HHS Public Access

Author manuscript

J Int Neuropsychol Soc. Author manuscript; available in PMC 2016 November 13.

Published in final edited form as:

J Int Neuropsychol Soc. 2016 March ; 22(3): 364–374. doi:10.1017/S135561771500137X.

Demographically Corrected Normative Standards for the Spanish Language Version of the NIH Toolbox Cognition Battery

Kaitlin B. Casaletto[1], **Anya Umlauf**[2], **Maria Marquine**[2], **Jennifer L. Beaumont**[3], **Daniel Mungas**[4], **Richard Gershon**[3], **Jerry Slotkin**[3], **Natacha Akshoomoff**[2], and **Robert K. Heaton**[2]

[1]SDSU/UCSD Joint Doctoral Program in Clinical Psychology; San Diego, California

[2]University of California, San Diego, Department of Psychiatry; San Diego, California

[3]Northwestern University, Department of Medical Social Sciences; Chicago, Illinois

[4]University of California, Davis, Department of Neurology; Sacramento, California

Abstract

Objectives—Hispanics are the fastest growing ethnicity in the United States, yet there are limited well-validated neuropsychological tools in Spanish, and an even greater paucity of normative standards representing this population. The Spanish NIH Toolbox Cognition Battery (NIHTB-CB) is a novel neurocognitive screener; however, the original norms were developed combining Spanish- and English-versions of the battery. We developed normative standards for the Spanish NIHTB-CB, fully adjusting for demographic variables and based entirely on a Spanish-speaking sample.

Methods—A total of 408 Spanish-speaking neurologically healthy adults (ages 18–85 years) and 496 children (ages 3–7 years) completed the NIH Toolbox norming project. We developed three types of scores: uncorrected based on the entire Spanish-speaking cohort, age-corrected, and fully demographically corrected (age, education, sex) scores for each of the seven NIHTB-CB tests and three composites (Fluid, Crystallized, Total Composites). Corrected scores were developed using polynomial regression models. Demographic factors demonstrated medium-to-large effects on uncorrected NIHTB-CB scores in a pattern that differed from that observed on the English NIHTB-CB. For example, in Spanish-speaking adults, education was more strongly associated with Fluid scores, but showed the strongest association with Crystallized scores among English-speaking adults.

Results—Demographic factors were no longer associated with fully corrected scores. The original norms were not successful in eliminating demographic effects, overestimating children's performances, and underestimating adults' performances on the Spanish NIHTB-CB.

Conclusions—The disparate pattern of demographic associations on the Spanish *versus* English NIHTB-CB supports the need for distinct normative standards developed separately for each

Correspondence and reprint requests to: Robert K. Heaton, Department of Psychiatry, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093-0603. rheaton@ucsd.edu.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S135561771500137X.

There are no conflicts of interest to report.

population. Fully adjusted scores presented here will aid in more accurately characterizing acquired brain dysfunction among U.S. Spanish-speakers.

Keywords

Neuropsychological test; Norms; Psychometrics; Assessment; Cross-cultural; Cognition

INTRODUCTION

Hispanics living in the United States are the largest ethnic minority, and one of the fastest growing segments of the U.S. population (U.S. Census Bureau, 2014). The majority of U.S. Hispanics are Spanish-speakers, representing 12% of the U.S. population aged 5 and over (Ryan, 2013). The increasing linguistic diversity of the United States is a particular challenge for neuropsychologists in determining the appropriate assessment tools for use in these multicultural contexts.

Accurate identification of neurocognitive impairment is an essential goal of cognitive testing, particularly for research studies aimed at identifying, tracking, and treating underlying brain dysfunction. Cultural and linguistic factors are known to impact cognitive test performance, and thus affect the ability of cognitive tests to accurately identify neurocognitive impairment due to *acquired* brain injury or illness. Even expert translations of cognitive tests do not fully address important cultural and linguistic differences across versions of instruments in various languages (Mungas, Reed, Marshall, & Gonzalez, 2000; Mungas, Reed, Crane, Haan, & Gonzalez, 2004; Mungas, Reed, Haan, & Gonzalez, 2005). In addition to the tests themselves, application of normative standards based on samples that closely resemble the key characteristics of the individual assessed, including language and cultural factors, are critical for accurate classification of neurocognitive impairment. For example, a study contrasting rates of cognitive impairments in a sample of “normal” Spanish-speakers using norms developed for English speakers, showed rates of impairment ranging from 30% to up to 68% among Spanish-speakers with low education (≤6 years). In contrast, impairment rates were close to the expected 16% base rate using norms specifically developed for Spanish-speakers (Cherner et al., 2007).

Recognizing the importance of measuring cognition in Spanish-speaking Americans, the NIH Toolbox for the Assessment of Neurological and Behavioral Function includes a brief (30-min) cognitive battery that is available in both English and Spanish. The NIH Toolbox Cognition Battery (NIHTB-CB) consists of seven tests measuring six neurocognitive domains (i.e., Attention, Executive Functions, Episodic Memory, Processing Speed, Working Memory, and Language; see Weintraub et al., 2013 for details), and three primary composite scores: Fluid, Crystallized, and Total (Heaton et al., 2014). Fluid abilities are flexible thinking skills that change throughout development with rapid gains in childhood that peak in early adulthood and decline with increasing age (Cattell, 1971); as with measures sensitive to the aging process, these are also the measures most sensitive to acquired brain injuries, and include episodic memory, processing speed, and executive functions (Cattell, 1971).

On the other hand, Crystallized skills are learned semantic knowledge (e.g., reading and vocabulary) and largely reflect one's educational, cultural, and life experiences. These latter abilities follow a different developmental trajectory with large gains in childhood that tend to stabilize in early adulthood, and do not tend to change with acquired brain injury/disease. The NIHTB-CB Total Composite reflects the average of one's Fluid and Crystallized abilities and may be viewed as indexing an individual's overall level of cognition, similar to an IQ score. The primary goal of the NIH Toolbox initiative was to develop assessment tools for clinical researchers using a common metric for cross-study comparisons. As such, the NIHTB-CB was not conceptualized as a substitute for in-depth, comprehensive neuropsychological batteries, or as a neurodiagnostic tool (Gershon et al., 2013). Therefore, although the NIHTB-CB may have potential use as a brief clinical neurocognitive screener to help identify individuals appropriate for referral for comprehensive neurological assessments, validation is still needed to determine its clinical utility, and it is most appropriately and well-position to be used in epidemiological and longitudinal clinical research at this time.

The original normative standards for the NIHTB-CB adjusted for age, education (or mother's education for children), sex, and race/ethnicity; yet, these corrections were calculated collapsing together participants tested in Spanish and English, as well as combining across children and adults. Combining Spanish- and English-speakers for normative standards is potentially problematic due to the distinct cultural effects and other background characteristics (e.g., socioeconomic status, quality of education) specific to Spanish speakers in the United States, as well as the differential relationships of these characteristics with neuropsychological test performances across language groups (Benson, de Felipe, Xiaodong, & Sano, 2014; Stricks, Pittman, Jacobs, Sano, & Stern, 1998). These distinct relationships cannot be adequately captured and accounted for when collapsing across language groups. Similarly, combining children and adults into one cohort for the development of norms may be problematic because there are different demographic relationships with neuropsychological test performance in children *versus* adults (e.g., education; Heaton, Miller, Taylor, & Grant, 2004).

