




Research Article

A psychometric evaluation of the NIH Toolbox fluid cognition tests adapted for Swahili and Dholuo languages in Kenyan children and adolescents

Megan S. McHenry^{1,2} , Anna Roose¹, Emily Abuonji², Mark Nyalumbe³, David Ayuku³, George Ayodo⁴, Tuan M. Tran⁵ and Aaron J. Kaat⁶ 

¹Ryan White Center for Pediatric Infectious Diseases and Global Health, Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN, USA, ²Academic Model Providing Access to Health (AMPATH), Eldoret, Kenya, ³College of Health Sciences, Moi University, Eldoret, Kenya, ⁴Jaramogi Oginga Odinga University of Science and Technology, Bondo, Kenya, ⁵Division of Infectious Diseases, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA and ⁶Northwestern University, Chicago, IL, USA

Abstract

Objective: Our objective was to evaluate the psychometric properties of the culturally adapted NIH Toolbox African Languages[®] when used in Swahili and Dholuo-speaking children in western Kenya. **Method:** Swahili-speaking participants were recruited from Eldoret and Dholuo-speaking participants from Ajigo; all were <14 years of age and enrolled in primary school. Participants completed a demographics questionnaire and five fluid cognition tests of the NIH Toolbox[®] African Languages program, including Flanker, Dimensional Change Card Sort (DCCS), Picture Sequence Memory, Pattern Comparison, and List Sorting tests. Statistical analyses examined aspects of reliability, including internal consistency (in both languages) and test–retest reliability (in Dholuo only). **Results:** Participants included 479 children ($n = 239$, Swahili-speaking; $n = 240$, Dholuo-speaking). Generally, the tests had acceptable psychometric properties for research use within Swahili- and Dholuo-speaking populations (mean age = 10.5; $SD = 2.3$). Issues related to shape identification and accuracy over speed limited the utility of DCCS for many participants, with approximately 25% of children unable to match based on shape. These cultural differences affected outcomes of reliability testing among the Dholuo-speaking cohort, where accuracy improved across all five tests, including speed. **Conclusions:** There is preliminary evidence that the NIH Toolbox[®] African Languages potentially offers a valid assessment of development and performance using tests of fluid cognition in Swahili and Dholuo among research settings. With piloting underway across other diverse settings, future research should gather additional evidence on the clinical utility and acceptability of these tests, specifically through the establishment of norming data among Kenyan regions and evaluating these psychometric properties.

Keywords: Cognitive assessments; pediatric; NIH Toolbox; adaptation; Kenya

(Received 5 May 2023; final revision 18 August 2023; accepted 21 September 2023)

Introduction

Cognition – the ability to learn, problem solve, remember, and retrieve information – is a critical part of the human experience. Cognition's bidirectional relationship with general achievement plays an important role in an individual's life (Peng & Kievit, 2020), impacting their education attainment, future economic potential, and overall physical and mental health (Lövdén et al., 2020; Ozawa et al., 2022; Wrulich et al., 2014). Over 500 million of the world's children live in sub-Saharan Africa, where numerous risk factors can negatively impact child development and cognitive outcomes (UNICEF, 2015). As cognition is the foundation for human development and potential, it is critical to study and understand the different risk factors impacting cognition across diverse populations to mitigate any harm and optimize developmental trajectories.

Valid tools are needed to evaluate cognition within diverse populations globally, but most high quality, internationally known tools have been designed and validated only for USA or European populations. Culturally adapted, evidence-based tools to measure cognition are needed in order to advance the evidence on factors associated with optimizing cognitive outcomes or to develop interventions supporting cognition in this region. The NIH Toolbox[®] is a set of standardized and normed tests in the cognitive, motor, sensory, and emotional domains that were publicly released in 2012 (Gershon et al., 2013) and released as an iPad application in 2014, developed by more than 250 scientists over 6 years and sponsored by the National Institutes of Health (Hodes et al., 2013). The NIH Toolbox[®] Cognition Battery was normed using US census data of nearly 3000 typically developing children and adolescents aged 3–17 years (Casaletto et al., 2015).

