Norming plans for the NIH Toolbox

Jennifer L. Beaumont, MS
Richard Havlik, MD
Karon F. Cook, PhD
Ron D. Hays, PhD
Kathleen Wallner-Allen, PhD
Samuel P. Korper, PhD
Jin-Shei Lai, PhD
Christine Nord, PhD
Nicholas Zill, PhD
Seung Choi, PhD
Kathleen J. Yost, PhD
Vitali Ustsinovich, MA
Pim Brouwers, MD
Howard J. Hoffman, MA
Richard Gershon, PhD

Correspondence to Ms. Beaumont: j-beaumont@northwestern.edu

ABSTRACT

Objective: The NIH Toolbox for Assessment of Neurological and Behavioral Function (NIH Toolbox) is a comprehensive battery of brief assessment tools. The purpose of this article is to describe plans to establish normative reference values for the NIH Toolbox measures.

Methods: A large sample will be obtained from the US population for the purpose of calculating normative values. The sample will be stratified by age (ages 3–85 years), sex, and language preference (English or Spanish) and have a total sample size of at least 4,205. The sample will include a minimum of 25–100 individuals in each targeted demographic and language subgroup.

Results: Norming methods will include poststratification adjustment calculated using iterative proportional fitting, also known as raking, so that the weighted sample will have the same distribution on key demographic variables as the US population described in the 2010 Census.

Conclusions: As with any set of norms, users should be mindful of the reference population and make conclusions consistent with the limitations of normative sampling, since it is not a probability-based sample. However, the NIH Toolbox norming study has been designed to minimize bias and maximize representativeness and precision of estimates. The availability of a "toolbox" of normed measures will be an important foundation for addressing critical research questions in neurologic and behavioral health. *Neurology*® 2013;80 (Suppl 3):S87–S92

The NIH Toolbox for Assessment of Neurological and Behavioral Function (NIH Toolbox) initiative was created in 2004 under the auspices of the NIH Blueprint for Neuroscience Research.[1] After evaluation of nearly 1,400 existing tests, 48 instruments were identified for development and inclusion in the NIH Toolbox, a comprehensive battery of brief assessment tools. Some of the selected instruments were available for immediate inclusion, while others are the results of refinements of existing instruments, and others still were newly developed in this effort. Collectively, these instruments will quickly and effectively measure motor, cognition, sensation, and emotion domains in individuals ranging in age from 3 to 85 years. The 48 NIH Toolbox instruments were developed using state-of-the-science methods to maximize their precision and efficiency and ensure their usefulness across the lifespan. The reliability and validity of the instruments are evaluated elsewhere (e.g., Rine et al.[2]). An additional objective of the NIH Toolbox collaborative was to facilitate interpretation of scores by creating reference values (norms). Example questions that norms answer are "How does this person's score compare to the general population?" "How does this person's score compare to the score of another person her age?" and "How does this score compare to this person's score at an earlier or later age?"

The development of norms requires several steps,[3,4] including 1) identifying the relevant population, 2) determining the statistics to be computed (e.g., medians, means, SDs, percentile ranks) and designing a sampling plan that yields allowable levels of sampling error and adequate sample size, 3) collecting the data, 4) computing normative values for all groups and subgroups of interest and developing tables that report norms, and 5) disseminating guidance for

From the Department of Medical Social Sciences (J.L.B., K.F.C., J.-S.L., S.C., V.U., R.G.), Northwestern University Feinberg School of Medicine, Chicago, IL; Westat (R.H., K.W.-A., S.P.K., C.N., N.Z.), Philadelphia, PA; University of California–Los Angeles (R.D.H.), Los Angeles; Department of Health Sciences Research (K.J.Y.), Mayo Clinic, Rochester, MN; and National Institutes of Health (P.B., H.J.H.), Bethesda, MD.
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.

interpreting the obtained norms. Presented in this report are the methodologic considerations and decisions used to develop a norming plan for the NIH Toolbox measures.

RELEVANT POPULATION The first decision made in developing a plan for NIH Toolbox measures was to collect data from both children and adults and both English and Spanish speakers. This norming sample reflects the intended use of the NIH Toolbox. The NIH Toolbox's target population is the estimated 285 million civilian, noninstitutionalized English- or Spanish-speaking individuals, ages 3–85, living in the United States.[5] More specifically, the target population includes persons with the following characteristics: 1) community-dwelling and noninstitutionalized, 2) ages 3–85 years, 3) capable of following test instructions (English or Spanish), and 4) able to give informed consent or, in the case of children age 8 or older, give assent with accompanying informed consent by proxy (i.e., parent/guardian). For a subset of measures, additional eligibility criteria include adequate visual, auditory, vestibular, or motor functioning (with or without assistance or assistive devices). In addition to national norms, norms are planned for age, sex, and primary or dominant language (English or Spanish) subgroups. Future studies may aim to create norms in different populations.

