NIH Toolbox for Assessment of Neurological and Behavioral Function

Richard C. Gershon, PhD
Molly V. Wagster, PhD
Hugh C. Hendrie, MB, ChB, DSc
Nathan A. Fox, PhD
Karon F. Cook, PhD
Cindy J. Nowinski, MD, PhD

Correspondence to
Dr. Gershon:
Gershon@northwestern.edu

Supplemental data at
www.neurology.org

At present, there are many studies that collect information on aspects of neurologic and behavioral function (cognition, sensation, movement, emotion), but with little uniformity among the measures used to capture these constructs. Further, available measures are generally expensive, normed on homogenous nondiverse populations, not easily administered, do not cover the lifespan (or have easily linked pediatric and adult counterparts for the purposes of longitudinal comparison), and not based on the current thinking in the neuroscience community. There is also a paucity of measurement tools to gauge normal children in the motor and sensation domain areas, and many of these measures rely heavily on proxy reporting. Investigators have expressed the need for brief assessment tools that could address these issues and be used as a form of "common currency" across diverse study designs and populations. This ability to assess functionality along a common metric and "crosswalk" across measures is essential to the process of being able to pool data, which is often necessary when a large and diverse sample is needed. When individual studies employ unique assessment batteries, comparisons between studies and combining data from multiple studies can be problematic. The contract for the NIH Toolbox for the Assessment of Neurological and Behavioral Function (www.nihtoolbox.org) was initiated by the NIH Blueprint for Neuroscience Research (www.neuroscienceblueprint.nih.gov) to develop a set of state-of-the-art measurement tools to enhance collection of data in large cohort studies and to advance the biomedical research enterprise.

The NIH Toolbox was not conceptualized as a substitute for the in-depth assessment of a behavioral domain or subdomain, and does not specifically target disease outcomes in its current format. As such, it is not intended for use as a diagnostic tool. Nonetheless, it is the hope that the normative data for NIH Toolbox performance in neurologic, psychiatric, and other disorders will be generated in the future through other research mechanisms. Developed via a systematic, iterative process that involved content experts and stakeholders, the NIH Toolbox was envisioned to incorporate the following characteristics:

- Multidimensionality within each domain area
- Versatility in terms of the types of studies where it can be employed, the portability of the measures across study designs, and the ability to crosswalk to existing and previous studies through the use of embedded benchmark items
- Brevity to ensure low respondent burden and to address needs of researchers conducting large cohort studies
- Methodologically sound
- State-of-the-art in terms of psychometric approaches and technologies
- Diversity in terms of having known measurement properties across racial and ethnic groups and numerous age ranges, as well as availability of English and Spanish versions
- Dynamic to demonstrate sensitivity to change over time, and to allow for the adaptation of the measures over time in response to advances in science or technology

Importantly, the construction of NIH Toolbox assessments was based, where possible, on item response theory and adapted for testing by computer. These characteristics helped to diminish floor/ceiling effects and practice effects and contributed positively to the goal of brevity in assessment. Notable deviations are the motor assessments and some of the sensory assessments that do not lend themselves well to this type of psychometric approach.

OVERVIEW OF DEVELOPMENT Project development was divided into 2 phases. Initial project goals included the identification of salient criteria for the measures to include, the determination of specific subdomains for each of the 4 primary content areas, the identification of existing measures that met the NIH Toolbox criteria, modification of existing measures to meet the criteria, or the development of new instruments

From the Northwestern University Feinberg School of Medicine (R.C.G., K.F.C., C.J.N.), Chicago, IL; National Institute on Aging (M.V.W.), Bethesda, MD; Center for Aging Research (H.C.H.), Indiana University School of Medicine, Indianapolis; and University of Maryland (N.A.F.), College Park.

Go to Neurology.org for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

This article was written in part by authors in their capacity as NIH employees, but the views expressed in the article do not necessarily represent those of the NIH.

where needed. In phase II, candidate measures underwent pilot testing and initial evaluation of psychometric properties. This second phase continued with additional measure refinement and Spanish translation in preparation for norming, and ultimately delivery of the final product and procedure manual in September, 2012.

Requests for information. The NIH Toolbox development included gathering background information and soliciting information from the expert community (a detailed description of this process including results can be found in the article on surveying the end-user research community later in this issue). After experts were identified, they were solicited through multiple formal requests for information (RFI). Literature and database reviews facilitated the process of identifying 1) subdomain level criterion for NIH Toolbox inclusion, 2) existing measures relevant to the project goals, and 3) clinical and domain area experts. For example, the literature review helped to refine the list of subdomains and defined the significance of each subdomain relative to the assessment of functionality in that area.

