HHS Public Access

Author manuscript

J Int Neuropsychol Soc. Author manuscript; available in PMC 2020 July 01.

Published in final edited form as:

J Int Neuropsychol Soc. 2020 July ; 26(6): 567–575. doi:10.1017/S1355617720000028.

The NIH Toolbox: Overview of Development for Use with Hispanic Populations

Richard C. Gershon[1,*], **Rina S. Fox**[1,*], **Jennifer J. Manly**[2], **Dan M. Mungas**[3], **Cindy J. Nowinski**[1], **Ellen M. Roney**[1], **Jerry Slotkin**[4]

[1]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[2]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[3]Department of Neurology, University of California, Davis, Davis, CA, USA

[4]The Center for Health Assessment Research and Translation, University of Delaware, Newark, DE, USA

Abstract

Objective: Hispanics/Latinos are the largest and fastest-growing minority population in the United States. To facilitate appropriate outcome assessment of this expanding population, the NIH Toolbox for Assessment of Neurological and Behavioral Function® (NIH Toolbox®) was developed with particular attention paid to the cultural and linguistic needs of English- and Spanish-speaking Hispanics/Latinos.

Methods: A Cultural Working Group ensured that all included measures were appropriate for use with Hispanics/Latinos in both English and Spanish. In addition, a Spanish Language Working Group assessed all English-language NIH Toolbox measures for translatability.

Results: Measures were translated following the Functional Assessment of Chronic Illness Therapy (FACIT) translation methodology for instances where language interpretation could impact scores, or a modified version thereof for more simplified translations. The Spanish versions of the NIH Toolbox Cognition Battery language measures (i.e., Picture Vocabulary Test, Oral Reading Recognition Test) were developed independently of their English counterparts.

Conclusions: The Spanish-language version of the NIH Toolbox provides a much-needed set of tools that can be selected as appropriate to complement existing protocols being conducted with the growing Hispanic/Latino population in the United States.

Keywords

Assessment; Language; Culturally Competent Care; Hispanic Americans

Corresponding Author: Richard C. Gershon, Ph.D.; Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 N. Michigan Ave, 27th Floor, Chicago, IL, USA; Phone: 312-503-3453; gershon@northwestern.edu.

*The first two authors contributed equally to this manuscript.

Introduction

The debate around cultural bias and appropriate assessment with minority populations has a lengthy history (Padilla & Medina, 1996). It is paramount to the effective interpretation of assessment results that researchers consider potential biases when developing and utilizing measures, and use culturally appropriate tests to improve scientific accuracy of research (Bravo, 2003). The NIH Toolbox for Assessment of Neurological and Behavioral Function® (NIH Toolbox®) was developed with particular sensitivity to the growing Hispanic/Latino populations, both Spanish and English speaking (Victorson et al., 2013). This is reflected in both the English and Spanish versions of these measures.

Hispanics/Latinos represented 18% of the United States population as of 2016 (Bureau), and account for approximately half of the nation's population growth since 2000 (A. Flores, 2017). Although English proficiency is increasing among Hispanics/Latinos, nearly three quarters (73%) of Hispanics/Latinos age five and older reported speaking Spanish at home as of 2013, and 12.5 million Hispanics/Latinos in the United States reported speaking English less than "very well" (Krogstad, Stepler, & Lopez, 2015). In addition to spoken language abilities, according to the 2011 Pew Hispanic National Survey of Latinos, approximately three quarters (78%) of Hispanics/Latinos reported being able to read at least "very well" in Spanish, while only 60% of Hispanics/Latinos in the United States reported being able to do so in English (Taylor, Lopez, Martinez, & Velasco, 2012). Additionally, although Hispanics/Latinos represent the fastest growing segment of the United States school-age population, this group often reports both low quantity (Gándara, 2010) and low quality (Gandara & Contreras, 2009) of education, which can negatively impact performance on tests of cognitive ability (Carvalho et al., 2014; Chin, Negash, Xie, Arnold, & Hamilton, 2012; Crowe et al., 2012).

The NIH Toolbox is comprised of a battery of 47 brief measurement tools initially commissioned by the NIH Blueprint for Neuroscience Research, a joint effort of 16 NIH Institutes, to facilitate large-scale data collection in epidemiologic cohort studies and in clinical research (Gershon et al., 2013). Comprised of four core domains: Sensation, Motor, Emotion, and Cognition, all included measures are available at minimal cost and have been normed for use across the lifespan (ages 3-85).