NORMING STATISTICS AND SAMPLING PLAN The NIH Toolbox team identified the need for several norming statistics (percentile ranks, medians, means, and SDs) computed for the reference population and for relevant subgroups.

Sampling design. A stratified sampling strategy is proposed for the NIH Toolbox norming study. In this approach, nonoverlapping categories called "strata" are defined for each demographic subgroup. For the NIH Toolbox norming study, strata are defined by age, sex, and primary language (tables 1 and 2). The language stratification was implemented due to the parallel English and Spanish versions of the NIH Toolbox and to ensure that a minimum number of Spanish-speaking participants were enrolled. Age was considered an important stratification factor because performance on many instruments of the NIH Toolbox was expected to vary greatly by age and we recognized a need to capture the extent of developmental change. Sex was included because of the ease of doing so and the important face validity of having equal participation of males and females in the study. Motor and Emotion Domain scores were also expected to differ by sex. Age is based on the last birthday (e.g., age 3 includes those 3 years and 0 days through 364 days). Census data estimates indicated that fewer than 2% of children 8–17 years of age have Spanish as their dominant language (i.e., report themselves as Spanish-speakers who either do not speak English or speak English "but not well")[5]; therefore, we will not attempt to populate these strata (0 sample size is entered for the Spanish columns for these 10 rows in table 2). While it was deemed cost-prohibitive to add further stratification factors, within each age stratum, target quotas were set relative to the US population distribution of race, ethnicity, and level of education (parents' education for children).

Sample size. A total sample size of 4,205 individuals is planned to ensure that at least 25 to 100 individuals

Table 1 Stratified sample of adults planned for the NIH Toolbox norming study

	English		Spanish		
Age, y	Male	Female	Male	Female	Total
18-29	50	50	25	25	150
30-39	50	50	25	25	150
40-49	50	50	25	25	150
50-59	50	50	25	25	150
60-69	50	50	25	25	150
70-79	50	50	25	25	250
80-85	50	50			
Total	350	350	150	150	1000

Table 2 Stratified sample of children planned for the NIH Toolbox norming study[a]

	English		Spanish		
Age, y	Male	Female	Male	Female	Total
3	50	50	50	50	200
4	50	50	50	50	200
5	50	50	50	50	200
6	50	50	50	50	200
7	50	50	50	50	200
8	100	100	0	0	200
9	100	100	0	0	200
10	100	100	0	0	200
11	100	100	0	0	200
12	100	100	0	0	200
13	100	100	0	0	200
14	100	100	0	0	200
15	100	100	0	0	200
16	100	100	0	0	200
17	100	100	0	0	200
Total	1,250	1,250	250	250	3,000

[a] Fewer than 1.5% of children 8–17 years of age have Spanish as their dominant language; therefore, children in these strata will not be enrolled.

per stratum of each targeted subgroup are included (tables 1 and 2). In addition to the 4,000 individuals depicted in tables 1 and 2, 105 pregnant women (35 pregnant less than 3 months and 70 pregnant 3 months or more) and 100 mothers (mothers of children listed in table 2) were included in the sample. Funding for these activities was provided jointly by the NIH Blueprint for Neuroscience Research and the National Children's Study. These proposed sample sizes will provide 95% confidence intervals within strata with approximate precision of ± 0.20 ($n = 100$) to ± 0.39 ($n = 25$) SD units. Additional analyses may combine strata to achieve greater levels of precision.

Measures. The NIH Toolbox comprised both primary and supplemental measures. The supplemental measures are ones endorsed by the study team but are not among the core measures of the NIH Toolbox. A measure could be designated as "supplemental" rather than as part of the NIH Toolbox for a number of reasons: 1) evidence supporting an instrument's reliability and validity is insufficient or only available for a limited age range, 2) its inclusion would increase the NIH Toolbox administration time beyond a level deemed acceptable to most researchers, or 3) its expense would make the NIH Toolbox cost-prohibitive. A matrix sampling design is planned to obtain data on supplemental measures. This design will result in smaller sample sizes for some instruments and, therefore, norms with wider confidence intervals than those obtained for the primary NIH Toolbox measures. However, this approach was judged to be cost-effective for nonprimary NIH Toolbox instruments.

To evaluate the representativeness of our norming sample, we will collect additional demographic variables including race/ethnicity and level of education of adult participants and parents of child participants. These results will be compared against known values in the reference population.