The first RFI was initiated in November 2006. More than 200 experts were solicited in order to gather data related to the assessment of the NIH Toolbox domain areas. A follow-up consensus meeting was held in January 2007 to discuss the criteria that affected instrument selection, creation, and norming. This included the members from the NIH Project Team, an external panel of content experts, and contract scientists and staff. Subsequent to this, expert interviews were undertaken to gather more detailed information from clinical and scientific experts to help further refine the list of possible subdomains. Considerations for subdomain selection included conceptual relevance across the lifespan and significance to health and function, as well as practical issues regarding existing measures.

A second consensus group meeting was held in May 2007 to discuss subdomain content and functional constructs that should be integrated into the NIH Toolbox. A second RFI was also sent to approximately 300 experts in February 2008 requesting feedback regarding the characteristics of the NIH Toolbox, with the goal of better understanding end-user preferences pertaining to setup and administration, including equipment costs.

DOMAIN STRUCTURE The results of the earlier described activities have directed the decision for the final NIH Toolbox to assess 4 core domain areas (cognitive, emotional, motor, and sensory health and function). Each domain is composed of multiple interrelated subdomains which, in turn, include multiple subcomponents that, in the NIH Toolbox framework, are the functional constructs that are measurable representations of the underlying domain. These subdomains and constructs were arrived upon based on the activities described earlier and were considered to be the most important to assess in order to meet the goals of the NIH Toolbox. In addition, based on early feedback from an external panel of experts and research by the NIH Toolbox Steering Committee, the NIH Project Team recognized there would be much to be gained by supporting the creation of 6 unique sensory teams and by inviting the new domain leads from each of these teams to serve on the Steering Committee. These new sensory teams, as had the original domain teams, researched and developed measures of one or more unique constructs that met the general goals of the NIH Toolbox (e.g., brief assessments targeted at ages 3–85). The constructs assessed within each domain and subdomain are presented in the table.

Instrument selection. During these project activities, more than 1,400 potential existing instruments were identified and summarized. The selection criteria for considering an existing measure to be appropriate for the NIH Toolbox included its applicability across the age span, lack of intellectual property constraints, psychometric soundness, brevity and ease of use, applicability in diverse settings and with different groups, along with a preference for instruments that already had been validated and normed for use with individuals between 3 and 85 years old. Results of the instrument selection process yielded draft development plans being established for 61 different measures. Many of the selected measures were designed to assess the same domain or subdomain. In this case, the assessments were later "horse raced" against other measures to determine which of the instruments would yield better psychometric properties across the target age range.

Organizational structure. Once the basic domain framework was determined and criteria for inclusion were established, a large structure was created to oversee overall development, while at the same time granting independence to numerous small groups charged with carrying out most of the early development work. The Steering Committee was increased in size to its current format consisting of the principal investigator, multiple coinvestigators representing 9 domain teams (cognition, motor, emotional health, vision, audition, taste, olfaction, vestibular balance, and somatosensation), the lead NIH Project Officer, and several additional coinvestigators with particular expertise in assessment, early childhood development, aging, and epidemiologic research. This group met on a monthly basis (primarily by teleconference) in conjunction with five "domain managers." The domain managers were all coinvestigators associated with Northwestern University who dedicated up to full time in the early years to coordinate the activities of each of the domain teams. The domain teams oversaw the initial research to define the assessment needs of each domain, to review the

Table Constructs assessed within each domain and subdomain

Domain/Subdomain	Functional constructs
Cognition	
Executive function	Inhibitory control and cognitive flexibility
Episodic memory	Visual episodic memory
Language	Vocabulary comprehension, reading decoding
Processing speed	Visual processing speed
Attention	Visual attention
Working memory	Memory for stimuli presented visually and auditorily
Emotion	
Negative affect	Sadness, fear, anger; a supplemental apathy measure is also available
Positive affect	Positive feeling states, life satisfaction, meaning
Social relationships	Social support, companionship, social distress, positive social development; a supplemental measure of social network integration is also available
Stress and coping	Perceived stress, self-efficacy, coping strategies
Motor	
Endurance	Cardiopulmonary function, biomechanical and neuromuscular function at a particular intensity
Locomotion	Movement of body from one place to another
Strength	Muscle ability to generate force against physical objects
Dexterity	Small muscle movements which occur in body parts; the ability to coordinate fingers to manipulate objects quickly and accurately
Balance	Orienting the body in space, maintaining upright posture under both static and dynamic conditions, move and walk without falling.
Sensation	
Olfaction	Odor identification
Vestibular	Vestibular ocular reflex
Audition	Words-in Noise; supplemental measures of hearing thresholds and a hearing handicap inventory are also available
Vision	Visual acuity; a supplemental vision function health related quality of life measure is also available
Taste	The ability to perceive taste in different regions of the oral cavity
Somatosensation	Pain intensity and pain interference; measures of texture discrimination and kinesthesia were included in the validation study but were not retained for the final Toolbox

