will be part of medical record and integrated with analytics relating these to other health variables	The NP data will be integrated into comprehensive model of patient; implications will be pushed to all care-team members and hypotheses fed back to NP clinicians for follow-up; 'big data' analytics will find new patterns to inform future evidence-based practice
Mobile platforms	Not used; not trusted	Passive monitoring will dramatically increase data flow; experience sampling will augment self reports	Marked increase in longitudinal repeated measures for self-reports and tests; new variables extracted from passive monitoring
Wearables	Not used; not trusted	Passive monitoring of diverse physiological, activity, and experiential data	Data previously available only in cross-sectional lab studies (sleep, EEG, cardiovascular) will be widely available and assessed longitudinally
Internet of Things (IOT)	Not used; not trusted	Passive monitoring of activities across multiple environments	Ecologically valid assessments will be done in real-world contexts; and environment can 'respond' with appropriate cues and assistance

Cognitive Abilities Screening Instrument, and the Community Screening instrument for Dementia, enabling direct comparison across studies and revealing that markedly different cut scores were applied across studies (Crane et al., 2008). Using test linking methods, the authors further showed that the MMSE is less accurate than the other screening tests at higher levels of ability, and they provided a simple look-up table enabling examiners to determine which scores are comparable across instruments (for example, an MMSE score of 22 is comparable to 3MS scores in the range of 54 to 60, further illustrating the increased sensitivity of 3MS at this level of ability). The IRT methods additionally enabled substantial improvements in change measurements, by considering the fact that an examinee moving from one score to another is necessarily moving from one level of measurement precision to another, a fact that is ignored in classical measurement approaches.

Linking methods further provide an opportunity to develop new tests (with new content), while maintaining back-compatibility with prior versions of the test. For example, the Educational Testing Service uses IRT to innovate and inject updated content in their development of high-stakes aptitude and achievement tests for which comparability over time and diverse administration settings is essential (Carlson & von Davier, 2013).³ This could help overcome the challenges that test developers have long faced when attempting to introduce novel NP instrumentation. There are more elaborate methods available for *data alignment* across tests that may enable integration of data even when there are no linking items (Asparouhov & Muthén, 2014; Marsh et al., 2018; Muthén & Asparouhov, 2014; Van De Schoot et al., 2013), and these methods may usefully augment the IRT-based linking methods.

Computerized adaptive testing (CAT)

Computerized adaptive testing offers the opportunity to dramatically enhance the efficiency of certain NP testing procedures. Most current NP tests contain a set of items that are to be administered in a prescribed order, often progressing from easier to more difficult items. Because the tests need to span a broad range of ability levels, an individual of a given ability level will be administered some items that are too easy for them to be informative, and some items that are too difficult for them to be informative. Some tests address this challenge by establishing a 'basal' level, starting at an intermediate-to-low level of difficulty, then continuing to more difficult items if responses are accurate, and 'backtracking' to simpler items if there are errors. This process can be cumbersome and error prone for human examiners, and even if executed successfully, it is not the most efficient method. Instead, CAT models enable selection of *the most informative item* on a given trial, based on prior estimates of the individual's ability (Waller & Reise, 1989). On an individual's first trial with a given test (assuming we know nothing else about the individual), the CAT model will select the item of median difficulty, assuming the person has 'average' ability. A correct response will update our estimate of that individual's ability, and lead to the selection of the *next most informative item*, which may not be the next more difficult item in the series, but instead is a more difficult item that might be passed by 50% of individuals who have ability at the 75th percentile, and if that is correct, the next item might be that passed

by 50% of people whose true score is at the 87th percentile, but if that is wrong, then the next item might focus on ability near the 81st percentile, and if that is correct, we will have arrived at an estimate of ability between the 81st and 87th percentiles after administering only four items.⁴ In general, when equipped with a bank of well-calibrated items, it is possible to generate an adaptive test that will use the most efficient method (the smallest number of items) to arrive at an estimate of the latent trait with a pre-specified level of precision. In contrast to standard tests where the number of items is fixed and precision varies, in a CAT the number of items administered varies depending on the preferred level of precision. Compared to typical fixed-length tests, the application of IRT-based CAT procedures typically leads to efficiency gains of 50–95% without a decrease in the quality of measurement (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Ebesutani et al., 2012; Gibbons, Clark, VonAmmon Cavanaugh, & Davis, 1985; Gibbons et al., 2008; Moore et al., 2015, 2016; Pilkonis et al., 2011; Waller & Reise, 1989).

The CATs can efficiently examine multidimensional or bifactor IRT models. For example, Gibbons et al., (2008) examined performance of a CAT that aimed to increase efficiency in the administration of the Mood and Anxiety Spectrum Scales, comprising 616 items with an average administration time of 115 minutes. To specify the general factor, only 24 to 30 items were needed (95% reduction in items administered), and the administration time was reduced to 22 min. To specify not only the general factor but also the four additional subscales, a total of only 98 to 118 items was needed – still an 85% reduction in items relative to the full scale. If comparable gains in efficiency could be achieved for clinical NP assessment, the typical 4- to 8-h assessment might be reduced to 1 or 2 h, without significant sacrifices in clinically useful information.

Differential item functioning (DIF)

Differential item functioning refers to an item that may behave differently in different groups that have the same true ability. DIF may be critically important in evaluating systematic differences in test performance among groups with distinct cultural or linguistic backgrounds. McHorney and Fleishman (2006) reviewed findings of DIF according to age, gender, race, ethnicity, socioeconomic status, language, nationality, and health care setting. It is important to note that findings of DIF do not necessarily mean that overall test scores are invalid (that is, there may be DIF without 'DTF' or differential test functioning, because individual items may cancel each other out or not impact total scores). But findings of DIF should prompt caution in the application of tests across different groups, and lead researchers to seek meaningful psychological explanations for these differences (McHorney & Fleishman, 2006; Reise & Waller, 2009). Given the importance of accurate NP assessment across multiple groups defined by linguistic, cultural, racial and ethnic backgrounds, the assessment of DIF holds great promise for more systematic identification of factors that may spuriously lead to false conclusions about differences in brain function and the diagnosis of neuropsychiatric disorders internationally. Dan Mungas and colleagues used IRT-based DIF analyses to develop the Spanish and English Neuropsychological Assessment Scales, which have

matched English and Spanish forms, and have demonstrated distinctive patterns of association among NP deficits and measures of brain structure attributable to different ethnoracial groups (Gavett et al., 2018; Meyer et al., 2018; Mungas, Reed, Crane, Haan, & González, 2004; Mungas, Reed, Haan, & González, 2005; Mungas, Reed, Marshall, & González, 2000; Mungas, Widaman, Reed, & Farias, 2011).