A comprehensive overview detailing the general development of the NIH Toolbox is available from Gershon and colleagues (Gershon et al., 2013). Numerous articles detail the development and validation of the individual NIH Toolbox domains for both adults and children (Coldwell et al., 2013; Cook et al., 2013; Dalton et al., 2013; Dunn et al., 2013; Reuben et al., 2013; Rine et al., 2013; Salsman et al., 2013; Varma, McKean-Cowdin, Vitale, Slotkin, & Hays, 2013; Weintraub et al., 2013; Zecker et al., 2013). Briefly, project development consisted of six phases, involving 1) identification of criteria for included measures (e.g., high interrater reliability, sensitivity to change, applicable to a broad age range (see (Nowinski, Victorson, Debb, & Gershon, 2013) for more information); 2) determination of the subdomains to include in each of the four primary domains (Gershon et al., 2013); 3) identification and/or modification of existing measures, or development of new measures, to meet the selected eligibility criteria; 4) pilot testing and preliminary evaluation

of the psychometric properties of candidate measures; 5) conducting a national norming study (Beaumont et al., 2013); and 6) distribution of the measures for research and potential clinical applications. Ultimately, 47 construct areas within 21 subdomains were identified as important to the comprehensive assessment of the four NIH Toolbox primary domains.

A primary consideration of the NIH Toolbox development process was to ensure the cultural appropriateness of the measures across all ages and major US race and ethnic groups. Separate pediatric, geriatric, cultural, disability, and Spanish language teams worked alongside each of the four domain groups as well as with each instrument development team. The Cultural Working Group ensured that the final NIH Toolbox would be appropriate for Hispanics/Latinos and other ethnocultural groups who prefer to speak, read, and write in English. The Spanish Language Working Group assumed responsibility for the development of a parallel version of the NIH Toolbox for use with those who identify Spanish as their primary language (Gershon et al., 2013).

Ultimately, the NIH Toolbox development effort produced and normed 54 measures in both English and Spanish. This paper details the specific development efforts made to ensure the cultural appropriateness of the English version of the NIH Toolbox for English-speaking Hispanics/Latinos, as well as the procedures followed to produce the Spanish versions of the measures. Both battery-wide and individual test considerations are detailed.

Initial Development

A survey of 150 NIH-funded investigators was conducted to determine initial eligibility criteria for measure inclusion (Nowinski et al., 2013). Criteria discussed were primarily related to psychometric considerations. Applicability to ethnic subgroups and having a Spanish-language version available were respectively rated as “very important” by 69% and 45% of survey respondents. Ethnic and language considerations did not factor into sub-domain selection (e.g., which areas of cognitive function should be considered). While overall NIH Toolbox measure development was described in a series of articles published in a special issue of *Neurology* (Coldwell et al., 2013; Cook et al., 2013; Dalton et al., 2013; Dunn et al., 2013; Reuben et al., 2013; Rine et al., 2013; Salsman et al., 2013; Varma et al., 2013; Weintraub et al., 2013; Zecker et al., 2013), these articles did not detail the dedicated attention to Hispanic/Latino cultural and linguistic considerations reflected in the initial item development phase.

Cultural Considerations

The Cultural Working Group was convened to ensure that all included measures were culturally and conceptually appropriate for use with diverse groups. An overview of the five criteria used to establish cultural competency of the NIH Toolbox measurement tools is described in detail by Victorson and colleagues (Victorson et al., 2013). Briefly, these criteria included 1) incorporating input from culturally diverse end-users into NIH Toolbox development; 2) ensuring conceptual, semantic, and linguistic equivalence across groups; 3) identifying quantitative approaches to ensure psychometric equivalence across groups; 4)

evaluating differential item functioning across groups; and 5) ensuring comparable utility of technical measurement properties, such as Likert-type scales, across groups.

The Cultural Working Group reviewed all English-language NIH Toolbox measures in-depth to identify barriers to cross-cultural validity and to ensure appropriateness for use with Hispanics/Latinos, as it was anticipated that many members of this group would elect to complete the NIH Toolbox in English.

Linguistic Translation, Adaptation, and Validation

The Spanish Language Working Group, comprised of individuals representing different Hispanic/Latino subgroups, was convened to conduct a translatability assessment of all English-language NIH Toolbox measures (Victorson et al., 2013). This group identified potential conceptual or linguistic difficulties in specific wording and offered alternative wordings more suitable for Hispanic/Latino populations that could be more easily and accurately translated. Although different approaches were adopted for each of the domains based on the specific included measures, translatability was generally evaluated according to: 1) universality, 2) cultural relevance, 3) figure of speech/jargon, 4) ambiguity, 5) register, 6) number of words, 7) translation reversal, 8) double-negative, 9) double-barrel (i.e., a question/statement that addresses more than one issue but only allows for one answer), 10) sex and number agreement, 11) parts of speech, 12) oral vs. written, and 13) mode of administration and technology (Victorson et al., 2013).

Language-specific content within the Sensation, Motor, and Cognition Batteries consists primarily of test administrator demonstration and script recitation. As such, these measures were translated following a modified version of the Functional Assessment of Chronic Illness Therapy (FACIT) translation methodology applicable for use in more simplified translations (Bonomi et al., 1996; Cella et al., 1998; Eremenco, Cella, & Arnold, 2005; Lent, Hahn, Eremenco, Webster, & Cella, 1999). This approach included one forward and one backward translation by two different native Spanish speakers. The process began with translation of the English source material into Spanish by one native Spanish speaker. A separate native Spanish speaker subsequently translated this version back into English to enable comparison of the new and original English-language versions. Additionally, a bilingual expert reviewed each translation. In instances where the potential for language interpretation could impact scores, such as with the Emotion measures and a limited number of survey measures from the other domains, a more rigorous translation and cultural adaptation process was used, as described in more detail below. Table 1 provides an overview of the translation methodology used for each NIH Toolbox measure.