The goal of the present study was to address these problems with the original normative standards of the NIHTB-CB. We analyzed raw data from the original norming cohort and created standardized scores adjusted for demographics (age, education, sex) separately for English and Spanish speakers, as well as children and adults. Here we present the development of normative standards for the Spanish-speaking children and adults. The normative standards for English-speakers, which were also developed independently per racial/ethnic group, are presented in a separate study (Casaletto et al., 2015).

METHODS

Participants

Participants were neurologically healthy, community-dwelling children and adults who elected to be evaluated in Spanish (Beaumont et al., 2013). Participants were recruited from 10 U.S. testing sites through online self-enrollment, enrollment events, and random telephone calls by market research companies, Delve, La Verdad, and Facts 'n Figures. A

stratified recruitment plan by the NIH Toolbox outlining the norming plans is available (Beaumont et al., 2013). All participants had adequate visual, auditory, vestibular, and motor functioning to complete all items on the Spanish version of the Toolbox test battery, or availability of assistance/assist devices to complete tasks, and were able to provide informed consent, or assent (i.e., children ages ≥8 years) accompanied by parental informed consent. Trained research personnel completed structured interviews and administered questionnaires to determine eligibility (see Casaletto et al., 2015, for more details). All data were gathered *via* self-reported paper-based questionnaires, interviewer-administered questionnaires, or PC-based objective assessments. This project was conducted in accordance to the Helsinki Declaration; written informed consent was obtained from all participants *via* a protocol that covered all testing sites approved by the institutional review board at Northwestern University.

Given the lack of U.S. school-aged children who would elect to be evaluated in Spanish (i.e., fewer than 2.5% of *Spanish-speaking* children ages 8–17 years speak Spanish as their primary language), the cost of recruiting such a low probability cohort, and the poor representation of such a normative cohort (e.g., not feasible through a random sampling and high variability with poor generalizability of this cohort), the normative study for Spanish-speakers prospectively only recruited those children ages 3–7 years old. For guidance on interpreting NIHTB-CB scores for Spanish speakers ages 8–17 years old, see the Discussion section. The entire NIH Toolbox Cognition Battery (NIHTB-CB) norming sample was comprised of a total of 3413 children and 1446 adults; of those, 496 children (ages 3–7 years) and 408 adults (ages 18–85 years) were administered the Spanish version of the NIHTB-CB and were included in the current study. Participants also completed self-report measures identifying race and ethnicity, age, sex, years of education, and background language information (this latter information was not gathered for children). For children, mother's education was used a proxy for the child's education. For the full demographically corrected standards, we determined that performances on the NIHTB-CB (age-corrected scores) were comparable across the Spanish-speaking races/ethnicity groups (ps > .05). Due to the small sample sizes within the races/ethnicities (other than Hispanic White) and comparability of performances, we chose to include all races/ethnicities in the final fully corrected norming parameters to be as representative as possible of the Spanish-speaking U.S. population (see Table 1)

Additionally, among the children tested in Spanish, 46.7% (n = 200) demonstrated a raw theta Oral Reading score of −11.04 or −10.98, indicating performances on the floor of the test (i.e., lowest possible scores). Given that almost half of the children scored the lowest possible scores, the raw Oral Reading data were so significantly skewed that achieving normality for normative standards was not possible when including this pre-reading subgroup of low scorers; furthermore, we found that, indeed, inclusion of these children who scored on the floor of the Oral Reading test would result in insensitivity to performances above the floor on this measure (i.e., all performances above the lowest possible scores would be classified as "normal"). Therefore, we elected to exclude children who demonstrated Oral Reading performances (thetas) ≤ − 10.98 from the development of the Oral Reading normative formula. Notably, children who tested on the floor of the Oral Reading test were significantly younger (4.2 *vs.* 5.8 years; t = 14.8; p < .001) and more

likely to be male (52.5 *vs.* 43.0% male; $\chi^2 = 3.9$; $p = .049$), but did not differ regarding years of mother's education (9.3 *vs.* 9.9; range, 0–20 years; $t = 1.4$; $p = .15$) compared to those with scores > − 10.98. Also, there were no significant differences between the pre-reading *versus* reading children on the Fluid Composite (Fluid fully corrected T-scores: 49.4 *vs.* 50.8; t = 0.95; $p = .34$) or Vocabulary test (fully corrected T-score: 49.1 *vs.* 50.9; $p = .052$). Of interest, for comparison purposes, no such floor effects were observed on raw Oral Reading scores for children in the same age group tested in English [i.e., $n = 4$ (<1%) scored two lowest possible values].

NIH Toolbox Cognition Battery (NIHTB-CB) measures—The NIHTB-CB is a 30-min computerized battery that includes seven measures and assesses six cognitive domains. See Weintraub et al. (2013) for more detailed descriptions of each measure and Heaton et al. (2014) for validation of the Composite Scores. In brief, the battery is comprised of five tests of Fluid abilities (Dimensional Change Card Sort, Flanker Inhibitory Control and Attention Test, Picture Sequence Memory Test, List Sorting Test, and Pattern Comparison Test) and two Crystallized (Picture Vocabulary and Oral Reading Recognition) measures. All Spanish Fluid measures were translated from the English version using a modified version of the FACIT translation methodology (Bonomi et al., 1996; Eremenco, Cella, & Arnold, 2005), which included: (1) one forward translation by a native Spanish-speaker; (2) back-translation by a native English-speaking translator; (3) comparison of source and back-translated versions to identify possible discrepancies and facilitate early harmonization; (4) reviews by one bilingual expert; (5) finalization by the Spanish language coordinator; (6) harmonization and quality assurance; and (7) formatting, typesetting, proofreading, and audio-proofing of translated materials.

The instructions for the Spanish Crystallized measures were similarly translated using the modified FACIT translation approach. However, the Crystallized (language tests) stimuli required distinct development approaches. The translation procedure for the Picture Vocabulary Test included: (1) one forward translation by a native Spanish speaker; (2) two reviews by bilingual experts for language relevance; (3) review for modifications of the item prompts (images presented), as needed; (4) finalization of each word based on all feedback by the language coordinator; and (5) development of audio-recorded prompts in a voice appropriate for a wide age range of the target audience.

On the Oral Reading Test, English translation was not applicable and Spanish-specific items were independently developed as follows: (1) identify a corpus of words that span a broad range of difficulty, frequency, and regularity/irregularity in Spanish; and (2) expert review with attention to linguistics education, and culture. After the initial development, both the Spanish Picture Vocabulary test were pre-tested in a wide ability sample ($N = 1329$) and item response theory (IRT) statistics were calculated to determine level of difficulty and appropriateness of each item in Spanish and to create the computer adaptive test format. The Spanish Oral Reading test was initially piloted in a sample of $N = 50$ and final calibrations for the computer adaptive test were derived based on the norming sample data ($N = 904$).

Language and background measures—Participants self-reported first language learned (English, Spanish, or some other language), and which language they mainly speak

at home (English, Spanish, English, and Spanish equally, or some other language). Participants also rated, separately, how frequently they use English and Spanish in everyday life using the following response options: 1 (none), 2 (rarely), 3 (often), 4 (every day). They also indicated if they went to school in the United States and if they were born outside of the United States.

Data Analyses

Uncorrected normalized standard scores—To create uncorrected scores that could be compared across individuals and tests on the same metric, we developed normally distributed standard scores based on the entire Spanish-speaking cohort (i.e., across both children and adults). Raw scores for each individual NIHTB-CB measure were converted into sample-based normalized standard scores ($M = 100$; $SD = 15$). These uncorrected scores represent an individual's performance compared to our Spanish-speaking normative sample (i.e., how far an individual deviates from the average in the cohort). They may be most useful when evaluating individuals longitudinally to determine absolute levels of neurocognitive change and/or in guiding everyday functioning recommendations (wherein absolute levels of cognitive capacity are important).