DIF offers one of multiple approaches to addressing general questions about *measurement invariance* (i.e. do two instruments behave the same way in different groups). There is a large literature using confirmatory factor analysis (CFA) to examine differences in means and covariance structures across groups, and there are multiple IRT-based approaches [e.g. likelihood ratio approaches and differential functioning of items and tests (DFIT) analyses] that may yield distinct results (Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002). We believe these methods are critically important for the development of instruments that can be used internationally, across linguistic and cultural boundaries, and offer enormous potential for development of globally integrated knowledge about brain-behavior relations.

Person fit statistics

Person fit statistics have been developed to identify aberrant patterns of item responses that do not fit with the overall trait level identified for a given individual or the overall patterns observed in other test takers. A good introduction to this topic is provided by Meijer, who identified multiple types of aberrant responses, including: 'sleeping behavior' (slow starters who miss and never check their answers to easy items but 'wake up' and do better on harder items); guessing behavior; cheating behavior (with more correct answers to difficult items that were copied from a more proficient neighbor); plodding behavior (where scores align with the ability level too perfectly, suggesting that the person is taking special care on each item before proceeding to the next item); alignment errors; extremely creative examinees (who fail easy items because they presume the correct answers must be more complicated); and deficiency of sub-abilities (if there is a special ability confounded with overall item difficulty) (Meijer, 1996). These methods already have been proposed for application in the detection of both faking good (impression management) and faking bad (malingering), along with simply unmotivated or unsympathetic test responding (sabotaging) (Reise & Flannery, 1996). There have been more recent developments in these methods to help identify invalid response patterns, or those in which the response pattern does not fit other data produced by the same person (Kim, Reise, & Bentler, 2018; Mansolf & Reise, 2018; Reise, Kim, Mansolf, & Widaman, 2016), and these may be augmented by inclusion of response times, in ways that have not yet been widely used in NP testing. For example, a participant's slow selection of a particularly unlikely response option may fit poorly with other responses by that participant (e.g. rapid, accurate responses to harder items). We believe these analyses can lead to multiple new proposals for data-driven embedded performance validity tests (PVTs), so that ultimately, every NP test will have built-in indicators of validity, both based on within-test patterns of performance, and relations of responses to responses from the same person during other tests.

Non-IRT item-level strategies

Item Response Theory has been applied most often to tests that include similarly styled items selected from a domain, like questionnaires and multiple-choice achievement tests. Many item-based NP tests are suitable targets for IRT, but others are not. For example, one may wonder how IRT can benefit tests that vary content across items in a graded sequence (as in Digit Span) or use multiple sequentially dependent trials (as in list-learning paradigms). Modifying NP tests may require more fundamental changes in the paradigms, beyond item selection based on IRT-derived difficulty and discrimination parameters. For example, an interesting modification of the Digit Span procedure is to change difficulty dynamically depending on the accuracy of prior responses. A computerized, adaptive digit span procedure yielded reduced variance, higher test-retest reliability, and stronger concurrent validity with other measures, relative to the conventional Digit Span procedure (Woods et al., 2011). Similar adaptive titration methods using up-down transform rules have been used to measure working memory capacity (Lencz et al., 2003). Alternative novel algorithms may be used to identify psychometrically robust short forms of existing procedures (e.g. a Poisson predictive model enabled reduction of the 60-item Raven Standard Progressive Matrices test to only nine items) (Bilker et al., 2012).

Increasing the efficiency of procedures that rely on sequentially dependent trials (e.g. list-learning) may be more challenging, and shortening these tests too much may reduce their sensitivity, reliability, and/or validity (Loring et al., 2018). But sequential analysis of items may have a profound impact. For example, there was once debate in schizophrenia research about whether it would be psychometrically defensible to use the 64-card short form, rather than the full 128-card version of the Wisconsin Card Sorting Test (WCST). Prentice, Gold, and Buchanan (2008) examined WCST responses trial-by-trial, and found that patients were significantly differentiated from healthy people within the first four cards; moreover, this analysis provided insights about a differential deficit among patients specifically in processing of negative feedback. If the goal is to identify problems in strategy-switching based on learning from negative feedback, it is not clear that administering more than four cards is either necessary or desirable, given that continued failure often frustrates and sometimes alienates the unsuccessful examinee. In another example, the intra-individual variability across consecutive trials on the Rey Auditory Verbal Learning Test was found to increase sensitivity relative to customary total score indicators (Sugarman et al., 2014). Analyses like these, even if not directly based on IRT, are enabled by the availability of trial-by-trial, item-level data. Unfortunately, few studies have carefully evaluated potentially relevant variables on a trial-by-trial basis, largely due to the absence of databases containing that level of response detail. In the past, it was also not feasible to administer tests that involved analyses of trial-by-trial data ‘on the fly’, but now that computers are ubiquitous, this potential is within reach.

Leveraging technology

The current practice of NP assessment relies principally on a print publishing model that is experiencing rapid change and is the target of innovation globally. We

purchase test materials in printed, boxed, and ground-shipped packages. We prepare packages of paper test forms, record responses using pencils on paper, flip pages of stimulus books, click stopwatches, and following hours of data gathering, manually sum up various scores, and then look up norms in printed manuals and record the results on yet another piece of paper or enter scores (again) on a computer that will perform additional calculations, which we then print, often inputting these data (one more time) into a score summary sheet. At last, we are prepared to interpret the data! This practice has remained essentially unchanged for decades. Many opportunities now exist to update our practices, including computerized testing; Web-based testing; health care informatics and bioinformatics; mobile platforms; wearables; and the internet of things (IOT).

Computerized testing

Despite the widespread availability of personal computers since the 1980s, and manifold advantages including precision in timing of stimulus presentation and response collection, automatic recording, scoring and database entry of responses, and automatic updating of software to new versions, computer administration is usually limited to a few tests and even these tests are still administered in essentially the same way they were in the 1970s. New methods are enabling acquisition of voice, facial expression, drawing and other three-dimensional actions, and other complex behaviors. Pearson's Q-interactive platform, ImPACT concussion testing, the NIH Toolbox, and the Penn Computerized Neurocognitive Battery (CNB) have been developed using either iPads or computers for certain aspects of administration and scoring, but so far, the penetration of these products in clinical NP assessment worldwide has been limited. There may be greater uptake of computerized products in physicians' offices, where non-psychologists are billing for computer testing services, mostly doing screening. It may ultimately be the task of clinical neuropsychologists to provide evidence that their assessments provide incremental value over these brief screening instruments.

Web-based testing

Despite sharing many of the advantages of testing on a local computer and additional advantages of scalability, Internet-based assessment remains infrequently used. The Penn CNB offers one alternative, as does Cambridge Cognition's CANTAB Connect. Online retailers of 'brain training' now also offer cognitive assessments (for example, see Lumosity.com, where the NeuroCognitive Performance Test or NCPT is offered to 'premium' subscribers), and there are numerous additional for-profit websites offering cognitive assessments of highly variable quality. There are also noteworthy not-for-profit developments, including TestMyBrain (www.testmybrain.org), which had included more than 1,700,000 participants as of 2017 (accessed 4 April 2018). None of these developments appears to have had substantial impact on the clinical NP assessment landscape. The promise of these assessments is that individuals might be able to monitor their own brain function, in the same way that we are accustomed to currently monitoring finances, diet, or exercise.