DATA COLLECTION Delve, Inc., a market research company, has been contracted to administer the NIH Toolbox measures to a sample randomly selected from existing databases maintained by Delve, La Verdad, and Facts 'n Figures market research companies following the NIH Toolbox sampling plan. These databases were assembled using a variety of methods including online self-enrollment, enrollment events hosted by the companies, and random telephone calls from market research representatives. Sites for the norming study (Atlanta, Chicago–Oak Brook, Cincinnati, Columbus, Dallas, Los Angeles, Minneapolis, Philadelphia, Phoenix, St. Louis) were selected to correspond with Delve office locations and to maximize the ability to meet sample cell size requirements.

Potential study participants will be randomly selected from the existing databases. Selected individuals will be called at home and screened to ascertain eligibility in sociodemographic and linguistic categories defined by the NIH Toolbox sampling plan. If eligible, a testing appointment will be scheduled at a nearby testing location. Scheduled respondents will be sent a package of detailed information regarding what they can expect on the day of testing, directions to the testing site, and a telephone number to call if they have questions about the study. They will also receive a reminder call about their appointment 2–3 days in advance and given the opportunity to reschedule if the scheduled appointment is no longer convenient. In addition, participants who do not keep their testing appointment will receive a follow-up call inviting them to reschedule.

Delve field technicians will be trained by the NIH Toolbox staff to administer all study measures to participants using a train-the-trainer model. Training will take place over 4½ days in Chicago. Technicians will return to their sites to practice for 1 week. Then they will go to St. Louis for live-testing observation by 2 professionals from the NIH Toolbox staff. Examiners will use a certification log documenting critical aspects of each test as feedback to these administrators. Upon completing the certification process, each technician will return to his or her site and train his or her own staff using the same model. After 1 week of training and a second week of practice, professional certifiers from the NIH Toolbox will be sent to each site to observe and certify the local administrators. Some individuals may require more practice or be deemed uncertifiable. If it is determined that a technician needs more practice, the individual will be retested for certification by a trained and certified regional manager. A month into the testing, the NIH Toolbox professionals or Delve staff trained and certified by the NIH Toolbox staff will audit the sites to ensure the tests are being administered correctly.

Examiners will do very little data entry, since most tests feed data directly into the assessment laptop. The only significant data entry effort will be entering data from the initial paper questionnaire into the database. Every data entry will be double checked by a person other than the one who made the entry. Weekly data extractions will be conducted by the NIH Toolbox team and administrator logs will be examined to identify assessment problems that need to be addressed.

Proxies (i.e., parents/guardians) will be included for child participants. The preferred proxy is the household member with the most knowledge about the child. Emotion measures will be completed by proxy for children ages 3–7. A subset of the emotion measures for children ages 8–12 will be completed by both self-report and proxy. For all pediatric participants (i.e., ages 3–17), proxies will complete

questions relating to personal and household demographics and health history. Proxy respondents for adult participants will not be allowed as development and validation of a separate proxy battery is outside the scope of this project.

We estimate the time required for adults and children ages 8–17 to complete the 4 modules of the NIH Toolbox to be approximately 2 hours. For children ages 3–7 years, we estimate testing time to be 1–1.5 hours. Supplemental NIH Toolbox instruments and additional questionnaires will add roughly 30–60 additional minutes to assessment time.

COMPUTING NORMATIVE VALUES

Methods to minimize and quantify nonresponse. Several strategies will be used to maximize response rates. As described under Data Collection, these include sending scheduled respondents a detailed package of information that includes directions to the testing site and information regarding what participants can expect on the day of testing, a reminder call about their appointment 2–3 days in advance, and the opportunity to reschedule a missed appointment. In accordance with Delve's standards, compensation of $120 will be provided to adult participants and $90 to families of child participants (given to the child if the child is old enough to provide assent) who complete testing.

The sample is designed to achieve target numbers within each cell. Normative values will be calculated within each cell. This will ensure that the impact of any nonresponse bias is contained and does not spread across cells. To evaluate the representativeness of the sample, we will compare weighted sample respondents to the US subpopulations (age, sex, and language) on a set of demographic variables such as geographical region, household income, education, race, and household size. We will compare early study participants with later participants to see if household characteristics and demographics are comparable.