Corresponding author: Megan S. McHenry; Email: msuhl@iu.edu

Cite this article: McHenry M.S., Roose A., Abuonji E., Nyalumbe M., Ayuku D., Ayodo G., Tran T.M., & Kaat A.J. (2023) A psychometric evaluation of the NIH Toolbox fluid cognition tests adapted for Swahili and Dholuo languages in Kenyan children and adolescents. *Journal of the International Neuropsychological Society*, 29: 933–942, <https://doi.org/10.1017/S1355617723000632>

© The Author(s), 2023. Published by Cambridge University Press on behalf of International Neuropsychological Society.

Table 1. NIH Toolbox® Dholuo fluid cognition tests

Test	Ages	Time to administer	Scoring Scale	Domains measured
Dimensional Change Card Sort test	7–85 years	4 min	0–10	Cognitive flexibility
Flanker test	7–85 years	3 min	0–3 ^a	Attention and inhibitory control
List Sorting	7–85 years	7 min	0–26	Working memory
Pattern Comparison	7–85 years	3 min	0–130	Processing speed
Picture Sequence Memory	7–85 years	7 min	0–31	Episodic memory

^aAdjusted scoring model for this tool.

While extensive evidence is available on the validity of the English-language version within the USA (Mungas et al., 2013; Weintraub et al., 2013) and some work supports its use in Spanish-speaking populations (Gershon et al., 2020; Victorson et al., 2013), it has not been used widely outside of this context. In 2020, The NIH Toolbox® African Languages, including Swahili and Dholuo languages, was created utilizing an extensive translation and cultural adaptation process (Duffey et al., 2022). This African Languages Toolbox included a battery of fluid cognition tasks, which are tasks that measure aspects less reliant on learned knowledge, including problem-solving and executive functioning (Ferrer et al., 2009).

However, NIH Toolbox® African Languages has yet to be evaluated to determine whether their psychometric properties support use within this population. This introduces a critical void in the accessibility of this tool, as there are many differences in culture to be expected between the cultures of well-resourced, English-speaking settings and low-resourced settings in sub-Saharan Africa. By evaluating the psychometric properties NIH Toolbox® African Languages, we will determine its potential use within this population for future research purposes. Among the available cognitive tools that have been adapted and used in this setting (Alcock et al., 2008; McHenry et al., 2023), the NIH Toolbox® African Languages has two primary benefits: (1) tasks are administered and scored by a tablet computer, making it scalable and easily administered; (2) it can evaluate a wide age range, including those 7 years and older, versus other tests that target specific populations with narrow age windows (McHenry et al., 2023). Without evaluating the psychometric properties of tools used to measure cognition within children in Kenya, we will be unable to produce high-quality evidence needed to develop critical interventions and optimize cognitive outcomes within this population.

The objective of this study was to evaluate the psychometric properties of the culturally adapted NIH Toolbox® African Languages when used in Swahili and Dholuo-speaking populations in western Kenya. We aimed to evaluate internal consistency and test-retest reliability, of which we hypothesized moderate correlations. While our primary aim was not to measure the overlapping effects of demographic information, we recognized its importance and sought to measure demographics as exploratory in nature. Given the age range of this sample, we also hypothesized a monotonic increasing relationship between the test scores and chronological age.

Methods

Participants

We aimed to recruit 480 children aged 7–14 years from two different cities of western Kenya. We recruited 240 Dholuo-speaking participants and 239 Swahili-speaking study participants, aged 7–14 years (with $n = 30$, per year age band), totaling 479 participants. The recruitment target per age band was aligned with

other psychometric studies in sub-Saharan Africa (Gladstone et al., 2010; Holding et al., 2004).