The development of new computerized and web-based testing raises many questions, including: What is the role of the examiner? We are accustomed to examiners providing qualitative observations about performance and supportive or corrective feedback to examinees. The current generation of computer and online tests are generally lacking these ‘uniquely human’ capacities. But we anticipate the rapid development of automated methods that include video, audio, and haptic capture, and these tools may make many qualitative and quantitative observations as well as humans do. Feedback too is improving rapidly, and humans are already ‘fooled’ by some apps that respond in surprisingly human ways. The human components of decision-making in our exams and interpretation are expected to remain robust for the near future, but there is good reason to expect that key elements of decision-making, including test selection and differential diagnosis, will increasingly be augmented by algorithms that have a comprehensive command of medical and neuropsychological knowledge. This should free us to focus on the uniquely human aspects of interpersonal communication, complex judgment, and creative planning that may take longer for computers to emulate.

Healthcare informatics and bioinformatics

The last two decades have witnessed an explosion of ‘big data’ repositories, the advent of mandatory electronic health records, and the creation of myriad new methods for data mining and ‘deep’ machine learning algorithms. There has been little uptake of these methods despite efforts to create bridges from NP constructs to other domains of biological science (Bilder, 2011; Jagaroo, 2009; Poldrack et al., 2011; Sabb et al., 2008). The electronic health records mandated by the Affordable Care Act may offer the clearest long-term path to aggregate NP data on a near-universal scale. If all NP data collected at major health systems were aggregated, this would enable enormous opportunities for NP data to be mined and analyzed with respect to all recorded medical illness categories, treatments received, and the results of other diagnostic tests including blood tests, genetic tests, and diverse imaging procedures. Meanwhile, the NIMH Data Archive (NDA) is already accumulating data on cognitive tests from research projects spanning autism, clinical trials, the Research Domains Criteria (RDoC) initiative, the Adolescent Brain and Cognitive Development (ABCD) study, and the Human Connectome Project (see <https://data-archive.nimh.nih.gov/>). The RDoC database already contains about 10,000 clinical records, and includes templates for item-level data recording on many of the most popular NP instruments. The NIH also launched the *All of Us* research program (see <https://allofus.nih.gov/>), which aims to engage 1,000,000 individuals across the United States and gather at least genomic and self-report data. It is anticipated that additional data will be accumulated from personal devices and sensors, and that protocol extensions will ultimately include cognitive characterization.

Mobile platforms

In the Spring of 2017, it was estimated that there are more than 237,000,000 cell phone users in the United States⁵ and by 2011, the number of wireless subscriptions

(>336,000,000) already exceeded the population of the United States, with 'wireless penetration' at 104.6%.⁶ These mobile devices are seldom more than a few feet away from their owners, who 'check' their devices on average 46 times per day,⁷ and passive monitoring of behavior on these devices is already ubiquitous and providing troves of data about our usage and physical locations, that are aggregated by cellular service carriers, internet service providers, Google, Apple, Amazon, Facebook, and other entities. Most of these data are being used for commercial purposes, but a growing number of research studies are dedicated to implementing both health-related assessments and interventions via mobile platforms. As part of the NIH 'Big Data to Knowledge' (BD2K) initiative, one Center of Excellence was established for Mobile Sensor Data-to-Knowledge (MD2K; see <https://mhealth.md2k.org/>). Not only are self-report ratings of cognitive function and specific cognitive tests now deployed on mobile devices, but *passive* monitoring of our mobile phone use may provide insights into our cognitive ability. For example, machine learning algorithms have extracted features from passive monitoring of mobile phone use, validated these features with respect to conventional psychometric assessments of vigilance, and found they can detect deviations in alertness as small as 11% (Abdullah et al., 2016). This technology is developing rapidly and could revolutionize NP assessment, particularly the ability to follow-up continuously with patients between more extensive exams. The patterns of mobile phone use could signal health care providers about changes in cognitive function, sleep patterns, mood, mobility, exploration of novel environments, social engagement, and other features that may provide critical indications of clinically meaningful change.

Wearables

There are now many wearable accessories that provide information similar to that recorded on our mobile phones (such as accelerometry, GPS, light-dark, sound/noise monitoring, speech detection, detection of social interactions), along with additional sensing of temperature, heart rate and other electrocardiographic (ECG) data, galvanic skin response, blood oxygenation, electromyographic (EMG) responses, facial expression, posture, sleep patterns, and respiratory rhythms. Further developments emerge daily. Consider the recent production of 'tattoo electrodes' that are designed to serve as ECG or EMG electrodes, but can be ink-jet printed, are less than one micron thick, and can be used for both recording and stimulation (Ferrari et al., 2018). While there so far remain substantial limitations in the nature, quality, and scalability of deployment, the future of 'portable' and wearable EEG, functional near infrared spectroscopy (fNIRS), magnetoencephalography (MEG), or even positron emission tomography (PET) devices seems within reach (Boto et al., 2018; Hocke, Duszynski, Debert, Dleikan, & Dunn, 2018; Melroy et al., 2017) and standards for large-scale data aggregation already have been proposed (Bigdely-Shamlo, Makeig, & Robbins, 2016). We should anticipate that in addition to typical NP assessment data, future knowledge-bases will contain a diversity of other continuously and passively-acquired brain-relevant signal data that may be useful for both prevention of disease and cognitive enhancement.

Internet of things (IOT)

While mobile and wearable technologies already demonstrate promise for augmenting NP assessment, there may be even more rapid growth in the ‘internet-of-things’, which includes remote sensing and transmission devices embedded in our environments (homes, workplaces, cars, city streets, public transportation, etc.). Maureen Schmitter-Edgecombe and her colleagues have been leaders in developing ‘smart home’ technologies that automatically and passively gather and identify a range of behavioral features potentially useful in longitudinal monitoring of cognitively and socially relevant functioning (Aramendi et al., 2018; Dawadi, Cook, & Schmitter-Edgecombe, 2013, 2016). While this work currently benefits from specialized sensors that are custom-installed for this purpose, the ever-expanding deployment of other sensors (e.g. Google Home, Amazon Echo with Alexa, Apple’s HomePod and HomeKit, myriad other home and public security systems, etc.) creates an enormous opportunity to develop large-scale projects that can advance NP assessment and intervention. Although this idea may soon become obsolete due to self-driving vehicles, there has long been an effort to use sensors in automobiles to track driver cognitive, sensory, and motor skills (Pompei, Sharon, Buckley, & Kemp, 2002), and there is a recent US Patent for a ‘System for Fatigue Detection Using a Suite of Physiological Measurement Devices’ (Grube, Thomas, Craig, & Gast, 2017), which seems eminently more reasonable than attempting to draw conclusions about peoples’ driving abilities based on NP tests given in the clinic or laboratory using methods that are far less ecologically valid (e.g. the Trail Making Test).