Age-corrected standard score derivation—Age-corrected scores were calculated for adults and children separately using the statistical software R (www.r-project.org) and R package mfp (Ambler & Benner, 2008). Raw test scores were converted to normalized scaled scores based on their standardized quantiles ($M = 10$; $SD = 3$). The normalized scaled scores were then regressed on age, using fractional polynomials. Fractional polynomials allow fitting non-linear terms, only if they explain variability in the outcome significantly better ($p < .05$) than a simple linear pattern (Royston & Altman, 1994). The algorithm could choose between 36 linear combinations of power transformations (e.g., $X^{-0.5}$, $\log_{(x)}$, $X^{0.5}$). The uncorrected residuals from the regression equations were then obtained, which represent the difference between the actual observed score and the expected scaled score for that individual's age.

Importantly, the residuals may have different spreads (variances) across age groups for various reasons (e.g., differences in range of ability levels, random chance). To create homogeneity of the variances, the residuals within each age group were adjusted for how far, on average, they fell from the expected values. Multiple fractional polynomials were used in this procedure to regress the absolute values of the residuals on age. The resulting curves estimated the smoothed, absolute average distance of the residuals across each age group. Finally, the original uncorrected residuals within each age group were divided by the smoothed mean distance estimated for that age group. This process brought residuals that were large (larger average distance) closer to the mean, while those that were small (small average distance) were extended further, so that on average, residuals for the whole sample had approximately equal variances across age (i.e., achieved homogeneity of variances across age).

The standardized, corrected residuals formed age-adjusted standard scores ($M = 100$; $SD = 15$). Age-corrected scores represent individuals' neurocognitive abilities as compared to

their developmentally matched peers, and may, therefore, be most useful in determining performances expected for one's age (e.g., school or work settings), or when comparing against other age-only adjusted performances on other cognitive instruments (e.g., IQ scores).

Fully demographically corrected T-score derivation—A standard T-score metric (M = 50; SD = 10) was chosen for all fully corrected values to clearly distinguish these scores from the uncorrected and age-corrected scores, and because these scores will be most applicable in a neuropsychological context in which T-scores are a commonly used metric.

The normative standards for the fully corrected scores were developed in adults and children, separately. Using R (www.r-project.org) and R package *mfp* (Ambler & Benner, 2008), raw test values for the normative groups were converted to normalized scores by obtaining their standardized quantiles and scaling them to have a mean of 10 and standard deviation of 3 (see Appendices 1 and 2). Scaled scores were regressed on the demographic characteristics, including age, education, and sex, using fractional polynomials. The residuals for each of the normative groups were obtained and adjusted to achieve homogeneity of the variances across all demographic characteristics, using the smoothing methods described in the previous section. Standardized corrected residuals formed demographically adjusted T-scores (M = 50 and SD = 10). These fully adjusted T-scores represent an individual's neurocognitive functioning relative to his/her age, education, and sex; these scores are most useful in determining decline from "expected" levels of performance due to acquired neurological injury or illness.

NIHTB-CB Composite Score creation—NIHTB-CB Composite Scores were developed separately for uncorrected, age-corrected, and fully demographically corrected scores. The following tests (unadjusted, normalized scores) were averaged for each Composite and then demographically adjusted using the fractional polynomial regression methods described above: (1) *Fluid Composite:* average of Flanker Inhibitory Control and Attention Test, Picture Sequence Memory Test, List Sorting Test, Pattern Comparison Test, and Dimensional Change Card Sort Test (DCCS); (2) *Crystallized Composite:* average of Oral Reading and Picture Vocabulary; and (3) *Total Cognition Composite:* average of the Fluid and Crystallized Composites. For all Composites, a score was only calculated if the individual completed all measures within the Composite.

See Appendices 1 and 2 for conversion of raw NIHTB-CB scores into standard scores and all normative formulas.

NIHTB-CB composite score "impairment" cut-point—A one standard deviation cut-point below the mean (T < 40) was chosen to classify "impairment" across the Fluid, Crystallized, and Total Composite scores (Taylor & Heaton, 2001). According to the normal curve, we expect approximately 84% specificity within a neurological healthy population (i.e., 16% "impairment").

Original versus new NIHTB-CB normative standards—Lastly, we compared the fully corrected, newly created scores to those fully corrected ones previously posted online

for the Spanish NIHTB-CB by calculating the absolute value of the difference between the new and original scores across each test and the Composite scores (e.g., |Original DCCS scores – New DCCS score|). We also explored any residual significant demographic effects on the original, fully corrected scores using correlational or analysis of variance analyses, as appropriate.

RESULTS

Normalized Uncorrected Standard Scores

Appendix 2 presents the formulas used to convert raw scores to uncorrected, normally distributed scores based on the current cohort of Spanish-speaking children and adults. The summary demographics across our entire sample (children and adults) were: 22.6 years old (*SD* = 22.5), 58.1% female, 10.2 years of education (*SD* = 4.1), 82.3% Hispanic White, 12.2% Native American, 3.1% Hispanic African American, 0.2% Asian, 0.2% Pacific Islander, and 0.1% Non-Hispanic White. The average performance across each test and Composite score demonstrated *M*s = 100 and *SD*s = 15. Adults generally fell in the above average range [Fluid *M* = 112.1 (*SD* = 8.8); Crystallized *M* = 113.9, (*SD* = 6.8)], while children had low average scores [Fluid *M* = 92.3 (*SD* = 9.9); Crystal *M* = 87.2 (*SD* = 7.0)] across the Composites, as expected.

Impact of Demographic Characteristics on the Spanish NIHTB-CB

Age effects—The uncorrected scores (broken lines) presented in Figure 1 demonstrate the significant impact of age on Fluid and Crystallized abilities. Rapid developmental gains of both Fluid (r = 0.76) and Crystallized (r = 0.69) skills were observed in early childhood (see Table 2). However, while Fluid abilities peaked in early adulthood (18–29 years old) and steadily declined with age (Fluid Composite adults, r = −0.50), Crystallized skills peaked slightly later (ages 40–49 years) and demonstrated a slight, but significant decline in older age (Crystallized Composite: ages 40–85 years; r = −0.33; $p < .001$). The Flanker test (attention) was one of the individual measures most strongly associated with age in both children (r = 0.67) and adults (r = −0.53). Picture Sequence Memory also demonstrated strong development effects in children (r = 0.70), and Pattern Comparison (processing speed) showed strong age-related decline in adults (r = −0.50).

The solid lines in Figure 1 illustrate Fluid and Crystallized scores corrected for such age effects; in both children and adults, age was not significantly related to any of the age-corrected NIHTB-CB measures or Composite scores (ps > .20).

Education effects—Among children and adults, education was positively associated with age-corrected NIHTB-CB scores (see left panel of Figure 2 and Table 2). In adults, education demonstrated strong, positive relationships with both Fluid (r = 0.53) and Crystallized (r = 0.31) abilities, with DCCS (executive functions; r = 0.43) and Flanker (attention; r = 0.43) showing the strongest educational effects among the individual measures. Children demonstrated positive, but smaller relationships between NIHTB-CB performances and mother's education, as expected. Overall, Fluid abilities showed the strongest relationship with mothers' education in children (r = 0.23); the latter was

especially driven by DCCS (executive functions; r = 0.17) and List Sorting (working memory; r = 0.15) abilities. On the other hand, mother's education was only modestly and nonsignificantly related to children's Crystallized abilities (r = 0.10).