Assessment of Sensory Functioning

The Sensation Domain of the NIH Toolbox includes assessments of olfaction, audition, vision, taste, and pain. Specific recommendations related to assessment of sensation made by the Cultural Working Group included placing sensory tests last in the battery to enable the administrator to build rapport with respondents, thus decreasing differential refusal across cultural groups. This was considered especially important for the sensory battery, given the need to interact with less commonly encountered stimuli (e.g., scratch-and-sniff

cards, swab saturated with strong-tasting solutions). Additionally, the Cultural Working Group specified the importance of familiarizing participants with sensory tests using video demonstrations prior to test initiation, to normalize the tests and afford participants an opportunity to refuse to participate after becoming familiar with the protocol. These videos were ultimately not developed due to limited resources. No specific recommendations were made with regard to the assessment of vision. The tests of taste, audition, and olfaction were evaluated for translatability by members of the Spanish Language Working Group. The tests of pain underwent more intensive translation, as described below.

Recommendations regarding other constructs assessed included:

Gustation.—Utilize non-scientific descriptors to identify stimuli (e.g., “sour taste” vs. “citric acid”) to increase the linguistic accessibility of instrument instructions.

Audition.—Use stimuli exclusively in Spanish (e.g., Spanish background noise for the NIH Toolbox Words-in-Noise Test).

Olfaction.—Remove odors that may not be as universally familiar (e.g., peppermint candy was removed). In addition, a pre-screening measure was developed and added for participants ages 3-9 to confirm familiarity with each odorant assessed.

Pain.—Unlike the remainder of the Sensation Battery, the pain intensity and pain interference measures are patient-reported outcome measures, and thus are more subject to respondent interpretation. Therefore, items from these assessments underwent full FACIT translation methodology (see assessment of Emotion below) (Bonomi et al., 1996; Cella et al., 1998; Eremenco et al., 2005; Lent et al., 1999).

Assessment of Motor Functioning

The Motor Domain of the NIH Toolbox includes assessments of endurance, locomotion, strength, dexterity, and balance. The Spanish Language Working Group evaluated early translations of select measures and identified no concerns regarding translatability. Coupled with the low linguistic demand of these measures, the remaining motor assessments were not reviewed. The Cultural Working Group recommendation that instructions for all timed tasks include information regarding both speed and accuracy because certain phrases may be culture-bound (e.g., “as quickly as you can” may not universally convey “as quickly and accurately as you can.”) was deemed not applicable for motor tasks, as these tests are not scored for accuracy.

Assessment of Emotional Well-Being

The Emotion Domain of the NIH Toolbox evaluates four theoretically derived composites – negative affect, social relationships, psychological well-being, and stress and self-efficacy – through 17 scales. Given the potential for language interpretation to impact scores within the Emotion domain more so than in other domains, a more rigorous review of appropriateness across cultures was undertaken. The Cultural Working Group broadly discussed the Emotion domain items as they relate to migration experience effects. For example, immigration can

impact social networks and availability of social support in both positive and negative ways, which could systematically influence responses to the emotion battery. Furthermore, the importance of including culturally relevant examples within items addressing social clubs and recreational groups was reinforced to increase the likelihood of comprehension among Hispanics/Latinos. However, given the depth of validation evidence for the extant Spanish versions of many of these measures, as many were adapted from existent measurement systems such as the Patient-Reported Outcomes Measurement Information System (PROMIS), these recommendations were not followed for the NIH Toolbox.

Items which were previously translated as part of the PROMIS development effort were retained without modification. The remaining emotion items were independently reviewed by at least three members of the Cultural Working Group. These members identified items that posed no cultural problem, those that posed a possible cultural problem requiring discussion, and those that posed a definite cultural problem requiring revision. These ratings were aggregated, with potentially problematic items modified as needed prior to translation. Following the overall translatability review, the emotion self-report and parent proxy report items were translated according to the FACIT translation methodology (Bonomi et al., 1996; Cella et al., 1998; Eremenco et al., 2005; Lent et al., 1999), which is consistent with the guidelines recommended by the International Society for Pharmacoeconomic and Outcomes Research (ISPOR) for translation of patient-reported outcomes instruments (Wild et al., 2005). This approach involves 1) two simultaneous forward translations by natives of the target language; 2) reconciliation of these translations into a single translation, conducted by a third independent translator; 3) back-translation by a native English-speaking translator; 4) comparison of source and back-translated versions to identify discrepancies and facilitate early harmonization; 5) reviews from three bilingual experts; 6) finalization by the language coordinator of the particular target language; 7) harmonization and quality assurance; 8) formatting, typesetting and proofreading; and 9) cognitive pre-testing of translations via interviews with participants from multiple Hispanic/Latino background groups who are native speakers of the target language. Each item was reviewed by at least 5 participants, who first responded to general questions regarding the item and subsequently answered more specific questions designed to ensure that their interpretation of the item text matched the intended English meaning. The acceptability of alternative items was also queried.