literature, to identify existing instruments, and to modify existing or develop new instruments destined to become part of the NIH Toolbox battery. In addition to the lead domain scientist and the domain manager, domain teams were made up of experts from institutions across the United States with expertise in the relevant constructs, as well as representatives from technology development, epidemiology, biostatistics, and pediatrics. One or more NIH Project Officers were also invited to each domain meeting to give oversight and to lend their personal expertise to each discussion. Over time, 30 additional instrument development teams were established by the domain teams to complete the instrument development and early validation studies. These temporary teams consisted of representatives from the domain teams, but primarily were populated by scientists new to the NIH Toolbox development effort and who were chosen for their expertise in assessing specific functional areas.

Several other teams supported the development process. The NIH Project Team, made up of 20 representatives from the NIH Institutes, Centers, and Offices that make up the NIH Neuroscience Blueprint, met by monthly teleconference to discuss issues and to give technical and administrative support for the NIH Toolbox. The Epidemiology/Biostatistics Team, made up of epidemiologists, statisticians, and psychometricians, met several times per month to establish common validation and norming goals. A member of this team participated in each of the 9 domain team meetings. A Technology Team, consisting of a full-time project manager, data architect, software developers, quality assurance, and customer service personnel, worked to automate the direct delivery and reporting for each of the assessments. The Spanish Language Team ensured that each of the instruments was as functional in Spanish as was is in English. The Multi-Cultural Team, made up of scientists who study cultural differences in assessment, reviewed literally thousands of items to ensure that they were appropriate for use across multiple cultural groups. Separate Pediatric and Geriatric Teams reviewed all assessments to ensure that content was as appropriate for 3-year-olds as it was for 85-year-olds. Finally, an Accessibility Team continuously reviewed items and the hardware and technology used to deliver them to insure compliance with Section 508 of the Rehabilitation Act, which requires federal agencies to make their electronic and information technology accessible to people with disabilities.[1] While it was our goal to provide assessments that are accessible to individuals with all disabilities, we realized early on that exceptions would have to made (e.g., blind people taking a vision test).

In total, more than 300 scientists and support personnel at over 60 institutions contributed to the development effort.

Field testing and validation. We anticipated that many of the identified existing instruments would not demonstrate acceptable validity across the complete NIH Toolbox age span. Further, we anticipated that some of the newly developed instruments would not survive rigorous test-retest and validation criterion against gold standard instruments. Gold standard instruments might have otherwise been included in the NIH Toolbox were it not for cost or concerns with total administration time (which in the NIH Toolbox was generally limited to 5 minutes or less per construct). We therefore created draft development plans for 61 new and existing measures to enable assessment of the 47 construct areas described above. Of these, 54 instruments were ultimately validated in sample sizes ranging from 300 to 700. Seven of the instruments had existing validation data across the age range. Overall, approximately 11,000 subjects participated in pretesting, validation, and calibration activities. All of the new emotional health items, the new vocabulary items, and the quality of life self-report scales for vision were calibrated using online panels. All other instruments were pretested and validated in face-to-face objective sessions at locations with specific domain-level expertise. Validation results for each domain are described in the domain articles that follow in this supplement.

In 2009, through funding opportunities realized by the American Recovery and Reinvestment Act, several research projects to validate and norm the instruments in clinical populations were awarded. One study (PI: V. Mark) evaluated the validity and feasibility of the NIH Toolbox in the acute neurologic inpatient rehabilitation environment. Another (PI: M. Husain) administered the NIH Toolbox to depressed and nondepressed patients with Parkinson's disease to assess validity, feasibility, and the unique and interactive effects of depression and Parkinson's disease on performance. A third project (PI: T. Jernigan) administered the NIH Toolbox Cognition Battery as part of a multi-institutional effort to build a shared database resource containing genetic, imaging, and neural phenotypic data for children and adolescents. Early reports from these groups have confirmed that the NIH Toolbox measures are valuable resources for assessing each respective area.