Obstacles

Given the advances enumerated above, what is preventing our discipline from making rapid progress in the development of novel NP assessment and intervention strategies? Sometimes test publishers are identified as the culprits, but test publishers often indicate they are eager to migrate to new software products (as has much of the rest of print publishing), and even with print products, test vendors tend to see advantages when they can sell new tests. It is mostly we, the customers, who continue to ask for ‘legacy’ products and resist new technology developments. Why?

Legacy (validity, familiarity)

Why would we, the customers, want to stick with the familiar? Other than this being a natural human tendency, the largest single factor may be our ethical responsibility to use the strongest, evidence-based methods available to evaluate our patients. This relatively unassailable assertion carries with it a corollary and mandatory inertia because our evidence base comprises *static publications* (either peer-reviewed research publications or publications of authoritative test manuals that summarize research studies conducted by the test publisher), which are fixed at the time of publication and only updated either by new publications or major test revisions (Bilder, 2011). To use a new instrument, a responsible neuropsychologist would want evidence that the new method is better than the old method. But to create a new method using the

current publication model, the new test designer needs to develop the new method and then demonstrate back-compatibility using a combination of convergent, divergent, and predictive validity studies, along with large-scale, demographically appropriate collection of normative data. This leads to the next challenge.

Money and time (and time = money)

Multiple scholarly organizations have developed working groups to consider how to advance our instrumentation and develop new tests. These working groups generate a wealth of outstanding ideas for progress, but to change the practice of clinical neuropsychology will demand a large-scale, sustained, coordinated effort, requiring substantial investments over 5 to 10 years. The investment is needed for both test development (which in turn requires not only considerable experience and knowledge, but considerable time and resources) and the subsequent normative and validity studies. Even our largest scholarly organizations (e.g. the American Psychological Association's Society for Clinical Neuropsychology, the National Academy of Neuropsychology, the American Academy of Clinical Neuropsychology, and the International Neuropsychological Society) so far do not have the budgets to support major test development efforts. Investment is necessary; ideally this investment would combine support from agencies that support data archives and have vested interests in population health (e.g. NIH, SAMHSA, the World Health Organization, other international foundations, along with state and regional agencies), together with companies that have proprietary rights to legacy test content and large-scale test development experience.

Ethical challenges

Many new technologies appear to be simultaneously 'cool' and 'creepy,' but as we gain experience and refine uses over time, both the extraordinary novelty and sense of alienation and fear tend to decrease. For example, just a few decades ago brain-machine interfaces were largely relegated to science fiction, but today, we routinely evaluate patients for deep brain stimulator implantation, and recent experiments claim to augment episodic memory encoding and recall in humans via direct patterned stimulation of hippocampal subfields using a prosthetic device (Hampson et al., 2018). Resolving these challenges may benefit from distinguishing what we *can* do from what we *should* do. For example, should genetic engineering using CRISPR or similar technologies be used to augment human cognitive ability? What if the same technology were used to 'correct' a genetic anomaly? The same concerns apply to privacy in the face of novel ubiquitous technologies and passive sensing systems. How should such data be used and who decides? Is it legally, ethically, or morally defensible to capture data about people, say facial images in public places, and draw inferences about their possible *future* behavior? Many of us might recoil at such a flagrant violation of our privacy and civil liberties, yet others might endorse the same systems if they effectively prevented terrorist acts. More directly relevant to advancing NP assessment using advanced technology: should our cars pull off the road and stop safely if

we appear fatigued or distracted? What about depositing sensitive personal health information into repositories? Can we ever assure that there will not be data breaches? How can we create genetics repositories that do not permit identification of individuals? We are indeed in a brave new world where these questions are no longer simply rhetorical or academic. Now, it is the time to consider these issues carefully and develop guidelines for the appropriate use of technologies that assess and intervene in human brain function.

A path forward

One potential path forward takes an incremental, step-by-step approach to building an infrastructure capable of replacing the current model for NP test development and breaking free of the traditional print publishing model within a 10-year period. This path updates ideas presented previously (Bilder, 2011) and focuses on a 'transition' strategy rather than a 'replacement' strategy. This plan takes the existing state-of-the-art (which we acknowledge may be outdated and suboptimal), provides a detailed specification of what we are measuring with current tools, and then makes those data universally available so that new methods can evolve.

This strategy requires: (a) substantial collaboration among multiple NP assessment centers internationally; (b) leveraging current efforts to support data sharing and knowledge aggregation; (c) development of credible, trustworthy strategies for aggregation of personal health data in a way that both protects individual privacy but enables sharing and use by qualified users for legitimate clinical and research purposes; (d) deployment of substantial expertise in applied psychometrics and statistical inference; (e) deployment of substantial expertise in technology, particularly (over the near term) in programming Web-applications; and (f) education and dissemination to enhance both understanding of and access to the new methods.

Create a repository of item-level NP test data

To achieve all the benefits associated with IRT and non-IRT trial-by-trial analytics, requires item-level data on large numbers of examinees spanning a full range of ability levels; in most cases, sample sizes in the range of 500 to 1000 are desirable (Jang, Wang, & Weiss, 2016). Unfortunately, standard practice often involves an administration, scoring, and recording process that culminates in dead-end filing cabinets, so the item-level data are lost to posterity. This reflects an enormous missed opportunity to capture and use for research purposes data from approximately 500,000 NP exams that are collected in the United States every year.

One solution would be to hire teams of data entry personnel; but this would require an enormous amount of redundant labor, with examiners recording item-level responses on a form at the time of acquisition, and then data entry staff transferring those entries from test form to database. A more pragmatic solution is to collect item-level data *at the point of testing* and directly transfer these data to an archive. Some individual clinics already use point-of-testing data input methods, using iPads instead of paper forms to record examinee responses. Pearson's Q-interactive serves a similar

function. Q-interactive is only available for selected tests, but these include leading selections according to Rabin et al. (2016), including the WAIS-IV, WMS-IV, CVLT-II, and D-KEFS tests. The primary problem with either of these solutions is that the data are not shared openly with the clinical and research communities for the common good. A secondary problem is that the point-of-testing software is not usually shared freely: individual clinics and/or companies may limit access or sell the software they paid to develop. How can we overcome these challenges?

An ideal solution would offer free software for point-of-testing data acquisition. This could be possible if either a consortium of NP organizations or other grant funding helped support the software development. Software maintenance would probably require some charge to the users for long-term use. For example, the NIH PROMIS (Patient Reported Outcomes Measurement Information System) initiative first received NIH support for development, but now charges only for ongoing use of Web-based, computerized adaptive test versions of the instruments, while many fixed instruments and items banks are available for free download.

