Input on NIH Toolbox inclusion criteria

Surveying the end-user community

Cindy J. Nowinski, MD, PhD
David Victorson, PhD
Scott M. Debb, EdD
Richard C. Gershon, PhD

Correspondence to
Dr. Nowinski:
c-nowinski@northwestern.edu

ABSTRACT

Objective: The NIH Toolbox is intended to be responsive to the needs of investigators evaluating neurologic and behavioral function in diverse settings. Early phases of the project involved gathering information and input from potential end users.

Methods: Information was collected through literature and instrument database reviews, requests for information, consensus meetings, and expert interviews and integrated into the NIH Toolbox development process in an iterative manner.

Results: Criteria for instrument inclusion, subdomains to be assessed, and preferences regarding instrument cost and length were obtained. Existing measures suitable for inclusion in the NIH Toolbox and areas requiring new measure development were identified.

Conclusion: The NIH Toolbox was developed with explicit input from potential end users regarding many of its key features. *Neurology*® 2013;80 (Suppl 3):S7–S12

GLOSSARY

RFI = request for information.

The phase I goals of the NIH Toolbox[1,2] development process included: 1) the identification of criteria for the acceptance of cognitive, emotional, motor, and sensory domain specific tasks in behavioral and neurologic research, 2) the identification of existing tests and measurement tools that could potentially be included in the NIH Toolbox, and 3) the selection of subdomain areas within each comprehensive domain to be targeted by the NIH Toolbox instruments. This article presents a summary of expert input obtained from potential research end users regarding these tasks. Please note that all domain-level results from phase I are not reported in this article, but are identified in the domain-specific articles found later in this supplement.[3–14]

METHODS AND SAMPLE DESCRIPTIONS

Overview. Obtaining phase I data was accomplished by conducting Medline literature searches and instrument database reviews, a formal online request for information (RFI) from the expert research community, an expert consensus group meeting, conducting a series of expert interviews, a second expert consensus group meeting, and then a second online RFI sent to additional clinical and domain-area experts who were identified during this process. The experts solicited during phase I activities were identified via literature searches, examination of the former Computer Retrieval of Information on Scientific Projects database (now known as the NIH Research Portfolio Online Reporting Tools), or by nomination of 1 of the 12 NIH science officers who comprised the NIH Toolbox Project Team at that time. Results from these activities were reviewed by individual domain teams, the NIH Steering Committee, external experts, and representatives from the 16 institutes that make up the NIH Blueprint for Neuroscience Research (the lead sponsor of the NIH Toolbox). Final recommendations about Toolbox inclusion criteria and content also incorporated feedback from NIH Project Team members.

Requests for information. The first RFI was distributed to content-area experts to identify the criteria for subdomain content and test selection. Respondents were asked to provide ratings regarding the importance of specific clinical components that should be incorporated into the NIH Toolbox. Measuring respondent-identified importance was accomplished by using a 4-point Likert scale that ranged from "Not important at all" to "Very

From the Department of Medical Social Sciences (C.J.N., D.V., R.C.G.), Northwestern University Feinberg School of Medicine, Chicago; and FACIT–Functional Assessment of Chronic Illness Therapy (S.M.D.), Elmhurst, IL.

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.

important." Characteristics that were rated included content of the test questions, assessment length, credibility of results, as well as quality, understandability, and interpretability of resulting information. Experts were also instructed to add other criteria they thought were omitted or were important given their particular areas of expertise.