Assessment of Cognitive Functioning

The Cognition Battery of the NIH Toolbox evaluates Fluid (attention, executive function, episodic memory, processing speed, working memory) and Crystallized (language) abilities through seven different tests (Heaton et al., 2014; Weintraub et al., 2013).

Fluid abilities.—The fluid ability tests generally minimized the use of language. Auditory stimuli were translated and audio-recorded in Spanish in separate versions, which were culturally appropriate for children versus adults. Instructions were delivered using the informal form of address for children and the formal form of address for adults. The Spanish Language Working Group then reviewed these recordings and either approved or recommended modifications as needed prior to finalization.

Crystallized abilities.—It was recognized early in the development of the NIH Toolbox that language development and usage differed greatly by culture. The acquisition of Spanish-based vocabulary does not match that of English on a word-for-word basis. While English-speaking children and adults may have difficulty pronouncing many words with idiographic spelling, Spanish-speaking first graders can correctly pronounce almost any correctly accented word in the dictionary. The Spanish-speaking versions of these tests were therefore developed independently of their English counterparts. Additionally, to ensure that these tests assessed the same constructs as the English versions, gold-standard Spanish-language measures of crystallized abilities were also administered to enable validation.

NIH Toolbox Picture Vocabulary Test.—This assessment involves auditory presentation of single words via audio file, with concurrent visual presentation of four images of objects, actions, and/or depictions of concepts (Gershon et al., 2014). The respondent must identify the picture that most closely matches the meaning of the spoken word. While respondents are not required to speak to complete the task, they must be able to hear and comprehend auditory stimuli. In the English-language version of the measure, the four response options presented for each item generally reflect a) a synonym (i.e., the correct image); b) an antonym (distractor); c) a look-/sound-alike word (distractor); and d) a close mislead (distractor). Each word reflects a standardized level of difficulty and is associated with a school grade level identified based on English-language education. Given that a single concept may not reflect equal levels of difficulty in English and in Spanish, translation alone is insufficient to yield equivalent assessments across languages. For example, the word “cactus” is likely to be acquired at a later age in English than in Spanish. To address these concerns, a multi-step process involving translation, expert feedback, and item-calibration in Spanish was employed to obtain a Spanish-language test that would be equivalent to its English-language counterpart.

Initially, all items included in the English-language version were translated into Spanish by a native speaker. Linguists/translators then verified the accuracy of the translation vis-à-vis the images, and assessed if terms used could be universally accepted by Spanish speakers from different countries. Six bilingual experts who had knowledge of cognitive processes (e.g., psycholinguists, clinical neuropsychologists) and/or translation, and who represented heterogeneous countries, independently reviewed the translated items. These expert reviewers provided feedback on issues such as age of acquisition, level of difficulty, cultural relevance, connection between the word and the images, and perceived lack of equivalence between the test in Spanish and English. For items where a direct translation of the English word was inappropriate, alternative words were proposed to enable usage of the same images across languages. A Spanish Language Coordinator aggregated all recommendations and proposed a final decision for each word potentially included in the measure. These final translated items were then audio-recorded in a voice appropriate for a wide age range and administered to a Spanish-speaking sample with a broad ability level via an online panel. Item Response Theory (IRT) statistics were calculated to ensure that each item was assigned the appropriate level of difficulty for this language, and to support delivery of the measure in a computer adaptive test format. This calibration process identified a list of over 30 words that remained problematic, which were subsequently qualitatively evaluated for potential

removal. Following review 402 items were included in the item bank, and ultimately the final version yielded 258 Spanish reading items.

NIH Toolbox Oral Reading Recognition Test.—The NIH Toolbox Oral Reading Recognition Test assesses one's ability to recognize and name letters and to properly pronounce individual, printed words out of context. For this test, respondents must read and correctly pronounce letters and words shown one at a time on a screen. The test includes words with irregular orthography and varying complexity of letter-sound relationships, as well as those that are infrequently encountered (Gershon et al., 2014). The Spanish version of the NIH Toolbox Oral Reading Recognition Test underwent de novo development mirroring the same principles of, but distinct from, the development of the English-language versions of this measure. All words included in the Spanish-language version of the measure are presented written in capital letter form without accents to diminish pronunciation cues. Therefore, words for which the meaning is changed by inclusion or exclusion of an accent were not included (e.g., PUBLICO, which could indicate público, publico, or publicó). The test was designed to include words reflecting a wide breadth of reading difficulty to enable assessment of reading levels ranging from very low to very high. Additionally, both irregularly stressed words, usually written with accents, and unambiguously pronounced words, stressed on the last syllable, were included to incorporate a broader range of difficulty. Words were considered irregular when 1) the accent of the word is placed three or more syllables away from the end of the word (e.g., película); 2) the word ends in the letter "n" or "s" and the accent is on the last syllable (e.g., francés); 3) the word ends in the letters "d," "l," "n," or "r" and the accent is not on the last syllable (e.g., difícil); or 4) the word ends in "ia" and the accent is not on the penultimate letter "i" (e.g., divisoria).