The NIH Toolbox contract has given support to these projects as part of a future goal to validate the NIH Toolbox in clinical settings. In addition, the GENORM project collected genetic material from all subjects in the norming sample to enable future research comparing genotypes with phenotypes represented in the NIH Toolbox.

Norming. Forty-seven instruments were administered to a national sample of persons ages 3–85, in both English and Spanish versions. The 47 instruments were combined into a single test battery that flows from both examiner and examinee perspectives. Instructions were "homogenized" to be presented in a common voice with prompts that are similar across instruments. Instruments with audio presentation were recorded by a single professional voice actor (in separate versions for adults and children, English and Spanish). Thousands of hours of software developer time were dedicated to ensure that all instruments are available in a common interface. Each subject response along with item level timing was stored automatically. The computer controlled the flow of test administration, automatically presenting the specific tests and test items appropriate for different age, language (English- and Spanish-speaking), or other subgroups. For research purposes the order of major domains was alternated.

Norming included a large English- and Spanish-speaking sample of at least 150 persons per age band (single year bands for children 3–17, and multiple-year age bands for adults 18–85). Five hundred sample members were readministered the entire battery 1 week later to assess test-retest reliability.

The final NIH Toolbox. The Technology Team and manual writing teams prepared the final release of the NIH Toolbox. Companion technology enables the administration and scoring of the total NIH Toolbox battery, individual domain batteries, or individual instruments. All of the norming data was centralized, cleaned, and analyzed to create population- and age-based norms for each of the instruments. Each domain team met again to confirm or modify scoring algorithms for each individual assessment, and in some cases recommended the creation of a "total domain" score (similar to a verbal or performance IQ). In some cases, instruments were modified slightly to improve administration or scoring but not to the extent that changes impacted the value of the normative data already collected.

WORK TO BE PRESENTED IN THIS SUPPLEMENT

The remaining articles in this supplement describe the NIH Toolbox construction process from conception through the current status of development and validation activities for each of the domains. Next is an article by Nowinski et al.,[2] titled *Input on NIH Toolbox inclusion criteria: Surveying the end-user community*, which describes the processes and recommendations produced by a series of surveys and consensus meetings regarding the content of the NIH Toolbox. An article by Victorson et al.,[3] *Using the NIH Toolbox in special populations: Considerations for the assessment of pediatric, geriatric, culturally diverse, non–English-speaking, and disabled individuals*, overviews the NIH Toolbox development processes that considered the importance of producing assessments that would be valid in English- and Spanish-speaking populations, across multiple cultural groups, and with particular attention to accessibility across numerous disabilities. The final series of articles provides detailed information about instrument development for each domain and includes test-retest reliability and validation evidence for most of the new assessments. The importance of each content area in the assessment of neurologic and behavioral function is described.[4–13] The last article presents a more detailed overview of the norming sample plan and procedures.[14]

DISCUSSION

The current NIH Toolbox is comprised of a core set of tasks that focuses on the cognitive, emotional, motor, and sensory function domains, and is a valid, reliable, multidimensional, and versatile tool that is also brief, diverse, state-of-the-art, and capable of being modified and updated in the future without losing the continuity or comparability of previously collected data. By using multiple constructs of each domain, the NIH Toolbox is capable of monitoring neurologic and behavioral function over time, and therefore can measure the domain constructs across developmental stages. This facilitates the study of functional changes across the lifespan, including evaluating intervention and treatment effectiveness. It is intended to be used as a set of selection tools that will complement and add to a given project, which will allow greater clarity and consistency in measurement across studies. This promotes comparability and aids in the development of a solid scientific base from which evidence-based practices can evolve.

AUTHOR CONTRIBUTIONS

Richard Gershon: drafting/revising the manuscript, analysis or interpretation of data, contribution of vital reagents/tools/patients. Molly Wagster: drafting/revising the manuscript, study concept or design, study supervision. Hugh Hendrie: drafting/revising the manuscript, study concept or design. Nathan Fox: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, study supervision. Karon Cook: drafting/revising the manuscript. Cindy Nowinski: drafting/revising the manuscript.

ACKNOWLEDGMENT

The authors thank everyone who contributed to the development of the NIH Toolbox for Assessment of Neurological and Behavioral Function. A list of contributors can be found in the online supplement at www.neurology.org.

STUDY FUNDING

This study is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research, NIH, under Contract No. HHS-N-260-2006-00007-C.