To determine what the constructs within each domain area consisted of, as well as to identify important measurement properties, we solicited input from 232 experts. Experts were predominantly NIH-funded investigators who were actively involved in neurologic or behavioral research. Each expert was sent a link via e-mail in November 2006 along with a follow-up request in January 2007 if necessary. The RFI itself was conducted online and consisted of a series of questions relating to the assessment of cognition, emotion, motor, and sensory function. Respondents who were not members of the NIH Toolbox Steering Committee or NIH employees were offered a $100 honorarium for their participation. Only 8 of the external experts requested compensation. The overall response rate for the first RFI was 65% (n = 150). Respondents identified themselves as having expertise in the assessment of cognition (70%), emotion (48%), motor function (37%), and sensation (29%). The sample averaged 25 years of research experience and 17 years of clinical experience, including experience conducting clinical (86%) or longitudinal/epidemiologic research (87%). One hundred twenty-four respondents (83%) indicated that they had been the principal investigator on a clinical trial or a longitudinal/epidemiologic study. Every respondent identified having been responsible for enrolling numerous people into clinical trials. The majority of the sample was male (58%). Respondents indicated their primary area of expertise as being adult only (57%), pediatric only (18%), or both (25%). Respondents were also asked to rate their familiarity with any of the 4 NIH Toolbox domain areas. The majority of the experts indicated they had familiarity with the Cognition domain (63%), followed by familiarity with the Emotion domain (43%), Motor domain (33%), and the Sensory domain (27%). The specific areas of expertise or specialization are listed in table 1.

Table 1 Expertise and specialization[a]

	RFI 1[b]	Expert interviews[c]	RFI 2[d]
Clinical/developmental psychology	38	18	30
Neuropsychology	25	14	22
Epidemiology	21	9	20
Cognitive neuroscience	20	14	16
Psychometrics	16	11	22
Gerontology/geriatrics	15	7	17
Medicine	12	7	8
Health/rehabilitative psychology	7	—	5
Occupational/physical therapy	7	23	11
Neurology	6	14	11
Audiology/otology	3	5	8
Pediatrics	3	5	6
Psychology (other than neuropsychology)	3	25	1
Clinical trials methodology	2	7	—
Biostatistics	2	—	10
Neurophysiology	2	—	—
Physiatry/PM&R	1	—	6
Nursing	1	—	—
Neuroscience (other than cognitive)	—	7	1
Experimental psychology	—	—	16

Abbreviations: PM&R = physical medicine and rehabilitation; RFI = request for information.
[a] Data are percentages.
[b] Other RFI 1 write-in areas included neuroimaging, substance abuse, bioengineering, movement disorders, psychiatry, pediatric neurosurgery, pediatric orthopedics, speech communication, psychophysics, sensory perception, medical sociology, genetic epidemiology, and Alzheimer disease.
[c] For the expert interviews, other write-in areas included developmental cognitive neurology, emotion, neurophysiological, psychoneuroimmunologist, and sensory systems.
[d] Other RFI 2 write-in areas included ophthalmology, nutrition, chemical senses, otolaryngology, motor development, pediatric orthopedics, demography, food sensory, experimental neurology, and motor control.

In February 2008, a second RFI was distributed. This RFI focused on better understanding end-user preferences regarding administration time and cost of using the NIH Toolbox. The second RFI questionnaire was sent electronically to 305 experts selected from a new expanded list of NIH-funded investigators.

One hundred forty-three of the scientists solicited (47%) replied. This panel identified themselves as having experience with adults (53%), pediatrics (20%), or both (37%). Fifty-nine percent claimed expertise in cognition, 38% in sensory function, 38% in emotion, and 36% in motor function. A slight majority of respondents (52%) indicated that they had experience as a principal investigator of a large longitudinal or epidemiologic research study. Table 1 reports the respondents' identified field(s) of expertise.

Expert interviews. We conducted interviews with clinicians and scientists who had expertise in 1 or more domain areas, or who had conducted large cohort studies of these domains, to identify subdomains and criteria for measure selection. Interviews were administered by phone or electronically, including an option to schedule a follow-up phone interview with a staff member to complete the electronic form, or to schedule a follow-up phone debriefing in order to discuss the nature of their responses. The survey consisted of asking respondents to rank germane subdomains within their domain (or domains for generalists) area of expertise in order of how important each would be to include in the NIH Toolbox, ranging

from "Most important" to "Least important" on a Likert scale. Additionally, respondents were asked to provide the names of measures that could be suitable for inclusion into the NIH Toolbox.

Of the 44 total expert interviews conducted across all 5 groups, 64% were with males. The majority of the sample had experience working in academia (77%), with some having clinical experience (23%), and less having experience in government (9%) or industry (7%). Five percent of the experts noted that they did not have experience working in 1 of these 4 specific areas. See table 2 regarding the identified area(s) of expertise for the clinicians and scientists interviewed and table 1 for identified area(s) of specialization.