To match the inclusion criteria for the English-language version of the measure, words were included with numbers of letters ranging from 2-14 (Gershon et al., 2014). Thirty words per word length were selected from the Corpus del Español (http://www.corpusdelespanol.org/), based on expert linguistic recommendation, to yield an initial set of 390 candidate words. This initial pool was then reduced by two members of the Spanish Language Working Group with expertise in translation, editing, and proofreading by deleting 1) words for which removing the accent would yield another word; 2) words that were only slightly different from other included words (e.g., plurals); 3) words presenting similar irregularities; and 4) words containing the letters "y," "r," or "v," as these letters are often pronounced differently by individuals from distinct regional origins. Specific efforts were made to retain words containing the letter combinations "ca," "ce," "ci," "co," "cu," "ga," "ge," "gi," "go," "gu," "gua," "gue," "gui," "k," "j," "y" (as a semi-vowel), "qu," or "x," as such spelling does not directly correlate with the regular rules of pronunciation in Spanish. In addition, efforts were taken to retain words containing more than one consonant or vowel within a single syllable, and words containing the letter "h" in the middle of the word. The same presentation format used for the English-language version of the measure was used for the Spanish-language version, with one item presented per screen (Gershon et al., 2014). The Spanish Oral Reading Recognition Test was originally pilot tested among a small sample ($N = 50$) of respondents. Final IRT statistics were calculated using the norming sample data to determine difficulty level and to support delivery as a computer adaptive test. Following review, 263

items were included in the item bank, and ultimately the final version yielded 162 Spanish reading items.

Sociodemographic Forms

In addition to reviewing the items evaluating the four primary NIH Toolbox domains, the Cultural Working Group discussed the cultural appropriateness of the sociodemographic forms used in norming. One primary consideration was the need to gather information relating to the level of formal education obtained in each language spoken. The importance of capturing the number of languages spoken in the home was also reviewed, and the impact of social desirability regarding language of study completion among bilingual individuals was discussed. It was recommended that Hispanic/Latino participants be given the opportunity to provide information regarding their national background group. Finally, recommendations for more in-depth assessment of immigration, acculturation, and socioeconomic status were made. For example, the Cultural Working Group recommended that number of years spent living in the United States, and parental country of origin, be evaluated in addition to participant country of origin. They also suggested that information regarding current living environment (e.g., ZIP code) be included in addition to household income to better capture socioeconomic status. To address these considerations, a question was added to the sociodemographic form regarding the number of years of school attended in one's country of origin. Additionally, parental country of origin was assessed for children who participated in the norming study. However, the remaining additional recommendations regarding the sociodemographic forms were not followed in an effort to minimize respondent burden and the length of battery administration.

Norming

Demographically corrected norms have been published for both the English (Casaletto et al., 2015) and Spanish (Casaletto et al., 2016) language versions of the NIH Toolbox Cognition Battery, and the impact of ethnicity and language on performance has been previously explored (I. Flores et al., 2017). Ultimately, 47 instruments were administered to a national sample ranging in age from 3-85 years (N = 4,859), with at least 150 persons included per age band (single-year age bands for children ages 3-17, and multiple-year age bands for adults ages 18-85). Hispanic/Latino participants made up 15.0% of the 2,917 children and 9.6% of the 1,038 adults who took the *English* version of the test battery. Initially, subjects were directed to the Spanish version of the battery if they identified Spanish as the primary language spoken in the home. However, it quickly became apparent that even if a subject was a fluent Spanish speaker, it did not mean that they had Spanish reading proficiency. Further, those Spanish-speaking individuals who preferred reading in English generally preferred to be assessed in English, and were more capable of completing the battery in English. Therefore, the final Spanish sample consisted of those children ($N = 496$) and adults ($N = 408$) who preferred *reading* Spanish.

Specific efforts were made to facilitate recruitment of Spanish-speaking participants for the Spanish version of the NIH Toolbox norming study. The market research firm La Verdad, which specializes in conducting "in-culture" and "in-language" marketing research, was

contracted to provide culture-specific recommendations and guidance. This firm also served as the Cincinnati recruitment site. Additional recruitment strategies specifically targeting the Spanish-speaking population were implemented, including in-person recruitment at community events, recruitment through community organizations/partners, social media advertising, and snowball sampling techniques. Given that less than 2% of Spanish-speaking children in the United States between the ages of 8 and 17 speak Spanish as their dominant language (Census.gov), it was anticipated that very few school-aged children would elect to complete study participation in Spanish versus English. Therefore, only Spanish-speaking children between the ages of 3 and 7, and adults between the ages of 18 and 85, were recruited to create norms in Spanish. However, it is important to note that all measures are still believed to be appropriate for use with Spanish-speaking individuals ages 8-17, despite the lack of language-specific normative data for this age range. Therefore, these measures are still appropriate for use in situations when norms are not needed and raw scores are appropriate, such as tracking an individual's performance over time or comparing an experimental group versus a control group. The NIH Toolbox norming study was approved by the institutional review board at Northwestern University through a protocol that covered all testing sites, and was completed in accordance with the Helsinki Declaration. Written informed consent was obtained from all adult participants. Parental informed consent was obtained from children age 3-7; assent was also obtained from children age 7.