Sex effects—Although more variable, sex demonstrated some significant and generally small-to-medium effect sizes with NIHTB-CB performances, the pattern of which differed between adults and children (Table 2). For example, in adults, males performed significantly better than females on DCCS and Flanker tests (d's = 0.27–0.31); while in children, females performed better than males only on the Picture Vocabulary test (d = 0.31).

Fully Demographically Corrected NIHTB-CB T-Scores

Demographically adjusted T-scores demonstrated $M = 50$ and $SD = 10$ for all measures and Composite scores. There were no residual significant associations observed between the corrected scores and any demographic factors. Using a T < 40 (−1 *SD*) cut-point, impairment rates across individual test measures ranged from 12.0 to 17.1% in children, and 14.0 to 18.5% in adults. Similarly, Composite indices indicated 16.0–16.6% impairment in children, and 15.3–17.0% impairment in adults (Figure 3).

Associations with language background factors in adults—Although our fully corrected normative standards were created exclusively within a U.S. Spanish-speaking cohort, there was a fair degree of variability regarding adults' language backgrounds. We aimed to explore how these factors may influence NIHTB-CB performances. Not surprisingly, and supporting the validity of the NIHTB-CB language measures, performances on the Crystallized composite were associated with all language background factors, with indicators of increased Spanish-speaking frequency and Spanish exposure positively associated with fully corrected Crystallized scores. Those who reported speaking Spanish at home (*vs.* English, d = 0.50), Spanish as their first language (d = 1.4), and were educated (d = 0.45) or born (d = 0.83) outside of the United States performed better on the Crystallized composite (*p*s < .001).

Of note, some language background factors were also associated with the corrected Fluid composite, but in the opposite pattern. Language spoken at home was associated with Fluid performances ($p = .001$), such that individuals who reported speaking both Spanish and English at home performed better than those who only spoke Spanish at home (d = 0.55). Additionally, those who completed some school in the United States (d = 0.37; $p < .01$) and/or were born in the United States (d = 0.29; $p = .04$) performed *better* on the fully corrected Fluid Composite. This pattern of associations was comparable on the uncorrected Fluid and Crystallized indices as well.

Fully Corrected NIHTB-CB Scores: Norms Originally Posted Online *versus* New Norms

Although the originally created fully corrected NIHTB-CB Composite scores demonstrated strong associations with the newly created ones in Spanish-speaking children (r = 0.83–0.93) and adults (r = 0.87–0.89), there were several distinct differences. Using the absolute value of the difference between the original and new fully corrected scores in adults, there was an average of 4.3 ($SD = 2.2$; range = 0.9–15.5) T-score point difference across the individual

measures. On the Composite scores, the differences averaged between 4.5 and 5.6 T-score points (range, 0.01–26.5). In children, these score differences were comparably disparate with an average of 4.7 (*SD* = 1.6; range = 0.2–10.4) T-score point difference across the individual measures, and 4.2–10.1 T-score point average difference across the Composites (range, 0.07–22.5).

Importantly, many associations still remained between demographic factors and the original fully demographically corrected NIHTB-CB scores. Among adults, age demonstrated significant *positive* associations with the original fully corrected Pattern Comparison scores (r = 0.14; $p = .004$), suggesting an over-correction, and was still negatively associated with the Flanker test (r = −0.20; $p = .002$) and the Crystallized Composite (r = −0.24; $p < .001$), suggesting an under-correction. Education was significantly and positively associated with *all* of the original fully corrected individual measures (r's = 0.18 to 0.45; ps < .002) and Composite scores (r's = 0.21 to 0.42; ps < .001). Additionally, there were significant sex effects on the originally created fully corrected Pattern Comparison scores that did not exist on the uncorrected scores (M > F; $F(1,344) = 7.5$; d = 0.33; $p = .007$).

Similarly, in children, age was still significantly, positively associated with Pattern Comparison (r = 0.18; $p = .002$) and Reading (r = 0.67; $n = 195$ post-literate children, $p < .001$) and now negatively associated with List Sort (r = −0.19; $p = .002$). Significant, negative relationships between mother's education and children's Vocabulary (r = − 0.14; $p = .008$) and overall Crystallized Composite (r = −0.17; p = .002) performances were also observed, again representing possible over-corrections. Finally, sex demonstrated significant effects on the original fully corrected scores, with females showing better scores on the List Sort ($F(1,283) = 7.1$; d = 0.32; $p < .01$), Vocabulary ($F(1,414) = 4.3$; d = 0.19 $p = .04$), and Crystallized ($F(1,358) = 34.8$; d = 0.63; $p < .001$), and Total composites ($F(1,179) = 23.0$; d = 0.72; $p < .001$), but worse scores on the Picture Sequence Memory (F(1,293) = 31.2; d = 0.65; $p < .001$).

When examining impairment rates, the original scores appeared to significantly overestimate children's performances and underestimate adult's performances on the Spanish NIHTB-CB Composite scores compared to the new scores (ps < .001; see Figure 4), ranging from 1.1% to 4.5% among children and 26.7% to 31.4% among adults.

DISCUSSION

Given the relatively limited selection of well-validated neuropsychological tools available in Spanish and the rate of growth of the Hispanic population as one of the fastest in the United States (U.S. Census Bureau, 2014), the Spanish version of the NIH Toolbox Cognition Battery (NIHTB-CB) is a particularly needed assessment tool to help characterize neurocognitive functioning in this increasing population of Spanish-speakers. There are several important cultural and background factors associated with speaking Spanish in the United States that are known to be associated with neuropsychological test performance (e.g., acculturation, place of birth and education, quality of education; Benson et al., 2014; Mungas et al., 2005; Stricks et al., 1998). As such, the normative standards for the Spanish NIHTB-CB presented here were exclusively developed within a Spanish-speaking normative

cohort, which is an important distinction from the norms originally created on this battery (and currently available online) that were developed combining both Spanish and English speakers.

As expected, the uncorrected scores demonstrated significant relationships with all demographic factors in both the Spanish-speaking children and adults. In adults, age demonstrated strong, negative associations with Fluid abilities, and a later developmental peak was observed on Crystallized abilities (ages 40–49 years). Of interest, following this peak, Crystallized abilities then showed small but negative associations with age in older Spanish-speaking adults (ages 40–85 years). This latter pattern of results is distinct from that observed among the adults tested in English on the NIHTB-CB (Casaletto et al., 2015) in which a small, but *positive* linear association was observed with increasing age and Crystallized performances.

Additionally, although there were consistent positive, medium associations between education and NIHTB-CB performances in the Spanish-speaking adults, the strength of these associations with Fluid and Crystallized abilities differed in the Spanish- *versus* the English-speaking adults. Education showed a stronger relationship with Fluid abilities in the Spanish-speaking cohort, but a stronger relationship with Crystallized abilities in the English cohort (Spanish: Fluid r = 0.53; Crystallized r = 0.31; *vs*. English: Fluid r = 0.21; Crystallized r = 0.41).

Lastly, although the preferential effect of female sex on episodic memory has been consistently reported in English-speaking adults (e.g., d = 33 Picture Sequence Memory; Casaletto et al., 2015), this was not observed in our adult Spanish-speaking cohort (d = 0.13). These differential patterns point to the complex relationships between demographics and neuropsychological test performances that may be unique to U.S. Spanish-speakers. Given the heterogeneity of language backgrounds in our U.S.-based Spanish-speaking cohort compared to the English cohort, it may not be surprising that these different patterns of demographic associations were observed, especially on the Crystallized (language) tests. Specifically, differing levels of Spanish fluency may increase the performance variability on the Spanish language tests, resulting in potentially smaller associations with other demographic factors than was observed in the English version.