Establish a neuropsychology data archive as an open-access global resource

What happens to the data acquired at the point-of-testing? The data would not be useful to the NP community if they are in private clinic or company databases. Fortunately, the NIMH has developed its own data archives specifically to provide shared access to the data of various types, including NP data, at the item level. Data templates already exist for the most widely used NP tests, including the WAIS-IV, WMS-IV, CVLT-II, D-KEFS, Wisconsin Card Sorting Test, Boston Naming Test, Finger Tapping Test, and more. Thus, the primary database infrastructure for this work does not need to be invented or constructed, and NIH already possesses a mechanism for its upkeep (although, if we start uploading 500,000 records per year, additional support might be needed). The NIMH Data Archive (NDA) already has developed policies for data sharing, de-identification, and security following federal guidelines (<https://data-archive.nimh.nih.gov/s/sharedcontent/about/policy>). Many university institutional review boards (IRBs) already have considered and approved the use of the Global Unique Identifier (GUID) system and deposition of data in the NDA for projects involving patients, in compliance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

Develop a global collaborative network

To obtain data rapidly on sufficiently large and diverse samples, a representative network of sites is needed. Given the aim to conduct IRT analyses, and that these analyses generally require at least 500 cases, about 1/1000th of NP cases seen globally each year need to be sampled to acquire useful data.

Major challenges for establishing a network involve managing the heterogeneity of assessment methods and ensuring that there are sufficient common grounds to enable cross-site data integration. Fortunately, despite the majority of us

neuropsychologists indicating that we use flexible approaches to assessment, there is considerable overlap in the actual instrumentation used (Rabin et al., 2016). For users of any collaborative database to have confidence, there must be established standards for data quality, and a minimal set of core demographic and basic clinical measures that participating sites agree to use. Network members need to agree to training standards, establish shared protocols including training manuals, semi-structured history-taking and interviewing guidelines, and adopt a single diagnostic system (e.g. International Classification of Disease and Health Conditions, 10th Edition).

The greatest diversity of practice is not actual NP testing but how historical and clinical interview data are collected. The NP network can overcome this challenge by using *common data elements* (CDEs) recommended by NIH, the World Health Organization (WHO), and other organizations. Among NIH common data elements (CDEs) for all health conditions (<https://www.nlm.nih.gov/cde>), the most relevant include: NINDS Common Data Elements (NINDS CDEs); Quality of Life Outcomes in Neurological Disorders (Neuro-QOL); NIDA Substance Abuse Electronic Health Record Data Elements; Patient Reported Outcomes Measurement Information System (PROMIS); the NIH Toolbox for Assessment for Neurological and Behavioral Function (NIH Toolbox), and the Consensus Measures for Phenotypes and eXposures (PhenX). The WHO created and distributes the WHODAS 2.0 (WHO Disability Assessment Scale, 2nd Edition), which enables assessment of disability spanning illness categories internationally, and has been endorsed as a CDE by the NIMH (Barch et al., 2016). These CDEs are freely available, and would enable assessment of demographics, medical history, medications prescribed, major psychiatric symptoms, and ratings of everyday functioning and disability.

To share clinical data in a repository involves a combination of local IRB and/or institutional compliance officers who govern medical records privacy and security. While the NP assessments may be done for clinical purposes, the sharing of de-identified NP assessment and clinical data is a research procedure and must be reviewed and approved by the appropriate IRBs. The approved procedures involve obtaining informed consent, but at most institutions this may be readily integrated with patients' consent to the terms and conditions of treatment. To facilitate this, the NIH developed a single IRB policy, and there is now a service to support collaborative research that many institutions already participate in, *Smart IRB* (<https://smartirb.org/>), that enables different institutions to have master reliance agreements and use a single IRB approval for multiple collaborating sites. These policy developments lower the burden for sites to participate.

This article has focused so far on patients, who provide data critically important for diagnostic validity studies, but for the development of new methods, we also need healthy comparison groups. A global network should study a number of healthy individuals each year. If spread evenly across sites world-wide, each site would only have to perform a handful of assessments each year (and not have to write reports about them).⁸ These contributions could complement the deposits of both healthy comparison group data along with clinical case data into repositories, which is being done already with existing research grants. In this way, a global network can serve as a test-bed for the introduction of novel NP assessment strategies and accelerate the delivery of new instruments to the global NP community.

Analyze the data to create new short forms, adaptive tests, evidence-based batteries, and dynamic interpretive evidence bases

The ultimate goal of developing NP tests of the future is to increase efficiency and validity so that more patients can access high-quality NP services. To achieve these goals, analyses of item-level data need to:

Specify the latent traits that are assessed by each test and identify the most efficient measurement models for each trait. We can then determine what alternatives exist to enhance efficiency. Simulations using these data can lead to validation studies for new procedures that are promising.

Examine measurement invariance across sites and groups defined by demographic or diagnostic characteristics. This process can identify differences between groups defined by age, sex, race, ethnicity, culture, education or income, and ultimately assure that we understand how the test metrics perform in discriminating between groups defined by diagnostic differences, and that are *not* due to confounds with individual or cultural differences (see also below, determining differential diagnosis).

Develop linking and data alignment procedures so that we can see how traits are measured redundantly across tests, enabling reduction in redundant assessment. We emphasized above how to improve efficiency *within* tests using IRT. There may be even greater benefit from improving efficiency *across* tests. Most NP batteries comprise tests with high intercorrelations and overlap in the constructs they are measuring. Measurement models using multidimensional or bifactor IRT can narrow test content to achieve specific assessment goals.

Determine what combinations (sequences) of tests best enable differential diagnostic classification, and/or individual characterization with precision sufficient for the purposes of differential recommendations. Consider our diagnostic questions in a Bayesian framework: 'Which procedure should I administer next to maximize the positive predictive power of the result?' Before we administer any test, we have some *a priori* hypothesis(es) about the diagnosis(es). We should select the diagnostic procedure that will maximize the likelihood of rejecting that prior and suggesting an alternate hypothesis. For example, in a dementia assessment, when the first test indicates impairment with high confidence, what additional testing is necessary to alter the final diagnosis and recommendations? Perhaps, no further information is needed if the differential diagnosis involves distinguishing only neurocognitive disorder versus no neurocognitive disorder. But different treatment recommendations might be made if dementia were due to Parkinson's disease compared to Alzheimer's disease; in that case, additional testing might be necessary to refine the diagnosis. Individual characterization questions are often even more complex, and clinical neuropsychology offers value in characterizing diverse abilities using objective, quantitative indicators. But how precise do those measures need to be to make different recommendations? How precisely do we need to define specific impairments to make differentiated recommendations? How do we balance the precision of measurement with the time and cost needed to obtain that level of precision? A global bank of item-level data will make it possible to address these questions rationally and make informed decisions.

Create standard reports for clinicians that summarize individual performance with respect to the entire current database, along with data from published normative and clinical comparison groups.