DISCLOSURE

R. Gershon has received personal compensation for activities as a speaker and consultant with Sylvan Learning, Rockman, and the American Board of Podiatric Surgery. He has several grants awarded by NIH: N01-AG-6-0007, 1U5AR057943-01, HHSN260200600007, 1U01DK082342-01, AG-260-06-01, HD05469, NINDS: U01 NS 056 975 02, NHLBI K23: K23HL085766, NIA: 1RC2AG036498-01, NIDRR: H133B090024, OppNet: N01-AG-6-0007. M. Wagster reports no disclosures. H. Hendrie currently receives research funding from NIH/NIA grants #R01AG009956, R24MH080827, 5R01AG026096-05, UF20303/U01AG022376, R01AG031222, R01AG019181, and R01AG029884. N. Fox is funded by NIH grants R37HD017899, MH074454, U01MH080759, R01MH091363, P50MH078105, and P01HD064653. He serves on the scientific board of the National Scientific Council for the Developing Child. K. Cook has received financial support from Center for Psychiatric Rehabilitation Boston University, InvivoData, Xenoport, BrightOucome, the NIH, Veteran's Affairs Research and Development, National Institute on Disability and Rehabilitation Research (NIDDR), and Agency for Healthcare Research and Quality (AHRQ). In addition to NIH Toolbox, Dr. Cook receives other funding from NIH (5RC1NR011804-02 and 1U5AR057943-01). She also is currently supported by grants from NIDDR (H133B090024) and AHRQ (1R03HS020700-01). C. Nowinski receives or has received research support from the NIH (contracts HHSN265200423601C, HHSN260200600007C, and HHSN267200700027C), the Department of Veteran's Affairs, the Analysis Group, Novartis, and Teva Pharmaceuticals. She has also received honoraria for writing an article for Medlink. Go to Neurology.org for full disclosures.

Received June 6, 2012. Accepted in final form July 6, 2012.

REFERENCES

1. Section 508 of the Rehabilitation Act, as amended by the Workforce Investment Act of 1998 (P.L. 105–220) [online]. Available at: http://www.section508.gov/. Accessed February 20, 2012.
2. Nowinski CJ, Victorson D, Debb SM, Gershon R. Input on NIH Toolbox inclusion critera: Sureveying the end-user community. Neurology 2013;80(suppl 3):S7–S12.
3. Victorson D, Manley J, Wallner-Allen K, et al. Using the NIH Toolbox in special populations: considerations for the assessment of pediatric, geriatric, culturally diverse, non-English-speaking, and disabled individuals. Neurology 2013;80(suppl 3):S13–S19.
4. Coldwell SE, Mennella JA, Duffy VB, et al. Gustation assessment using the NIH Toolbox. Neurology 2013;80 (suppl 3):S20–S24.
5. Rine RM, Schubert MC, Whitney SL, et al. Vestibular function assessment using the NIH Toolbox. Neurology 2013;80(suppl 3):S25–S31.
6. Dalton P, Mennella JA, Doty RL, et al. Olfaction assessment using the NIH Toolbox. Neurology 2013;80(suppl 3): S32–S36.
7. Varma R, McKean-Cowdin R, Hays RD, Vitale S, Slotkin J. Vision assessment using the NIH Toolbox. Neurology 2013;80(suppl 3):S37–S40.
8. Dunn W, Griffith JW, Morrison MT, et al. Somatosensation assessment using the NIH Toolbox. Neurology 2013; 80(suppl 3):S41–S44.
9. Zecker SG, Hoffman HJ, Frisina R, et al. Audition assessment using the NIH Toolbox. Neurology 2013;80 (suppl 3): S45–S48.
10. Cook KF, Dunn W, Griffith JW, et al. Pain assessment using the NIH Toolbox. Neurology 2013;80(suppl 3): S49–S53.
11. Weintraub S, Dikmen SS, Heaton RK, et al. Cognition assessment using the NIH Toolbox. Neurology 2013;80 (suppl 3):S54–S64.
12. Reuben D, Magasi S, McCreath H, et al. Motor assessment using the NIH Toolbox. Neurology 2013;80(suppl 3): S65–S75.
13. Salsman JM, Butt Z, Pilkonis PA, et al. Emotion assessment using the NIH Toolbox. Neurology 2013;80(suppl 3):S76–S86.
14. Beaumont JL, Havlik R, Cook KF, et al. Norming plans for the NIH Toolbox. Neurology 2013;80(suppl 3):S87–S92.