Measure Availability

The NIH Toolbox is now distributed as an administrator-assisted iPad app and is available for download in the Apple App Store. The measures have been cited in more than 200 articles and have been used in more than 130 clinical trials. As of October 2018, the NIH Toolbox app had been licensed for use by more than 900 institutions (and used on as many as 40 iPads at each institution). *NIH Toolbox en Español* is used at 63 of those institutions. Of these, 58 users (92%) were located in the United States, two were located in Spain, and three were within Latin America. This indicates that the NIH Toolbox has been used relatively widely to assess cognitive, emotional, sensory, and/or motor functioning among Spanish-speaking individuals.

Discussion

Certain conditions should be noted when implementing the Spanish-language version of the NIH Toolbox. For example, because the English- and Spanish-language versions of the Picture Vocabulary Test and the Oral Reading Recognition Test were developed as entirely distinct measures, their scores cannot be compared or combined within a single sample. Further, while extensive efforts were taken to ensure the appropriateness of the NIH Toolbox for use with diverse cultures, challenges remain. Measures included in the Cognition Battery that assess reaction time require participants to place their hand in a specific location between trials to better standardize response times across items. However, this concept may be unfamiliar to individuals from various cultural backgrounds, and therefore additional instruction and reinforcement may be required. Finally, while all test administration materials are available for test subjects in Spanish, the instructions for the administrators and the applicable support materials were originally only available in English. Currently, efforts

are underway to provide the entire administrative package in Spanish and to provide instructional materials in Spanish to increase the usability of the NIH Toolbox for monolingual Spanish-speaking investigators.

The English version of the NIH Toolbox was designed to be culturally sensitive to English-speaking Hispanics/Latinos. The Spanish-language version of the NIH Toolbox is comprised of a series of measures designed to assess sensory, motor, emotional, and cognitive functioning. All included measures were thoroughly evaluated for cultural appropriateness with Hispanics/Latinos, among other underrepresented groups, in both English and Spanish. An extensive translation process was undertaken to develop the Spanish-language version, and when translation was impractical or unlikely to yield a high-quality tool, a more rigorous development process was utilized. A forthcoming article will outline the reliability and validity of the NIH Toolbox Spanish measures. Overall, the Spanish-language version of the NIH Toolbox provides a much-needed set of tools that can be selected as appropriate to complement existing research and clinical protocols being conducted with the growing Hispanic/Latino population in the United States.

Acknowledgments

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. The authors report no relevant conflicts of interest to disclose. The authors would like to thank Jennifer Beaumont, Helena Correia, and David Victorson for providing additional details to facilitate the development of this manuscript. Additionally, the authors would like to thank the following NIH Toolbox Domain Chairs for their valuable contributions to the development of the NIH Toolbox: David Cella, Susan Coldwell, Pamela Dalton, Winnie Dunn, Paul Pilkonis, David Reuben, Rose Marie Rine, W. Zev Rymer, Rohit Varma, Sandra Weintraub, & Steven Zecker. Finally, the authors would like to thank the participants of the NIH Toolbox norming study for their important contributions.

References

Beaumont JL, Havlik R, Cook KF, Hays RD, Wallner-Allen K, Korper SP, … Gershon R (2013). Norming plans for the NIH Toolbox. Neurology, 80, S87–S92. doi:10.1212/WNL.0b013e3182872e70 [PubMed: 23479550]

Bonomi AE, Cella DF, Hahn EA, Bjordal K, Sperner-Unterweger B, Gangeri L, … Zittoun R (1996). Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. Quality of Life Research, 5, 309–320. [PubMed: 8763799]

Bravo M (2003). Instrument development: Cultural adaptations for ethnic minority research In Bernal G, Trimble JE, Burlew AK, & Leong FTL (Eds.), Handbook of racial & ethnic minority psychology (pp. 220–236). Thousand Oaks, CA: Sage.