Computing norms. Sample weights will be constructed using the following 4 variables: 1) sex (male, female), 2) age (see tables 1 and 2 for strata), 3) race/ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, and non-Hispanic other), and 4) education (less than high school, high school diploma/general educational development, and more than high school; using parents' education for children). Iterative proportional fitting, or raking, will be used because of the sparseness of some cells in a 4-way cross-tabulation. The method of raking requires using an iterative proportional fitting procedure under marginal constraints. The marginal population distributions of sex, age, and race/ethnicity will be obtained from Census 2010 Summary File 1.[6] The education distribution will be estimated by analyzing the 2010 American Community Survey. The iterative proportional fitting procedure was first introduced by Deming and Stephan,[7] and more details can be found in Bishop et al.,[8] Fienberg,[9] and Little and Rubin.[10] Before the computation of the weights begins, we will impute random values for cases with missing data on race/ethnicity or education using probabilities proportional to observed distributions.

Missing data will not be imputed. Normative values will be based on available data only. Order of administration of the tasks is randomized to control for the impact of missing data that might occur due to time constraints. Summary tables of norms for each of the NIH Toolbox instruments will be prepared by age, sex, and language using the weighted data. Descriptive statistics presented in the summary tables will include means and SDs, percentiles (minimum, 5th percentile, 25th percentile, median, 75th percentile, 95th percentile, maximum), and the frequency of respondents at the floor and ceiling. Because the scores for the NIH Toolbox instruments are on many different measurement scales, rank-based normalized scores will also be calculated and scaled to have a mean of 100 and a SD of 15 to allow for simplified interpretation and comparison among instruments. Normalized scores adjusted for age and multiple demographic variables (e.g., age, sex, education, language) will be calculated using linear regression techniques. A user's guide will be prepared that explains the methods and includes normative reference tables. The public release of the NIH Toolbox software will provide data exports that include these normalized and adjusted scores in addition to the instrument raw scores on their original scale. The full dataset for this norming study will also be de-identified and available on request for researchers who may wish to calculate normative values using an alternative methodology.

GUIDELINES FOR USE OF THE NIH TOOLBOX NORMS

The NIH Toolbox is a battery for assessing neurologic and behavioral function that will facilitate comparison across future studies. The availability of norms for the NIH Toolbox measures will be a unique resource for investigators. However, NIH Toolbox users must be aware of limitations in interpretation of these norms. This is especially true when making comparisons in extreme percentile ranges (e.g., 5% and 95%). The NIH Toolbox was designed to discriminate within the general population and not to differentiate "normal" from "abnormal" scores. Particular caution should be exercised in interpreting norms that will be based on smaller sample sizes (e.g., some subgroups and some supplemental NIH Toolbox measures).

When interpreting the NIH Toolbox norms, investigators should keep in mind the reference population: noninstitutionalized individuals age 3–85 dwelling in

the community, and cognitively able to give informed consent or assent where appropriate. For some measures, participants will have to have adequate visual, auditory, vestibular, or motor functioning (with or without assistance or assistive devices) to complete measures.

During norming, additional information on race, education, medical status, and other relevant factors will be collected that will allow us to compare the normative sample, even though it is not a probability-based sample, to the US population. These results will assist in the proper interpretation and use of the data. Substantial effort will be expended to obtain a highly representative sample; however, it should be noted that this sample will be collected using quota sampling techniques as opposed to fully random sampling techniques. Thus, the estimated standard errors may differ from the true standard errors. Furthermore, participation rates are expected to vary by age and other demographic variables, reducing certainty of calculated normative values for some subgroups.

Though caution is advisable in interpreting norms from any study, the NIH Toolbox norming study has been designed to minimize bias and maximize representativeness and precision of estimates. Knowing the expected distributions and variability of test results will allow investigators to calculate appropriate sample size estimates for future studies. Even more important, the availability of a "toolbox" of normed measures will be an important foundation for addressing critical research questions regarding neurologic and behavioral health.

AUTHOR CONTRIBUTIONS

Richard Havlik: drafting/revising the manuscript for content, study concept or design. Karon F. Cook: drafting/revising the manuscript for content, study concept or design. Ron D. Hays: drafting/revising the manuscript for content, study concept or design. Kathleen Wallner-Allen: drafting/revising the manuscript for content, study concept or design. Samuel P. Korper: drafting/revising the manuscript for content, study concept or design. Jin-Shei Lai: drafting/revising the manuscript for content, study concept or design. Christine Nord: drafting/revising the manuscript for content, study concept or design. Nicholas Zill: drafting/revising the manuscript for content, study concept or design. Seung Choi: drafting/revising the manuscript for content, study concept or design. Kathleen J. Yost: drafting/revising the manuscript for content, study concept or design. Vitali Ustsinovich: drafting/revising the manuscript for content, study concept or design. Pim Brouwers: drafting/revising the manuscript for content, study concept or design. Howard J. Hoffman: drafting/revising the manuscript for content, study concept or design. Richard Gershon: drafting/revising the manuscript for content, study concept or design.