In fact, we did find that several language background factors were associated with corrected NIHTB-CB test performances, such that greater Spanish use frequency and exposure to Spanish was associated with *better* Crystallized performances (Spanish vocabulary and reading), but *poorer* Fluid scores. It is important to note that in the current study, all participants were community-dwelling, meaning that they resided in the United States and, therefore, likely accurately represent the background heterogeneity of Spanish speakers in the United States. Nonetheless, given their persisting effects, such background and acculturation factors should be taken into consideration during test score interpretation. Specifically, among individuals with lower U.S. acculturation factors (e.g., born/educated in Mexico), it may be important to consider that Crystallized scores may be *higher* than expected for age/education whereas Fluid scores be *lower* than expected for age/education. Acknowledging the influential role that language familiarity and acculturation play on

interpretation of even linguistically and demographically adjusted scores both on the NIHTB-CB and in other multicultural neuropsychological contexts is an important ethical role of a neuropsychologist.

Among Spanish-speaking children (ages 3–7 years), significant associations between demographics and uncorrected NIHTB-CB scores were observed, but again, somewhat different patterns emerged when compared to children tested in English. Specifically, age demonstrated strong, positive associations with all NIHTB-CB scores among the children tested in Spanish, but these relationships were slightly smaller than those observed in children from the English cohort (ages 3–7 years: Spanish range r = 0.50 to r = 0.70 *vs.* English range r = 0.49 to r = 0.84). Additionally, there was a significant, medium sex effect on the Vocabulary measure among Spanish-speaking children favoring females, whereas in children tested in English, males tended to show slightly better Crystallized abilities (Vocabulary Spanish F > M Cohen's d = 0.31 *vs.* English M > F Cohen's d = 0.12).

Of interest, regarding Crystallized abilities, almost half of the Spanish-speaking child cohort evidenced Oral Reading skills on the floor of the test (i.e., appeared to be pre-literation) and these Crystallized abilities did not appear to be importantly related to mother's education (r = 0.10; comparable to r = 0.06 relationship between mother's education and Crystallized performances among children ages 3–7 years tested in English). These findings may speak to the heterogeneity of language skills in primarily Spanish-speaking children in the United States, as well as culture-specific effects in language development and introduction to reading. For example, Spanish reading differs in important ways to reading in English. In Spanish, reading words correctly is based on knowledge of which syllable the intonation should be placed, as well as standard language rules (e.g., use of the "tilde") in addition to exposure to the words; the latter of which is important on English reading tests which include irregularly spelled words.

Additionally, there may be other multicultural factors specific to Spanish language development in the United States, such as reduced availability of Spanish language books for young children. For instance, a recent study demonstrated that although Latino children (tested in English and Spanish) exhibit comparable oral language abilities relative to their non-Hispanic White peers at 9 months old, these abilities began to significantly lag behind at 2 years of age (Fuller, Bein, Kim, Rabe-Hesketh, 2015). Importantly, language development among Latino children was associated with higher socioeconomic status (i.e., living above or below the poverty line) and more frequent learning activities (i.e., mother reading to child), factors that were less frequently observed among foreign-born Mexican American mothers. Therefore, although these factors were not assessed in the current study, there appear to be a variety of important acculturation and socioeconomic factors that may be contributing to the language performances of young Hispanic children, that are both more diverse and distinct from those children tested in English.

Taken together, the significant impact of demographics on neuropsychological test performances, and, especially, the unique pattern of these associations observed specifically within the Spanish-speaking adults and children, highlight the importance and need for normative standards distinctly developed for this cohort of individuals. Conceivably, even

more “full” demographic normative corrections for Hispanics in the United States might also include some indicators of language use and acculturation since these appear related to performances on Fluid measures typically used to detect and characterize acquired brain dysfunction.

For Spanish-speakers, in our newly presented norms, impairment rates showed appropriate levels of specificity at 1 *SD* below the mean for neurologically healthy individuals, ranging from 83.0 to 84.7% and 83.4 to 84.0% on the NIHTB-CB Composite measures among adults and children, respectively; these values are commensurate with the expected levels of specificity at −1 *SD* given a normal distribution. On the other hand, when compared to our newly developed fully corrected norms, the original norms demonstrated large overestimations of normal children’s performances (only 1–3% “impairment” rates) and underestimations of normal adults’ performances (27–31% “impairment” rates; Figure 4). In other words, application of the original norms would over-classify impairment when applied to Spanish-speaking children, but under-classify impairment when applied to Spanish-speaking adults. In addition, these original norms also maintained significant associations with age, sex, and especially education in both the Spanish-speaking children and adults, indicating that these relationships were not fully accounted for by the original norming method.

Still, there are several important limitations to consider when applying the newly created Spanish normative standards. First, individuals included in our normative cohort elected to be tested in Spanish, but were not objectively assessed for English *versus* Spanish language proficiencies (e.g., no *a priori* verbal fluency testing). As a result, the current cohort may represent individuals with fairly heterogenous levels of Spanish proficiency and bilingualism, which can impact NIHTB-CB test performances. However, as indicated in Table 1, the vast majority of adults indicated that they learned Spanish as their first language (94.2%) and that they speak at least some Spanish at home (87.2%). Additionally, this cohort represents those who indicated a preference to be tested in Spanish, which is likely representative of how the Spanish NIHTB-CB will be applied in real-life research and clinical settings (i.e., self-identify a Spanish preference), and is representative of the true heterogeneity of language proficiency among U.S. Spanish-speakers; both of which increase the ecological validity of the currently developed norms.

Importantly, we also lacked information regarding significant background and cultural factors (e.g., immigration status, acculturation, country of origin), especially among the Spanish-speaking children (e.g., preschool status, mother’s location of education), that could play an important role in understanding test performances in this cohort. In particular, undetected speech or developmental disorders may have impacted language (e.g., reading), as well as fluid performances. More in-depth and systematic investigations into how these and other factors that may impact NIHTB-CB performances are needed.

Additionally, we also do not have data representing Spanish-speaking children ages 8–17 years, which may particularly limit longitudinal lifespan analyses on the Spanish NIHTB-CB; however, given the scarcity of such school-aged monolingual Spanish-speaking children in the United States (<2.5% that would elect to be tested in Spanish), such individuals are

not very representative of Spanish speakers in America and, therefore, would likely be an infrequent issue. Ideally, a normative cohort of Spanish-speakers ages 8–17 years would include a range of variability in acculturation, language, and other cultural background factors that are known to impact neurocognitive performance. Indeed, future normative work focused on this cohort may benefit from potentially correcting for such important background factors.

For the purposes of the current NIHTB-CB, however, we provide several recommendations in guiding researchers and clinicians around the multicultural use of the NIHTB-CB in children 8–17 years old. First, among all individuals with diverse or multicultural backgrounds, we recommend a careful evaluation of language familiarity, past and current language use (e.g., first language learned, language spoken with peers *vs.* family), language preference, and educational (e.g., country completed education, years completed and quality) and other acculturation (e.g., country born) factors. Consideration of these background factors during interpretation of neuropsychological data is critical given the body of literature, including the current study, demonstrating their significant impact on cognitive test performances (i.e., greater U.S. acculturation factors, better fluid cognition performances).