Consensus meetings. The first consensus meeting was held in January 2007. The goal of this meeting was to select criteria that would have an impact on instrument selection, creation, and norming. Attendees included members of the NIH Toolbox Steering Committee, the NIH Toolbox Project Team, and internal staff. The meeting pertained to the "non-negotiable" criteria that were considered essential for the NIH Toolbox to incorporate, as well as other criteria that could potentially be used, including: 1) providing consistent results even when administered by different people (interrater reliability), 2) responsiveness to real change, 3) being stable over time unless there is a true change in what is being measured, 4) having an equivalent Spanish translation, and 5) providing lifespan coverage of the construct. Group participants were asked which additional criteria they thought would enhance the NIH Toolbox's acceptability for use in large-scale, longitudinal, and epidemiologic studies, and clinical trials. Each expert created a list of these criteria, which was subsequently addressed during the meeting in detail. Some of the main focal points of the discussion included self-report vs performance-based measures, the need for instruments to be brief but still sensitive enough to measure functioning, ensuring validity (accurately assessing real-world functioning), ease of administration and scoring, the need for common scales across measures, computerized testing, and the ability to define the level of functionality by score.

The group was then instructed to select and rate the 5 most important criteria to the production of effective tests for the NIH Toolbox. The results were sorted by meaning ascribed to a criterion and strength of the votes each criterion received. The group discussed each criterion to a) determine broad themes, b) identify criteria that could be grouped within those themes, c) remove components that should be excluded from consideration, and d) keep criterion that should remain unchanged.

A second consensus meeting was held in May 2007 to determine what subdomains and functional constructs would be incorporated into the NIH Toolbox. Attendees included members of the NIH Toolbox Steering Committee, the NIH Toolbox Project Team, and an invited group of domain-specific experts identified by the NIH. Domain teams made recommendations based on the RFI, consensus meeting, expert interview data, and review of the literature. The group as a whole reviewed all available phase I data gathered.

Table 2 Expert interview-identified area of expertise

Ages	
Pediatric (3–5 y)	32%
Pediatric (6–12 y)	43%
Pediatric (13–17 y)	45%
Adult (18–34 y)	43%
Adult (35–64 y)	62%
Adult (65–85 y)	45%
Research	
Longitudinal	64%
Epidemiologic	32%
Clinical	55%

RESULTS

Request for information #1. Using a Likert scale ranging from "Very important" to "Not important at all," respondents were asked to rate how important certain characteristics were for the NIH Toolbox to include. The most important characteristics (>80%) were the following: 1) the NIH Toolbox be stable over time unless there is a true change in what is being measured (83%); 2) the Toolbox measure what it is supposed to measure (89%); 3) the Toolbox be responsive to real change (93%); and 4) interrater reliability (95%). Table 3 indicates the top characteristics identified for each response category.

Consensus meeting #1. During the first consensus meeting, the following broad themes were identified as the most important elements of the NIH Toolbox.

Validity. Participants thought the NIH Toolbox should be able to 1) predict gold-standard criteria for current and later function, 2) provide a strong conceptual and theoretical foundation, 3) allow for the ability to cross-walk to other scales, 4) provide results that predict real-world functioning, and 5) ensure usability for cross-study analyses and discussions.

Precision/accuracy. This addresses the need for NIH Toolbox measures to be sensitive across a full range of abilities and levels of functioning. This includes 1) capturing the healthy population, 2) being reliable, 3) helping to define the normal range and cross-over, and 4) covering the full range of each construct.