A major limitation of current NP assessments is that interpretation is based on static norms acquired when the test was published (often 5 to 20 years earlier). The proposed archive will enable comparisons to newly acquired normative data and to diverse clinical samples of ever-increasing size and more current than the original standardization and clinical validation data. The archive can also include newly published data, to which individual scores can be compared. This approach has been adopted by the company NeuroPsychNorms (<http://www.npnorms.com/>; accessed 6 April 2018). Future data aggregation software may even automatically extract key statistical parameters from published papers without human effort (Liu, Chu, Sabb, Parker, & Bilder, 2014).

Introduce novel items and tests

A global archive can lay the empirical foundation onto which new measurement methods can evolve. If new tests are based upon objective, quantifiable relations to previously defined tests and the constructs these measure, this directly addresses the ‘back-compatibility’ problem. New items and methods can evolve incrementally, as is conventionally done for scholastic aptitude tests, by introducing experimental content. Alternatively, entirely new tests may be introduced, and using IRT-based linking procedures, their relations to existing constructs can be defined and evaluated. As validity data accrue, entirely new test procedures may be found to possess greater validity with respect to diagnostic outcomes or neuroscientific measures, even if they are *not* clearly linked to previously defined constructs. Validating new methods directly with respect to clinical questions or biological substrates is superior to cross-validating a new test with respect to a previously defined test that we already know has imperfect validity.

A global archive can enable us further to examine the validity of novel methods that are *not* based on NP tests (for example, data from genetic testing or neuroimaging results, or data gleaned from medical records, mobile devices, wearables, or the internet-of-things). These developments can help align clinical neuropsychology to personalized medicine initiatives that are expanding rapidly in other domains of biomedical science.

Beyond psychometrics and technology

The best psychometric and technological strategies will be useful only if we know what we want to measure (Borsboom, 2005). NP approaches to understanding measurement need to reconsider what we mean by *validity*. The classical psychometric approach depended largely on ideas of a ‘nomological network’, that is the idea that we can understand a construct based on the correlations among variables thought to measure that construct (Borsboom, Mellenbergh, & van Heerden, 2004). Instead, we may be better served by a focus on causality that helps us understand cognitive

mechanisms, and determine if our measures explain external variables, including measures of brain circuit function, and real-world outcomes.

In *The Attack of the Psychometricians*, Borsboom (2006) writes that we in psychology may be suffering from '... a shortage of substantive theory that is sufficiently detailed to drive informed psychometric modeling' (p. 428). Borsboom continues:

This may be the central problem of psychometrics: psychological theory does not motivate specific psychometric models. It does not say how theoretical attributes are structured, how observables are related to them, or what the functional form of that relation is. It is often silent even on whether that relation is directional and, if so, what its direction is. It only says that certain attributes and certain observables have something to do with each other. But that is simply not enough to build a measurement model (p. 435).

Clinical neuropsychology has a compelling advantage over some disciplines because we possess a substantive theory rooted in the structure and function of the brain. But we still have difficult choices to make. We need to determine what will best advance our understanding of brain-behavior relations, in the context of specific assessment goals we have in particular cases. These goals will differ if our aim is to characterize functional capacities to recommend vocational placement, or if we need to advise a neurosurgeon about the likelihood that removing certain brain tissue will impair language skills. Neither case demands the *a priori* definition of latent traits that would then inform measurement methods. Instead, both aims are better served by measures that possess the best predictive validity with respect to specific outcomes. In the former case, the anatomic issues are moot and ecological validity is paramount, while in the latter case the functional-anatomic questions are specific, both in terms of the functional speech outcome and the anatomic delineation that may help decide the locus of irreversible surgery.

From this perspective, the current neuropsychological practice of defining 'domains' (e.g. 'language', 'verbal learning and memory', 'executive functions', etc.), may be hindering rather than helping our field. Given that the domain labels are fluid, but the actual test variables are defined operationally, it is critically important that we keep a clear focus on the measurement level, and document exactly how specific test variables relate to specific outcomes of interest (Bilder, 2012; Poldrack et al., 2011; Sabb et al., 2008). For example, Digit Span was referred to by Ebbinghaus as a measure of 'repetition', by Galton and Jacobs as a measure of 'prehension', and then variously during subsequent decades of analysis as an index of 'attention', 'freedom from distractibility', and 'working memory' or 'working with memory' (Bilder et al., 2009b). The correlates of Digit Span performance are not the same as the correlates of other individual measures of working memory (e.g. results of a Sternberg item recognition test), nor are the correlates of either of these tests the same as those with a factor reflecting diverse tests of working memory. Progress may be maximized if clinical neuropsychology invests now in a 'bottom-up' approach, where we let the data speak for themselves. Construct labels we use widely today may fail to reflect what we actually tested, and too easily lead to errors in conceptualization that can be avoided if we stick with the original variables. If the Digit Span test helps identify a person who has deficits impacting his schoolwork, that is more important than saying he has a

problem with 'repetition', 'prehension', or 'working memory', each of which is a generalization that may fail depending on the case.

This bottom-up approach and generation of large-scale datasets that can inform both definition of constructs and practice parameters based on empirical evidence is highly consistent with the current directions proposed for the Research Domains Criteria (RDoC) initiative and to enhance precision medicine by NIMH Director Joshua Gordon.⁹ Further, in line with Gordon's vision is the focus on computational methods to advance our specification of neurobehavioral measures that more closely reflect underlying brain mechanism and more clearly specify that our current diagnostic categories emerge from the signs and symptoms used to diagnose them, not the other way around. Indeed, there is not persuasive evidence that specific disease entities cause psychiatric symptoms, and good evidence to believe that is not true (Borsboom & Cramer, 2013; Friston, Redish, & Gordon, 2017; Redish & Gordon, 2016). The current emphasis on computational approaches in psychiatry research echoes Borsboom's call for a renaissance of mathematical psychology and for its integration with psychometrics; thereby, helping to increase the formalization of concepts and methods in psychology (Borsboom, 2006). Clinical neuropsychology is poised for a revolution in its concepts and methods, and it is hoped that some of the suggestions made above will help accelerate that process.

Notes

1. IRT models also can consider free-response items where there are in theory an infinite number of responses (i.e. in responses to the Vocabulary subtest, many different kinds of responses are observed in practice); the challenge is for test designers and examiners to capture and evaluate the different responses rapidly enough to be useful.
2. There are notable exceptions (e.g. partial credit is given for certain responses on Vocabulary, Similarities, or the Rey-Osterrieth Complex Figure Test), and different scores are given for other WAIS subtests if completed within certain pre-specified time limits.
3. It is noteworthy that one of the original leaders in the development of IRT was Fred Lord, a leader of development at ETS (Carlson & von Davier, 2013).
4. This example uses some plausible percentile levels but the actual CAT algorithm will select items based on item-information curves, not by picking percentiles.
5. Nielsen Scarborough (n.d.). Number of cell phone users in the United States from spring 2008 to spring 2017 (in millions). In Statista - The Statistics Portal. Retrieved 4 April 2018, from <https://www.statista.com/statistics/231612/number-of-cell-phone-users-usa/>.
6. CTIA, the International Association for the Wireless Telecommunications Industry. Retrieved 4 April 2018 from https://web.archive.org/web/20120820013725/http://www.ctia.org/consumer_info/index.cfm/AID/10323.
7. Deloitte Technology Survey, retrieved 4 April 2018 from <https://www2.deloitte.com/us/en/pages/technology-media-and-telecommunications/articles/global-mobile-consumer-survey-us-edition.html>.
8. Estimating 500,000 NP assessments each year in the United States, to obtain 1000 healthy comparison cases per year would only demand a healthy person be included for every 500 clinical NP assessments. Even assuming that only a fraction of all clinics participate, the distributed burden remains relatively low.
9. Gordon, J. The Future of RDoC By Joshua Gordon on 5 June 2017. Accessed 4/10/2018 at <https://www.nimh.nih.gov/about/director/messages/2017/the-future-of-rdoc.shtml>.