Second, should a researcher or clinician be presented with a *bilingual* (English/Spanish) child between ages 8 and 17 years, we recommend use of the English NIHTB-CB given that the normative standards for the English battery include such bilingual speakers. Nonetheless, in such assessments, careful attention should be paid to the individual's reading and vocabulary performances when interpreting Fluid scores. Reading performance, in particular, can provide an objective indicator of prior education quality (e.g., Manly, Jocobs, Tourdji, Samll, & Stern, 2002; Manly, Byrd, Touradji, & Stern, 2004), which consistently demonstrates strong, positive effects on Fluid cognition and can, therefore, aid users in understanding and anticipating deviations from expected levels of performance.

Lastly, at the moment, considerable clinical judgment would be needed to interpret test scores in monolingual Spanish children ages 8–17 years old. In this case, a researcher or clinician could apply both the 7- and 18-year-old normative standards and, given that age demonstrates a linear, positive effect on neurocognitive performances in childhood (both in the current study and in ages 8–17 years on the English NIHTB-CB norms; see Casaletto et al., 2015), come to an informed estimate of performance range. Nonetheless, we continue to recommend conservative allowances for differences in backgrounds that may impact performance in these children (e.g., educational quality, acculturation factors).

Taken together, the Spanish NIHTB-CB norms described here were developed exclusively for a U.S. Spanish-speaking population and may best represent this cohort of individuals. As illustrated, there were important differences in the relationships between demographics and NIHTB-CB performances among the Spanish-speakers (compared to English-speakers) that were not accounted for in the originally developed norms. As with the English normative standards (Casaletto et al., 2015), the NIH Toolbox initiative plans to incorporate the Spanish NIHTB-CB norms presented here into the scoring system, however, they are not yet currently available online. Therefore, in the interim, they will be available for use *via* an

Excel program, which can be obtained by emailing the authors. Given the complexity of the normative formulas (Appendix 2), we recommend that users use the Excel program (with embedded formulas) rather than program and apply the formulas independently. The NIHTB-CB itself can be accessed at www.nihtoolbox.org.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The Original NIH Toolbox development contract and adult norming was supported by Blueprint for Neuroscience Research and the Office of Behavioral and Social Sciences Research, National Institutes of Health, under Contract No. HHS-N-260-2006-00007-C. The child and parent norming (includes some of the data used here for adults) was supported by Health Measurement Network for the National Children's Study. National Institutes of Health – NICHD, HHSN267200700027C. This work was also supported by the National Institute for Health grants F31-DA035708 and (in part) by a Foundation for Rehabilitation Psychology Dissertation Award.

REFERENCES

Ambler, G.; Benner, A. mfp: Multivariable Fractional Polynomials. 2008. Retrieved from http://stat.ethz.ch/CRAN/

Beaumont JL, Havlik R, Cook KF, Hays RD, Wallner-Allen K, Korper SP, Gershon R. Norming plans for the NIH Toolbox. Neurology. 2013; 80(11 Suppl 3):S87–S92. doi:10.1212/WNL.0b013e3182872e70. [PubMed: 23479550]

Benson G, de Felipe J, Xiaodong L, Sano M. Performance of Spanish-speaking community-dwelling elders in the United States on the Uniform Data Set (UDS). Alzheimers & Dementia. 2014; 10(S5):S338–S343. doi:10.1016/j.jalz.2013.09.002.

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

Casaletto KB, Umlauf A, Beaumont J, Gershon R, Slotkin J, Akshoomoff N, Heaton RK. Demographically corrected normative standards for the English version of the NIH toolbox cognition battery. Journal of the International Neuropsychological Society. 2015; 21:1–14. [PubMed: 25399546]

Cattell, RB. Abilities: Their structure, growth, and action. Cambridge University Press; Cambridge: 1971.

Cherner M, Suarez P, Lazzaretto D, Fortuny LA, Mindt MR, Dawes S, Heaton R. Demographically corrected norms for the Brief Visuospatial Memory Test-revised and Hopkins Verbal Learning Test-revised in monolingual Spanish speakers from the U.S.-Mexico border region. Archives of Clinical Neuropsychology. 2007; 22(3):343–353. doi:S0887-6177(07) 00016-9 [pii]10.1016/j.acn.2007.01.009. [PubMed: 17293078]

Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Evaluation & The Health Professions. 2005; 28(2):212–232. [PubMed: 15851774]

Fuller B, Bein E, Kim Y, Rabe-Hesketh S. Differing cognitive trajectories of Mexican American toddlers: The role of class, nativity, and maternal practices. Hispanic Journal of Behavioral Sciences. 2015; 37(2):139–169. doi:10.1177/0739986315571113.

Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. Neurology. 2013; 80(11 Suppl 3):S2–S6. [PubMed: 23479538]

Heaton, RK.; Miller, SW.; Taylor, JT.; Grant, I. Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults. Psychological Assessment Resources, Inc.; Lutz, FL: 2004.

Heaton RK, Akshoomoff N, Tulsky D, Mungas D, Weintraub S, Dikmen S, Gershon R. Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. Journal of the International Neuropsychological Society. 2014; 20(6):588–598. doi:10.1017/S1355617714000241. [PubMed: 24960398]

Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y. Reading level attenuates differences in neuropsychological test performance between African American and White elders. Journal of the International Neuropsychological Society. 2002; 8(3):341–348. [PubMed: 11939693]

Manly JJ, Byrd DA, Touradji P, Stern Y. Acculturation, reading level, and neuropsychological test performance among African American elders. Applied Neuropsychology. 2004; 11(1):37–46. [PubMed: 15471745]

Mungas D, Reed BR, Crane PK, Haan MN, Gonzalez H. Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics. Psychological Assessment. 2004; 16(4):347–359. doi:10.1037/1040-3590.16.4.347. [PubMed: 15584794]

Mungas D, Reed BR, Haan MN, Gonzalez H. Spanish and English neuropsychological assessment scales: Relationship to demographics, language, cognition, and independent function. Neuropsychology. 2005; 19(4):466–475. doi:10.1037/0894-4105.19.4.466. [PubMed: 16060821]

Mungas D, Reed BR, Marshall SC, Gonzalez HM. Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology. 2000; 14(2): 209–223. [PubMed: 10791861]

Muñoz-Sandoval, AF.; Woodcock, RW.; McGrew, KS.; Mather, N. Bateria III Woodcock-Muñoz. Riverside Publishing; Itasca, IL: 2005.

Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling. Applied Statistics. 1994; 43(3):429–467.

Ryan, C. Language Use in the United States: 2011. American Community Survey Reports. 2013. Retrieved from http://www.census.gov/prod/2013pubs/acs-22.pdf

Stricks L, Pittman J, Jacobs DM, Sano M, Stern Y. Normative data for a brief neuropsychological battery administered to English- and Spanish-speaking community-dwelling elders. Journal of the International Neuropsychological Society. 1998; 4(4):311–318. [PubMed: 9656604]

Taylor MJ, Heaton RK. Sensitivity and specificity of WAIS-III/WMS-III demographically corrected factor scores in neuropsychological assessment. Journal of the International Neuropsychological Society. 2001; 7:867–874. [PubMed: 11771630]

U.S. Census Bureau. Population estimates. National characteristics: Vintage 2013. 2014. Retrieved from http://www.census.gov/popest/data/national/asrh/2013/index.html