Table 3 Characteristics of the NIH Toolbox

Most frequent rating	% Rating very important
Very important	
Interrater reliability	95
Be responsive to real change	93
Measure what it is supposed to measure	89
Be stable over time unless there is a true change in what is being measured	83
Be suitable for use with a variety of racial and ethnic groups	69
Provide results that predict real-world functioning	64
Provide results that are clinically relevant	59
Have established norms for different age groups	58
Provide scores/results that are easy to understand	52
Concurrent validity	50
Internal consistency	49
Have a Spanish translation	45
Be easy to administer	45
Somewhat important	
Discriminant validity	58
Be available in separate but equivalent forms	56
Provide statistical corrections if practice effects can occur	54
Be able to assess the full spectrum of ability or severity for a given concept	52
Convergent validity	49
Provide comprehensive coverage for a given concept	48
Predictive validity	48
Place minimal burden on the test taker	47
Face validity	46
A little important	
Require no special training to administer	40
Be suitable for self-administration	36
Be available at no or minimal cost	35
Be suitable for interviewer administration	35
Be suitable for use with proxy respondents	34
Require no special equipment to administer	32
Evaluate what is important from the patient's perspective	30
Be easy to score	28
Be available in paper-and-pencil format	27
Not important at all	
Be suitable for self-administration	18
Be available in paper-and-pencil format	17
Require no special equipment to administer	14
Require no special training to administer	12
Be suitable for use with proxy respondents	12

Usability. This refers to facilitating the actual process of using the NIH Toolbox in a clinical setting so that it reflects 1) low participant burden and ease of administration and scoring, 2) availability at no or minimal cost, 3) having results that are easy to understand, and 4) having available instruments without intellectual property issues.

Innovative methodology. This produces flexibility of testing, addresses floor and ceiling concerns, and scores individuals and items on the same metric. This includes 1) instruments being brief and practical, 2) measures being adaptable over time, 3) instruments being able to incorporate computerized adaptive testing and scoring, and 4) measures utilizing item response theory.

Suitability for diverse populations. Group members believed that the NIH Toolbox needed to be able to be valid and reliable across numerous cultural and sociodemographic groups. This includes 1) suitability for use with a variety of racial/ethnic groups, education levels, and a broad population range, 2) usability across literacy levels, and 3) having norms for age, education, literacy, diverse groups, and racial/ethnic populations.

Additional miscellaneous criteria. The following other areas were mentioned: 1) having a variety of formats, 2) the availability of nonverbally based instruments, and 3) the evaluation of what is important from the patient's phenomenologic perspective.

Expert interviews. Expert interview results were domain specific. Cognition experts ranked subdomains in the following order of importance, with 1 being most important: executive function (mean = 1.89, SD = 0.93); processing speed (mean = 2.44, SD = 1.01); learning and memory (mean = 2.89, SD = 1.62); attention and working memory (mean = 3.11, SD = 1.45); language (mean = 4.22, SD = 0.97); and visuospatial function (mean = 5.44, SD = 1.01). Expert rankings of the subdomains of emotional health included negative affect (mean = 1.88, SD = 1.25); positive affect (mean = 3, SD = 2.33); emotion regulation (mean = 3.33, SD = 2.25); attachment (mean = 4, SD = 1.63); social integration (mean = 4, SD = 1.67); coping/resilience (mean = 4.43, SD = 2.23); externalizing problems (mean = 4.67, SD = 1.86); and self-efficacy (mean = 5.67, SD = 2.73). Motor experts indicated that locomotion was most important to assess (mean = 1.75, SD = 1.42) followed by upper extremity function (mean = 1.83, SD = 1.11); strength (mean = 4.25, SD = 1.76); balance (mean = 4.33, SD = 1.23); endurance (mean = 4.58, SD = 2.02); dexterity (mean = 5.09, SD = 2.21); and flexibility (mean = 6.58, SD = 1.31). With respect to sensory function, respondents believed that vision (mean = 1.57, SD = 0.79) and hearing were most important to evaluate (mean = 2.71, SD = 2.14) followed by balance (mean = 3.67, SD = 2.34); pain (mean = 3.86, SD = 0.69); proprioception (mean = 5.86, SD = 2.12); olfaction (mean = 5.86, SD = 1.46); taste

Table 4 Maximum time per domain

	10 min, %	20 min, %	30 min, %	40 min, %	50 min, %	60 min, %	Would not assess domain area, n
Cognitive function	11	21	26	13	5	24	4
Emotional health	28	35	19	8	2	8	6
Motor function	34	37	23	5	2	0	12
Sensory function	38	34	19	4	2	4	11

(mean = 6.00, SD = 2.08); and touch (mean = 6.33, SD = 1.21).