STUDY FUNDING

This study was 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

J. Beaumont served as a consultant for NorthShore University HealthSystems, FACIT.org, and Georgia Gastroenterology Group PC; and has received funding for travel as an invited speaker at the North American Neuroendocrine Tumor Symposium. R. Havlik reports no disclosures. K. Cook has received financial support from Center for Psychiatric Rehabilitation Boston University, InvivoData, Xenoport, BrightOutcome, 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 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). R. Hays received research funding from the NIA (AG020679-01, P30AG021684, P30-AG028748), NIAMS (UAR057936A, AR052177), NCMHD (2P20MD000182), and the Agency for Healthcare Research and Quality (U18 HS016980). He also received consulting money from Allergan, UBC, the VA, and SciMetrika. K. Wallner-Allen and S. Korper report no disclosures. J.-S. Lai has received research support from the NIH, Agency for Healthcare Research and Quality, and Pfizer, Inc. C. Nord reports no disclosures. N. Zill received research support from the National Council for Adoption, the Brookings Institution, and the Marriage and Religion Research Institute. He served as a reviewer for an NIH SBIR Review Panel. He holds a TIAA-CREF Retirement Annuity Contract that invests in US Treasury Bonds and an International Stock Index Fund. He received consulting income and income from selling stock and exercising stock options from Westat, an employee-owned S Corporation. He holds an IRA and brokerage account with Vanguard that includes holdings in the Vanguard Health Care, Precious Metals and Mining, High-Yield Corporate Bond, High-Yield Tax Exempt Bond, and Long-Term Corporate Bond mutual funds, as well as the Vanguard Consumer Discretionary, Consumer Staples, FTSE International Small Cap, and Corporate Long-Term Bond Exchange Traded Funds. He has holdings in a number of closed-ended mutual funds, including the Aberdeen Asia Pacific Income Fund, Templeton Global High Income Fund, Templeton Dragon Fund, Alliance-Bernstein Global High-Income Fund, Alliance-Bernstein Corporate Income Fund, Blackrock Income Opportunity Trust, Putnam Master Intermediate Income Trust, Nuveen Floating Rate Income Opportunity Fund, India Fund, Morgan Stanley India Investment Fund, Latin American Discovery Fund, Aberdeen Latin American Equity Fund, Singapore Fund, and Turkish Investment Fund. He has stock holdings in AT&T, Boardwalk Pipeline Partners, Banco de Columbia, Consolidated Edison, Exxon, Frontier Communications, Glaxo-Smith Kline, Honeywell, Huntington Ingalls Industries, IBM, Intel, 3M, MeadWestvaco, Newell Rubbermaid, Northrop Grumman, Occidental Petroleum, SCANA, Sherwin-Williams, Siemens, Southern, Verizon, Johnson Controls, American Superconductor, Ocean Power Technology, and Maxwell Technology. Dr. Zill's wife, Karen, prepared discussion guides for the Independent Lens program on PBS. She holds a TIAA Retirement Annuity Contract that invests in US Treasury Bonds. S. Choi has received research support from Boehringer-Ingelheim, Novartis, and the NIH. K. Yost, V. Ustsinovich, P. Brouwers, and H. Hoffman report 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.

EDITOR'S NOTE

NIH Toolbox norming accrual has now concluded. Instrument-level results are available in the technical manuals associated with each instrument. See http://www.nihtoolbox.org/HowDoI/TechnicalManual/Pages/default.aspx.

Received June 6, 2012. Accepted in final form October 18, 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. Rine RM, Roberts D, Corbin BA, et al. New portable tool to screen vestibular and visual function—National Institutes of Health Toolbox initiative. J Rehabil Res Dev. 2012;49(2):209–220.
3. Croker L, Algina J. Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart and Winston; 1986.
4. Pedhazur E, Pedhazur-Schmelkin L. Measurement, Design, and Analysis: An Integrated Approach. Hillsdale: Lawrence Erlbaum Associates; 1991.
5. Census.gov. American Community Survey: 2006–2008. Available at: http://www.census.gov/acs. Accessed September 7, 2011.
6. 2010 Census Summary File 1: United States. In: Bureau USC, ed. 2011.
7. Deming WE, Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann Math Stat 1940;11:427–444.
8. Bishop YMM, Fienberg SE, Holland PW. Discrete Multivariate Analysis: Theory and Practice. Cambridge: MIT Press; 1975.
9. Fienberg SE. The Analysis of Cross-Classified Categorical Data, 2nd ed. Cambridge: MIT Press; 1980.
10. Little RJA, Rubin DB. Statistical Analysis With Missing Data. New York: John Wiley & Sons; 1987.