Acknowledgements

The authors are grateful for the input from our collaborators, including Russell Bauer, Daniel Drane, James Holdnack, David Loring, and David Sabsevitz.

Funding

Preparation of this manuscript was supported by grants from the National Institute of Mental Health (R01MH101478, R03MH106922, and U01MH105578).

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Abdullah, S., Murnane, E. L., Matthews, M., Kay, M., Kientz, J. A., Gay, G., & Choudhury, T. (2016). *Cognitive rhythms: Unobtrusive and continuous sensing of alertness using a mobile phone*. Paper presented at the Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany.
- Aramendi, A. A., Weakley, A., Schmitter-Edgecombe, M., Cook, D. J., Goenaga, A. A., Basarab, A., & Carrasco, M. B. (2018). Smart home-based prediction of multi-domain symptoms related to Alzheimer's disease. *IEEE Journal of Biomedical and Health Informatics*, 99(1), e91.
- Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 495–508.
- Barch, D. M., Gotlib, I. H., Bilder, R. M., Pine, D. S., Smoller, J. W., Brown, C. H., ... Farber, G. K. (2016). Common measures for National Institute of Mental Health funded research. *Biological Psychiatry*, 79(12), e91–e96.
- Bigdely-Shamlo, N., Makeig, S., & Robbins, K. A. (2016). Preparing laboratory and real-world EEG data for large-scale analysis: A containerized approach. *Frontiers in Neuroinformatics*, 10, 7.
- Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society*, 17(1), 7–13.
- Bilder, R. M. (2012). Executive control: Balancing stability and flexibility via the duality of evolutionary neuroanatomical trends. *Dialogues in Clinical Neuroscience*, 14(1), 39.
- Bilder, R. M., Sabb, F., Cannon, T. D., London, E. D., Jentsch, J. D., Parker, D. S., ... Freimer, N. B. (2009). Phenomics: The systematic study of phenotypes on a genome-wide scale. *Neuroscience*, 164(1), 30–42. doi:10.1016/j.neuroscience.2009.01.027.
- Bilder, R. M., Sabb, F. W., Parker, D. S., Kalar, D., Chu, W. W., Fox, J., ... Poldrack, R. A. (2009). Cognitive ontologies for neuropsychiatric phenomics research. *Cognitive Neuropsychiatry*, 14(4–5), 419–450. doi:10.1080/13546800902787180.
- Bilker, W. B., Hansen, J. A., Bressinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, 19(3), 354–369. doi:10.1177/1073191112446655.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29–51.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425. doi:10.1007/s11336-006-1447-6.

- Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*, 91–121. doi:10.1146/annurev-clinpsy-050212-185608.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061.
- Boto, E., Holmes, N., Leggett, J., Roberts, G., Shah, V., Meyer, S. S., ... & Barnes, G. R. (2018). Moving magnetoencephalography towards real-world applications with a wearable system. *Nature, 555*(7698), 657.
- Carlson, J., & von Davier, M. (2013). *Item response theory (ETS R&D Scientific and Policy Contribution Series ETS SPC-13-05)*. Princeton, NJ: Educational Testing Service.
- Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*(1), 125–136. doi:10.1007/s11136-009-9560-5.
- Collins, F. S., & Riley, W. T. (2016). NIH's transformative opportunities for the behavioral and social sciences. *Science Translational Medicine, 8*(366), 366ed314–366ed314. doi:10.1126/scitranslmed.aai9374.
- Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., ... van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology, 61*(10), 1018–1027. e1019.
- Dawadi, P. N., Cook, D. J., & Schmitter-Edgecombe, M. (2013). Automated cognitive health assessment using smart home monitoring of complex tasks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43*(6), 1302–1313.
- Dawadi, P. N., Cook, D. J., & Schmitter-Edgecombe, M. (2016). Automated cognitive health assessment from smart home-based behavior data. *IEEE Journal of Biomedical and Health Informatics, 20*(4), 1188–1194.
- Ebesutani, C., Reise, S. P., Chorpita, B. F., Ale, C., Regan, J., Young, J., ... Weisz, J. R. (2012). The revised child anxiety and depression scale-short version: Scale reduction via exploratory bifactor modeling of the broad anxiety factor. *Psychological Assessment, 24*(4), 833–845. doi:10.1037/a0027283.
- Ferrari, L. M., Sudha, S., Tarantino, S., Esposti, R., Bolzoni, F., Cavallari, P., ... & Greco, F. (2018). Ultraconformable temporary tattoo electrodes for electrophysiology. *Advanced Science, 5*(3), 1–11.
- Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational nosology and precision psychiatry. *Computational Psychiatry, 1*, 2–23.
- Gavett, B. E., Fletcher, E., Harvey, D., Farias, S. T., Olichney, J., Beckett, L., & Mungas, D. (2018). Ethnoracial differences in brain structure change and cognitive change. *Neuropsychology, 32*(5), 529–540.
- Gibbons, R. D., Clark, D. C., VonAmmon Cavanaugh, S., & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19*(1), 43–55.
- Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., ... Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*(4), 361–368. doi:59/4/361 [pii] 10.1176/appi.ps.59.4.361.
- Gold, J., Barch, D., Carter, C., Dakin, S., Luck, S., MacDonald, A., ... Strauss, M. (2012). Clinical, functional, and intertask correlations of measures developed by the Cognitive Neuroscience Test Reliability and Clinical Applications for Schizophrenia Consortium. *Schizophrenia Bulletin, 38*(1), 144–152. doi:10.1093/schbul/sbr142.
- Gregory, R. J. (2016). *Psychological Testing: History, Principles and Applications* (Updated 7th ed.). Boston, MA: Pearson Education Inc.
- Grube, R. W., Thomas, L. C., Craig, K. M., & Gast, C. M. (2017). *US Patent No. US 9,771,081 B2*. Washington, DC: U. S. P. Office.
- Hampson, R. E., Song, D., Robinson, B. S., Fetterhoff, D., Dakos, A. S., Roeder, B. M., ... Couture, D. E. (2018). Developing a hippocampal neural prosthetic to facilitate human memory encoding and recall. *Journal of Neural Engineering, 15*(3), 036014.