Consensus meeting #2. After a full review of the results of the previous consensus meeting and the results of the expert interviews, the second consensus meeting identified 5 to 6 major areas for ongoing concentration by each of the domain teams. The Cognition domain was directed to focus on the areas of Attention, Memory, Executive Function, Processing Speed, and Vocabulary. The Emotional Health domain aimed to assess Positive Affect, Negative Affect, Stress and Coping, and Social Relationships. The Motor domain team was directed to focus on the assessment of Endurance, Dexterity, Locomotion, and Upper/Lower body Strength. The Sensory Function domain aimed to assess Taste, Audition, Somatosensation, Olfaction, Vestibular Balance, and Vision. It should be noted that subsequent to this meeting, 6 additional domain teams were created to further the work of each of the major areas identified in Sensory Function.

Objective measurements were recommended for all domains except for emotional health where subject self-report (with proxy reports given for young children) is in keeping with primary practice within the emotional health community. Subdomain selection for each of these areas is discussed in the domain-specific articles found later in this supplement.[3–14]

Request for information #2. This sample of experts was asked how much cost there should be for the equipment needed to administer the NIH Toolbox, as well as per-subject costs to gather data. In relation to cost, most respondents (67%) were in acceptance of domain team recommendations of total equipment costs of $4,000. Eighteen percent wanted to limit equipment costs to $2,000 and recommended changing the type of equipment to control the cost, and 10% preferred to limit equipment costs to $2,000 by reducing the number of domains.

Most respondents (82%) indicated that the maximum amount of time allotted for the setup, administration, and cleanup of all 4 domain areas should not exceed 2 hours for older children and adults. Sixty-one percent said that 60 minutes was the maximum allowable time for children who are 3 through 5 years of age, with another 20% indicating that 90 minutes was the maximum amount of time. Respondents were also asked to indicate the maximum time that they felt comfortable for the allotment of individual domain batteries (see table 4).

CONCLUSION

Phase I of the development for the NIH Toolbox included numerous stages of data collection, including soliciting information from domain-area experts from the clinical and research communities, as well as reviewing existing relevant literature. Data obtained were compiled and discussed during multiple consensus group meetings, and results were directly applied to the development of the NIH Toolbox. The iterative and multistep procedures used during phase I are consistent with the NIH's desire to ensure that the NIH Toolbox was developed using methodologies with explicit input from diverse and multidisciplinary research communities who will be the likely end users of this final product.

AUTHOR CONTRIBUTIONS

Cindy Nowinski: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, study supervision. David Victorson: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, study supervision, obtaining funding. Scott Debb: drafting/revising the manuscript, analysis or interpretation of data, statistical analysis. Richard Gershon: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, contribution of vital reagents/tools/patients.

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

C. Nowinski receives or has received research support from the NIH (contracts HHSN265200423601C and HHSN260200600007C), Teva Pharmaceuticals, and Novartis. She has also received honoraria for writing an article for Medlink. D. Victorson holds stock options in Eli Lilly and Company, received an honoraria for serving on the Steering Committee of the Reeve Neuro-Recovery Network, was funded by NIH contracts HHSN265200423601C and HHS-N-260-2006-00007-C and grants R01HD054569-02NIDRR, 1U01NS056975-01, and R01 CA104883, received support from the American Cancer Society (national and Illinois Division) for research in prostate cancer, received institutional support from NorthShore University HealthCare System for research in prostate cancer, received institutional support from the Medical University of South Carolina for sarcoidosis research, and received institutional support from the Northwestern Medical Faculty Foundation for urology research. S. Debb reports no disclosures. 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; National Institute of Neurological Disorders and Stroke: U01 NS 056 975 02; NHLBI K23: K23HL085766; NIA: 1RC2AG036498-01; NIDRR: H133B090024; OppNet: N01-AG-6-0007. Go to Neurology.org for full disclosures.

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

REFERENCES

1. Gershon RC, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. Assessment of neurological and behavioural function: the NIH Toolbox. Lancet Neurol 2010;9:138–139.
2. Gershon RC, Wagster MV, Hendrie HC, Fox N, Cook KF, Nowinski CJ. NIH Toolbox for Assessment of Neurological and Behavioral Function: introduction. Neurology 2013;80(suppl 3):S2–S6.
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.