Swahili and English are the official languages in Kenya and the compulsory languages in most Kenyan academic systems. However, many of the 42 tribes across Kenya have their own mother tongue; thus, some participants may speak multiple languages. Inclusion criteria were as follows: (1) between 7 and 14 years of age; (2) fluently spoke the specific language of interest for that study site (Swahili for Eldoret, Dholuo for Ajigo); (3) and had a primary caregiver who spoke either Swahili or Dholuo and was available for consent and completion of a questionnaire. The questionnaire included information about study participants' age, sex, water, sanitation, maternal education, and household income to determine general sociodemographic status. Additionally, participant health was evaluated within the questionnaire with items inquiring about health status at the time of testing, determined by the absence of fever, cough, rash, or any other symptoms within 2 weeks of the study visit.

Measures

Each of the NIH Toolbox® tests were developed and scored to target specific aspects of fluid cognition. The specific adaptations made to the original NIH Toolbox® to create the NIH Toolbox® African Language are outlined in depth within Duffey et al. (2022). Adaptations were necessary to ensure the displayed objects were representative of items known within sub-Saharan African (replacement of food and animal items), and that the translated language accurately conveyed the constructs and ideas intended within the original English version (Duffey et al., 2022). The adapted Fluid Cognition Tests on this iPad application include: Pattern Comparison (PC), the Flanker test, Dimensional Change Card Sort (DCCS) test, Picture Sequence Memory (PSM), and List Sorting (LS) (Table 1). Prior to administration of the battery, a practice tutorial was completed. This tutorial required participants to practice skills needed to complete the battery, such as clicking on an image or dragging an image inside or outside of a box. After successful completion of the tutorial, each battery, following the same order, was presented in correspondence with the age of the child.

In the Flanker test, directional fish (for younger children) and arrows (for older children) are placed in a row, and the participant must indicate the direction in which the middle object (fish/arrow) is pointing. This test of attention and inhibitory control is scored on accuracy and response time. Participants are instructed to work as fast as they can without making mistakes; while this instruction appears to work appropriately in Western cultures, we recognized that participants of the African Languages version made different decisions on the speed-accuracy trade-off. Therefore, we chose to calculate the rate-correct score (Liesefeld & Janczyk, 2019), which has been recommended for other tests involving a trade-off between speed and accuracy. This approach has also been used in

other investigations of the NIH Toolbox[®] (Shields et al., 2020), though it is not part of the traditional scoring approach.

DCCS assesses cognitive flexibility and set-shifting by presenting three images. One image is the bivalent target item (e.g., a blue ball), and the other two items are options that either match by shape (e.g., a red ball) or color (e.g., a blue star). Participants are then asked to match the target item to the other option by color or shape. This test is scored on accuracy and response time.

The PSM measures episodic memory by presenting a series of images related to a story or action, with each image associated with an auditory cue. Then the images are scrambled, and the participant is asked to re-order them correctly. This test is scored on accuracy of the location of the items, including whether two adjacent items are correctly located next to one another (Bauer et al., 2013; Dikmen et al., 2014). LS evaluated working memory by measuring the participant's ability to recall and re-arrange a series of objects presented visually with verbal cues and then verbally repeat them back to the examiner. This test is scored on accuracy of items.

PC examines processing speed by measuring the timed accuracy of the participant to determine items correctly identified within an 85-s period. For tests where response time is a part of the scoring, a "home base" – a set mark placed in a fixed distance away from the iPad – is used to standardize the distance each participant moves between the start of an item presentation and the screen and to keep the response time consistent across trials and participants.