Carvalho JO, Tommet D, Crane PK, Thomas ML, Claxton A, Habeck C, … Romero HR (2014). Deconstructing racial differences: The effects of quality of education and cerebrovascular risk factors. The Journals of Gerontology: Series B, 70(4), 545–556. doi:10.1093/geronb/gbu086

Casaletto KB, Umlauf A, Beaumont J, Gershon R, Slotkin J, Akshoomoff N, & Heaton RK (2015). Demographically Corrected Normative Standards for the English Version of the NIH Toolbox Cognition Battery. Journal of the International Neuropsychological Society, 21, 378–391. doi:10.1017/s1355617715000351 [PubMed: 26030001]

Casaletto KB, Umlauf A, Marquine M, Beaumont JL, Mungas D, Gershon R, … Heaton RK (2016). Demographically Corrected Normative Standards for the Spanish Language Version of the NIH Toolbox Cognition Battery. Journal of the International Neuropsychological Society, 22, 364–374. doi:10.1017/s135561771500137x [PubMed: 26817924]

Cella D, Hernandez L, Bonomi AE, Corona M, Vaquero M, Shiomoto G, & Baez L (1998). Spanish language translation and initial validation of the functional assessment of cancer therapy quality-of-

life instrument. Medical Care, 36, 1407–1418. doi:10.1097/00005650-199809000-00012 [PubMed: 9749663]

Chin AL, Negash S, Xie S, Arnold SE, & Hamilton R (2012). Quality, and not just quantity, of education accounts for differences in psychometric performance between African Americans and White Non-Hispanics with Alzheimer's Disease. Journal of the International Neuropsychological Society, 18(2), 277–285. doi:10.1017/S1355617711001688 [PubMed: 22300593]

Coldwell SE, Mennella JA, Duffy VB, Pelchat ML, Griffith JW, Smutzer G, … Hoffman HJ (2013). Gustation assessment using the NIH Toolbox. Neurology, 80, S20–S24. doi:10.1212/WNL.0b013e3182872e38

Cook KF, Dunn W, Griffith JW, Morrison MT, Tanquary J, Sabata D, … Gershon RC (2013). Pain assessment using the NIH Toolbox. Neurology, 80, S49–S53. doi:10.1212/WNL.0b013e3182872e80 [PubMed: 23479545]

Crowe M, Clay OJ, Martin RC, Howard VJ, Wadley VG, Sawyer P, & Allman RM (2012). Indicators of childhood quality of education in relation to cognitive function in older adulthood. The Journals of Gerontology: Series A, 68(2), 198–204. doi:10.1093/gerona/gls122

Dalton P, Doty RL, Murphy C, Frank R, Hoffman HJ, Maute C, … Slotkin J (2013). Olfactory assessment using the NIH Toolbox. Neurology, 80, S32–S36. doi:10.1212/WNL.0b013e3182872eb4 [PubMed: 23479541]

Dunn W, Griffith JW, Morrison MT, Tanquary J, Sabata D, Victorson D, … Gershon RC (2013). Somatosensation assessment using the NIH Toolbox. Neurology, 80, S41–S44. doi:10.1212/WNL.0b013e3182872c54 [PubMed: 23479543]

Eremenco SL, Cella D, & Arnold BJ (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Evaluation & the Health Professions, 28, 212–232. doi:10.1177/0163278705275342 [PubMed: 15851774]

Flores A (2017). How the U.S. Hispanic population is changing. Retrieved from https://www.pewresearch.org/fact-tank/2017/09/18/how-the-u-s-hispanic-population-is-changing/

Flores I, Casaletto KB, Marquine MJ, Umlauf A, Moore DJ, Mungas D, … Heaton RK (2017). Performance of Hispanics and Non-Hispanic Whites on the NIH Toolbox Cognition Battery: the roles of ethnicity and language backgrounds. The Clinical Neuropsychologist, 31, 783–797. doi:10.1080/13854046.2016.1276216 [PubMed: 28080261]

Gándara P (2010). The Latino education crisis. Educational Leadership, 67(5), 24–30.

Gandara PC, & Contreras F (2009). The Latino education crisis: The consequences of failed social policies. Cambridge, MA: Harvard University Press.

Gershon RC, Cook KF, Mungas D, Manly JJ, Slotkin J, Beaumont JL, & Weintraub S (2014). Language Measures of the NIH Toolbox Cognition Battery. Journal of the International Neuropsychological Society, 20, 642–651. doi:10.1017/s1355617714000411 [PubMed: 24960128]

Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, & Nowinski CJ (2013). NIH Toolbox for Assessment of Neurological and Behavioral Function. Neurology, 80, S2–S6. doi:10.1212/WNL.0b013e3182872e5f [PubMed: 23479538]

Heaton RK, Akshoomoff N, Tulsky D, Mungas D, Weintraub S, Dikmen S, … Gershon R (2014). Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. Journal of the International Neuropsychological Society, 20, 588–598. doi:10.1017/s1355617714000241 [PubMed: 24960398]

Krogstad JM, Stepler R, & Lopez MH (2015). English proficiency on the risk among Latinos: U.S. born driving language changes. Retrieved from https://www.pewhispanic.org/2015/05/12/english-proficiency-on-the-rise-among-latinos/

Lent L, Hahn E, Eremenco S, Webster K, & Cella D (1999). Using cross-cultural input to adapt the Functional Assessment of Chronic Illness Therapy (FACIT) scales. Acta Oncologica, 38(6), 695–702. doi:10.1080/028418699432842 [PubMed: 10522759]

Nowinski CJ, Victorson D, Debb SM, & Gershon RC (2013). Input on NIH Toolbox inclusion criteria: surveying the end-user community. Neurology, 80, S7–S12. doi:10.1212/WNL.0b013e3182872e4c [PubMed: 23479548]

Padilla AM, & Medina A (1996). Cross-cultural sensitivity in assessment: Using tests in culturally appropriate ways In Suzuki LA, Meller PJ, & Ponterotto JG (Eds.), Handbook of multicultural

assessment: Clinical, psychological, and educational applications (pp. 3–28). San Francisco: Jossey-Bass Publishers.