- Hocke, L. M., Duszynski, C. C., Debert, C. T., Dleikan, D., & Dunn, J. F. (2018). Reduced functional connectivity in adults with persistent post-concussion symptoms: A functional near-infrared spectroscopy study. *Journal of Neurotrauma*, 35(11), 1224–1232.
- Jagaroo, V. (2009). Obstacles and aids to neuroinformatics in neuropsychology. In *Neuroinformatics for neuropsychology* (pp. 85–93). Springer, New York, NY.
- Jang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, 7, 109. doi: [10.3389/fpsyg.2016.00109](https://doi.org/10.3389/fpsyg.2016.00109).
- Kim, D. S., Reise, S. P., & Bentler, P. M. (2018). Identifying aberrant data in structural equation models with IRLS-ADF. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 343–358.
- Lencz, T., Bilder, R., Turkel, E., Goldman, R., Robinson, D., Kane, J., & Lieberman, J. (2003). Impairments in perceptual competency and maintenance on a visual delayed match-to-sample test in first-episode schizophrenia. *Archives of General Psychiatry*, 60(3), 238–243. doi: [10.1001/archpsyc.60.3.238](https://doi.org/10.1001/archpsyc.60.3.238)
- Liu, C., Chu, W. W., Sabb, F., Parker, D. S., & Bilder, R. (2014). Path knowledge discovery: Multilevel text mining as a methodology for phenomics. In Chu, W. W., (Ed.), *Data mining and knowledge discovery for big data* (pp. 153–192). Berlin: Springer.
- Loring, D. W., Bowden, S. C., Staikova, E., Bishop, J. A., Drane, D. L., & Goldstein, F. C. (2018). NIH toolbox picture sequence memory test for assessing clinical memory function: Diagnostic relationship to the rey auditory verbal learning test. *Archives of Clinical Neuropsychology*. 1–9
- Mansolf, M., & Reise, S. P. (2018). Case diagnostics for factor analysis of ordered categorical data with applications to person-fit measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(1), 86–100.
- Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multigroup factor analysis comparison of latent means across many groups. *Psychological Methods*, 524–545. Advance online publication. doi:[10.1037/met0000113](https://doi.org/10.1037/met0000113).
- McHorney, C. A., & Cohen, A. S. (2000). Equating health status measures with item response theory: Illustrations with functional status items. *Medical Care*, 38, 1143–1159.
- McHorney, C. A., & Fleishman, J. A. (2006). Assessing and understanding measurement equivalence in health outcome measures: Issues for further quantitative and qualitative inquiry. *Medical Care*, 44(11), S205–S210.
- Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361–388.
- Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, 9(1), 3–8.
- Melroy, S., Bauer, C., McHugh, M., Carden, G., Stolin, A., Majewski, S., ... Wuest, T. (2017). Development and design of next-generation head-mounted ambulatory microdose positron-emission tomography (AM-PET) system. *Sensors*, 17(5), 1164.
- Meyer, O. L., Mungas, D., King, J., Hinton, L., Farias, S., Reed, B., ... Beckett, L. (2018). Neighborhood socioeconomic status and cognitive trajectories in a diverse longitudinal cohort. *Clinical Gerontologist*, 41(1), 82–93.
- Moore, T. M., Reise, S. P., Roalf, D. R., Satterthwaite, T. D., Davatzikos, C., Bilker, W. B., ... Gur, R. C. (2016). Development of an itemwise efficiency scoring method: Concurrent, convergent, discriminant, and neuroimaging-based predictive validity assessed in a large community sample. *Psychological Assessment*, 28(12), 1529–1542. doi:[10.1037/pas0000284](https://doi.org/10.1037/pas0000284).
- Moore, T. M., Scott, J. C., Reise, S. P., Port, A. M., Jackson, C. T., Ruparel, K., ... Gur, R. C. (2015). Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. *Psychological Assessment*, 27(3), 955–964. doi: [10.1037/pas0000102](https://doi.org/10.1037/pas0000102).
- Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & González, H. (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics. *Psychological Assessment*, 16(4), 347.

- Mungas, D., Reed, B. R., Haan, M. N., & González, H. (2005). Spanish and English Neuropsychological Assessment Scales: Relationship to demographics, language, cognition, and independent function. *Neuropsychology, 19*(4), 466.
- Mungas, D., Reed, B. R., Marshall, S. C., & González, H. M. (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. *Neuropsychology, 14*(2), 209.
- Mungas, D., Widaman, K. F., Reed, B. R., & Farias, S. T. (2011). Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology, 25*(2), 260.
- Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 978.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18. doi:10.1016/0022-2496(66)90002-2.
- Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & Group, P. C. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)): Depression, anxiety, and anger. *Assessment, 18*(3), 263–283. doi:10.1177/1073191111411667.
- Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., ... Bilder, R. M. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics, 5*, 17.
- Pompei, F. J., Sharon, T., Buckley, S. J., & Kemp, J. (2002). *An automobile-integrated system for assessing and reacting to driver cognitive load (No. 2002-21-0061)* (SAE Technical Paper, Warrendale, PA).
- Prentice, K. J., Gold, J. M., & Buchanan, R. W. (2008). The Wisconsin Card Sorting impairment in schizophrenia is evident in the first four trials. *Schizophrenia Research, 106*(1), 81–87.
- Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*(3), 523–550.
- Preston, K. S. J., & Reise, S. P. (2014). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement, 74*(3), 377–399.
- Preston, K. S. J., & Reise, S. P. (2015). Detecting faulty within-item category functioning with the nominal response model. In S. Reise & D. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 386–406). New York: Routledge.
- Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*(1), 33–65. doi:10.1016/j.acn.2004.02.005.
- Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology, 31*(3), 206–230.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517.
- Redish, A. D., & Gordon, J. A. (2016). *Computational psychiatry: New perspectives on mental illness*. Cambridge, MA: MIT Press.
- Reise, S., & Flannery, P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education, 9*(1), 9–26.
- Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg self-esteem scale. *Multivariate Behavioral Research, 51*(6), 818–838. doi:10.1080/00273171.2016.1243461.
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*(1), 27–48. doi:10.1146/annurev.clinpsy.032408.153553.
- Sabb, F. W., Bearden, C. E., Glahn, D. C., Parker, D. S., Freimer, N., & Bilder, R. M. (2008). A collaborative knowledge base for cognitive phenomics. *Molecular Psychiatry, 13*(4), 350–360. doi:10.1038/sj.mp.4002124.

- Sugarman, M. A., Woodard, J. L., Nielson, K. A., Smith, J. C., Seidenberg, M., Durgerian, S., Rao, S. M. (2014). Performance variability during a multitrial list-learning task as a predictor of future cognitive decline in healthy elders. *Journal of Clinical and Experimental Neuropsychology*, 36(3), 236–243.
- Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, 13(3), 201–214.
- Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4, 770.
- Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology*, 57(6), 1051–1058.
- Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., ... Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, 33(1), 101–111.