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

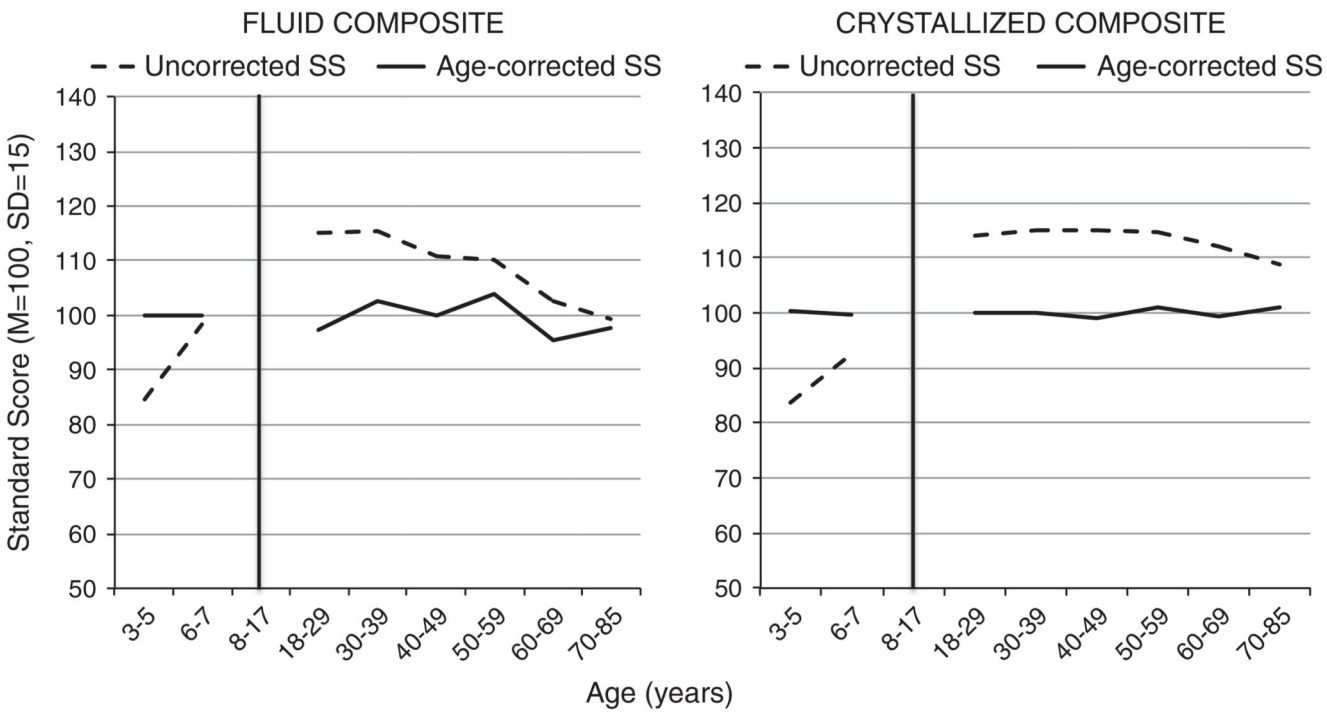

Fig. 1.

Uncorrected and age-corrected Spanish NIH Toolbox Cognition Battery Fluid and Crystallized composite performances by age.

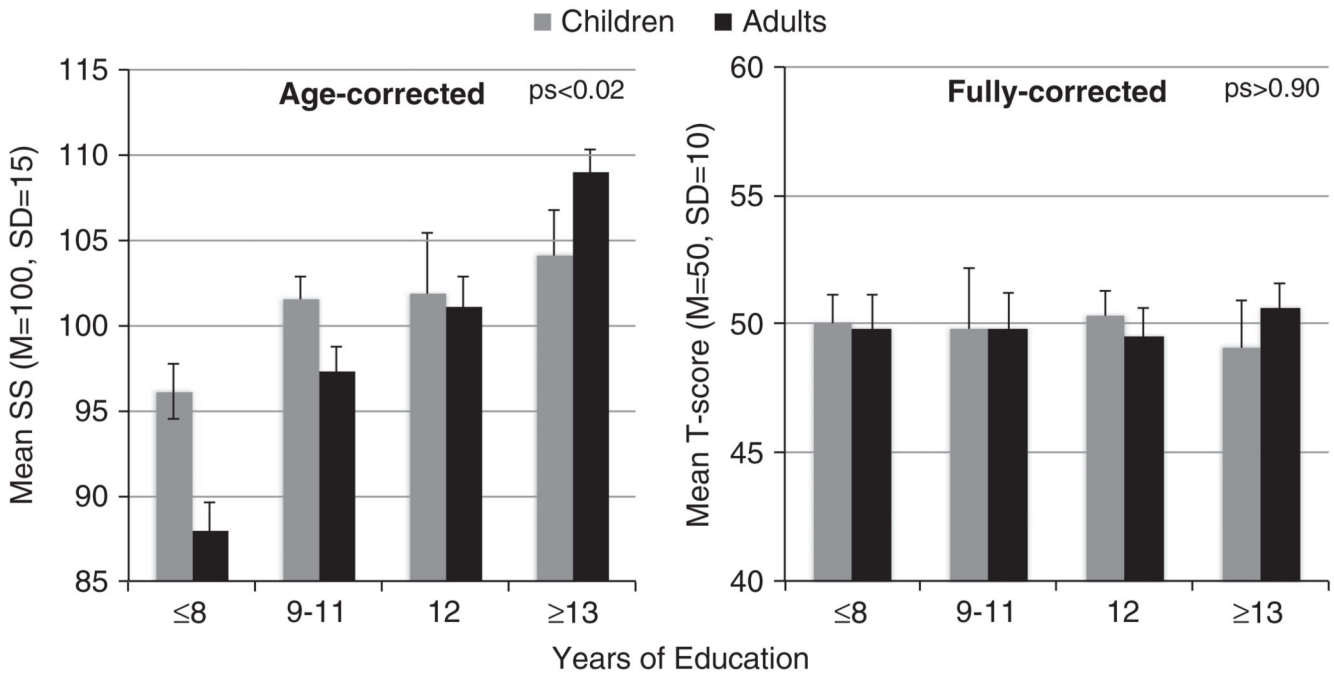

Fig. 2.

Age-corrected *versus* fully corrected Fluid composite scores by education across Spanish-speaking adults and children. Note. For children (ages 3–7), "Years of Education" refers to mothers' educational levels.