Reuben DB, Magasi S, McCreath HE, Bohannon RW, Wang YC, Bubela DJ, … Gershon RC (2013). Motor assessment using the NIH Toolbox. Neurology, 80, S65–S75. doi:10.1212/WNL.0b013e3182872e01 [PubMed: 23479547]

Rine RM, Schubert MC, Whitney SL, Roberts D, Redfern MS, Musolino MC, … Slotkin J (2013). Vestibular function assessment using the NIH Toolbox. Neurology, 80, S25–S31. doi:10.1212/WNL.0b013e3182872c6a [PubMed: 23479540]

Salsman JM, Butt Z, Pilkonis PA, Cyranowski JM, Zill N, Hendrie HC, … Cella D (2013). Emotion assessment using the NIH Toolbox. Neurology, 80, S76–S86. doi:10.1212/WNL.0b013e3182872e11 [PubMed: 23479549]

Taylor P, Lopez MH, Martinez J, & Velasco G (2012). Language use among Latinos. Retrieved from http://www.pewhispanic.org/2012/04/04/iv-language-use-among-latinos/

Varma R, McKean-Cowdin R, Vitale S, Slotkin J, & Hays RD (2013). Vision assessment using the NIH Toolbox. Neurology, 80, S37–S40. doi:10.1212/WNL.0b013e3182876e0a [PubMed: 23479542]

United States Census. American Community Survey: 2006-2008. Retrieved from: https://www.census.gov/programs-surveys/acs/

United States Census Bureau. QuickFacts Population Estimates. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045216

Victorson D, Manly J, Wallner-Allen K, Fox N, Purnell C, Hendrie H, … Gershon R (2013). Using the NIH Toolbox in special populations: considerations for assessment of pediatric, geriatric, culturally diverse, non-English-speaking, and disabled individuals. Neurology, 80, S13–S19. doi:10.1212/WNL.0b013e3182872e26

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, … Gershon RC (2013). Cognition assessment using the NIH Toolbox. Neurology, 80, S54–S64. doi:10.1212/WNL.0b013e3182872ded [PubMed: 23479546]

Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, & Erikson P (2005). Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in Health, 8, 94–104. doi:10.1111/j.1524-4733.2005.04054.x [PubMed: 15804318]

Zecker SG, Hoffman HJ, Frisina R, Dubno JR, Dhar S, Wallhagen M, Wilson RH (2013). Audition assessment using the NIH Toolbox. Neurology, 80, S45–S48. doi:10.1212/WNL.0b013e3182872dd2

Table 1.

Procedures for translating NIH Toolbox measures

NIH Toolbox measure	Translation Methodology	Audio Recordings included?	Cognitive Debriefing conducted?	Item Calibration conducted?
Sensation Domain				
WIN	Modified	No	No	No
Taste – Instructions	Modified	No	No	No
Visual Acuity – Instructions	Modified	No	No	No
Odor ID – Instructions	Modified	No	No	No
Pain Interference	Full	No	Yes	No
Pain Intensity	Full	No	Yes	No
Motor Domain				
9-hole Pegboard – Instructions	Modified	No	No	No
Grip Strength – Instructions	Modified	No	No	No
Standing Balance – Instructions	Modified	No	No	No
4-meter Walk – Instructions	Modified	No	No	No
2-minute Walk – Instructions	Modified	No	No	No
Emotion Domain				
Emotional Health Items – Self	Full	No	Yes	No
Emotional Health Items – Proxy	Modified	No	Yes	No
Emotional Health – Instructions	Modified	No	No	No
Cognition Domain				
DCCS	Modified	Yes	No	No
Flanker	Modified	Yes	No	No
List Sort	Modified	Yes	No	No
PSM	Modified	Yes	No	No
Pattern Comp	Modified	Yes	No	No
PVT – Instructions	Modified	Yes	No	No
PVT	Unique	Yes	No	Yes
ORRT – Instructions	Modified	No	No	No
ORRT	None	No	Yes	Yes

Note. Modified = Modified FACIT Translation methodology. Full = Full FACIT translation methodology. WIN = Words-In-Noise Test; Odor ID = Odor Identification; DCCS = Dimensional Change Card Sort Test; Flanker = Flanker Inhibitory Control and Attention Test; List Sort = List Sorting Working Memory Test; PSM = Picture Sequence Memory Test; Pattern Comp = Pattern Comparison Processing Speed Test; PVT = Picture Vocabulary Test; ORRT = Oral Reading Recognition Test.