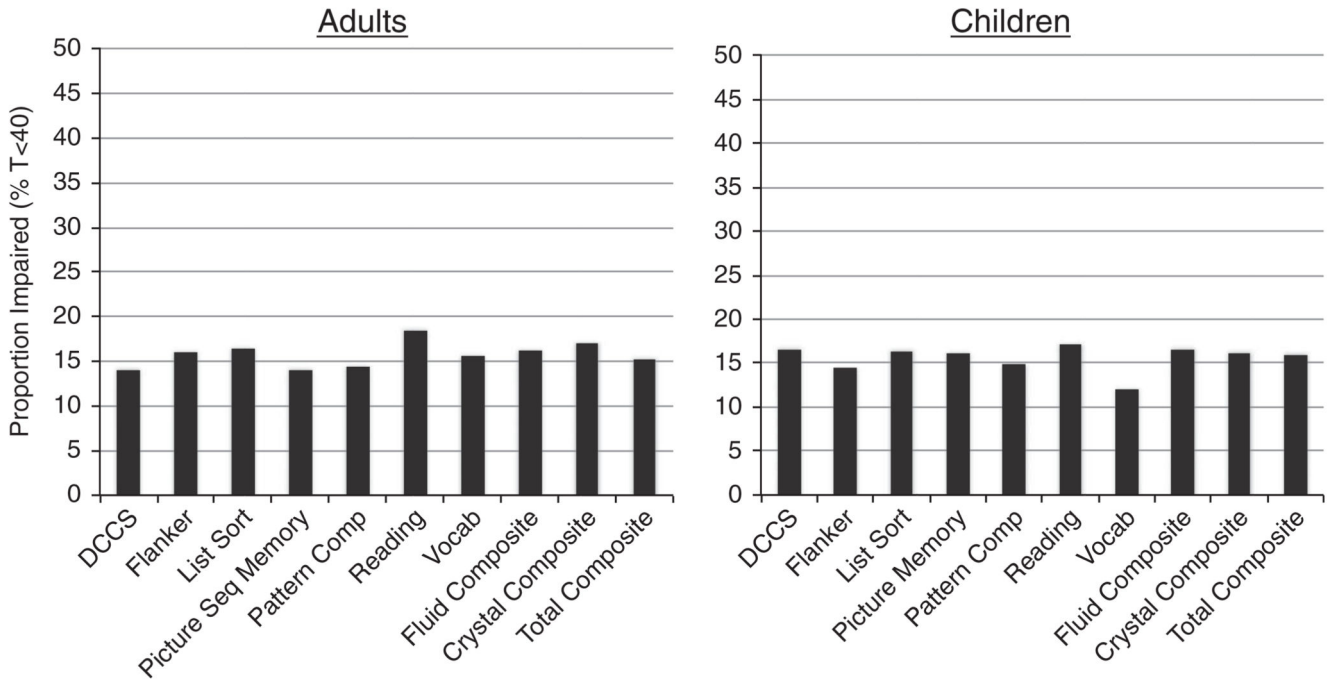

Fig. 3.

Fully demographically adjusted (age, education, sex) Spanish NIH Toolbox Cognition Battery impairment rates across children and adults. Note: "Impairment" classified as *T* score <40.

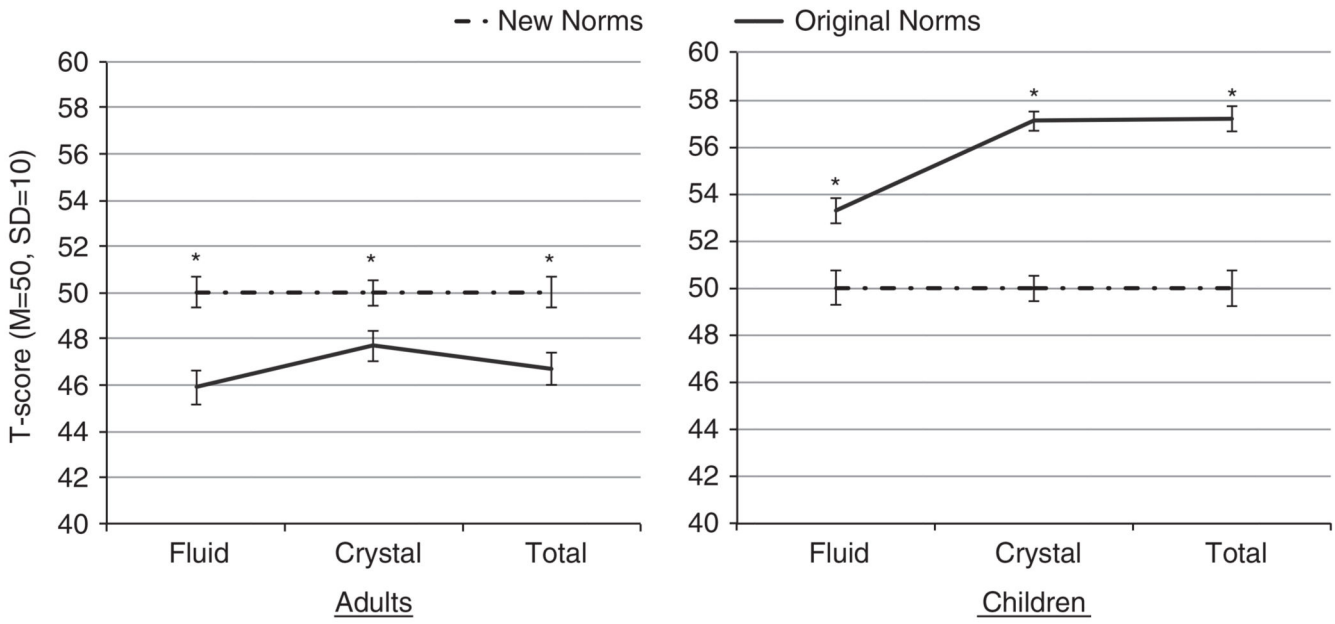

Fig. 4.

Spanish NIH Toolbox Cognition Battery fully demographically adjusted T-scores: Original norms *versus* new norms. *$p < .05$ Original Norms differ from New Norms.

Table 1

Demographic, cultural, and language backgrounds of the Spanish-speaking normative cohorts

	Adults (*n* = 408)	Children (*n* = 496)
Age	44.1 (16.7)	4.9 (1.4)
	range: 18–85	range: 3–7
Education/mother's education	10.7 (4.3)	9.7* (3.9)
Sex (% M)	35.0% (143)	47.6% (236)
Race/ethnicity (%, *n*)		
Hispanic White	77.0% (314)	78.4% (389)
Native American	9.6% (39)	15.9% (65)
Hispanic AfAm	5.4% (22)	1.0% (5)
Multiracial	1.7% (7)	1.6% (8)
Asian	0.5% (2)	—
Pacific Islander	0.5% (2)	—
Non-Hispanic White	—	0.2% (1)
Other/nonspecified	5.4% (22)	5.6% (28)
Born in the U.S. missing (*n* = 50)	22.1% (79)	—
Any school in the U.S. Missing (*n* = 53)	45.4% (161)	—
First language learned		
Spanish	94.2% (338)	
English	5.6% (20)	—
Other	0.3% (1)	
Missing (*n* = 49)		
Language spoken at home		—
Spanish	68.0% (244)	
English	12.8% (46)	
English and Spanish	19.2% (69)	
Missing (*n* = 49)		

*Mother's education.

Table 2

Linear univariable effects of demographic factors on NIH Toolbox Cognition Battery performances

Adults tested in Spanish (*N* = 408)			
	Age[a] (*r*)	**Education[b] (*r*)**	**Sex[b] (Cohen's *d*)[c]**
DCCS	−0.46**	0.43**	0.27* M> F
Flanker	−0.53**	0.43**	0.31* M> F
List Sort	−0.43**	0.30**	0.05
Pattern Comparison	−0.50**	0.23**	0.19
Picture Sequence Memory	−0.42**	0.34**	0.13
Oral Reading	−0.23**	0.27**	0.06
Picture Vocabulary	−0.19**	0.32**	0.02
Fluid Composite	−0.50**	0.53**	0.22
Crystallized Composite	−0.21*	0.31**	0.04
Total Composite	−0.38**	0.54**	0.12

Children tested in Spanish (ages 3–7; *N* = 496)			
	Age[a] (*r*)	**Education[b] (Mother's education) (*r*)**	**Sex[b] (Cohen's d)[c]**
DCCS	0.50**	0.17**	0.09
Flanker	0.67**	0.11*	0.08
List Sort	0.55**	0.15*	0.20
Pattern Comparison	0.59**	0.06	0.10
Picture Sequence Memory	0.70**	0.07	0.17
Oral Reading	0.61**	0.04	0.08
Picture Vocabulary	0.61**	0.06	0.31** F>M
Fluid Composite	0.76**	0.23**	0.03
Crystallized Composite	0.69**	0.10	0.21
Total Composite	0.77**	0.12	0.15

[a] Values reflect relationships with uncorrected normalized test scores.

[b] Values reflect relationships with age-corrected test scores.

[c] Cohen's d represent absolute values.

[**] $p<0.001$.

[*] $p< 0.05$.