Procedure

Setting

This psychometric cohort study was performed at two study sites within western Kenya between October 2020 and February 2021. These sites – Eldoret and Ajigo – were also the setting of prior work focused on the translation and cultural adaptation of the NIH Toolbox[®] cognitive tests into Swahili and Dholuo languages (Duffey et al., 2022). In the municipality of Eldoret, recruitment and study activities took place at a government-funded public central school, with the research team based within the Academic Model Proving Access To Healthcare (AMPATH) Program. AMPATH is a long-standing academic partnership between Moi University School of Medicine, Moi Teaching and Referral Hospital (MTRH), and a consortium of North American academic centers, led by Indiana University. In the Ajigo ward in the county of Siaya, recruitment took place at a local rural primary school, study activities took place at the Gobei Health Center, and the Ajigo research team was based within the Kenya Medical Research Institute (KEMRI) Kisumu program. This study received ethical approvals from Moi University, KEMRI, and Indiana University, and all human data included in this manuscript was obtained in compliance with regulations of these institutions.

Study activities

Research was completed in accordance with the Declaration of Helsinki. Our research teams worked collaboratively with the primary schools engaged for recruitment. An in-person meeting was first held with the head teacher at each school to discuss the details of the study and seek approval, of which all schools agreed. The head teacher at each school then provided a student register with child ages. For each age group, students' names were randomized within a list and then contacted in subsequent order to provide study information to a primary caregiver through the headmasters and teachers at the school. No written advertisements

were used. We continued to recruit within each age group until 30 participants were recruited. Within the 7-year-old age band in the Swahili group, only 29 participants were recruited due to small class size related to the COVID-19 pandemic. Prior to initiation of study activities, information sheets and informed consents were reviewed with a research assistant at each study site, written consent was obtained from caregivers, and written assent was obtained from 13- and 14-year-old study participants; all components of the informed consent process took place in the caregivers' preferred language.

Together, Swahili and Dholuo-speaking primary caregivers and study participants completed a questionnaire containing information about the child's age, sex, and maternal education. Household data, including assets, household income, and access to improved water and sanitation (as defined by WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene), were included in this questionnaire given the understanding that they may limit opportunities for cognitive growth (UNICEF, 2015). These household covariates were primarily collected for a different aspect of the research project.

After completion, a research assistant took the study participant to a quiet, private room for cognitive testing. COVID-19 precautions included: mask wearing, hand sanitation before and after testing, and iPad, chair, and surface cleaning between participants. Research assistants were trained to administer the NIH Toolbox[®] African Languages application. There were two research assistants administering the application among the Swahili sample, and one among the Dholuo participants. This training included completion of the NIH Toolbox[®] Online training curriculum. Additionally, each assistant reviewed standard operating procedures with the study principal investigator (MSM) for 6-8 hours of one-on-one training. Study visits took approximately 60–90 min.

Among Swahili-speaking participants, a single cross-sectional study visit was performed, which was within the scope of our planned cultural adaptation and pilot study. For the Dholuo-speaking participants, test-retest reliability evidence was also collected due to additional aspects of the primary study. After the initial study visit, a random number generator was used to select 120 participants for a follow-up visit for repeat NIH Toolbox[®] testing, occurring approximately 8 weeks (median = 41 days, range 39–75 days) after initial visit. All participants received 400 Kenyan shillings (approximately \$3.50 USD) for their time required to participate in the study, which was the standard rate approved by the Kenya ethics board for such procedures.

Data analysis

The primary analyses were psychometric evaluations of each of the NIH Toolbox[®] African Languages fluid cognition measures. We examined the relationship between scores and various participant demographics as an exploratory analysis. This included correlations and chi-square tests, depending on the distribution of the relevant demographic variable. Alpha was set at 0.05.

Convergent validity

Fluid cognition scores are expected to increase rapidly during early childhood, stabilize in adolescence and early adulthood, and then start to decline in older age. A high correlation provides validity evidence for expected developmental curves. We evaluated the relationship between test scores and age using Spearman rank-

Table 2. Study participant characteristics

	Total	Swahili-speaking	Dholuo-speaking	<i>p</i> -Value
	<i>N</i> = 479	<i>N</i> = 239	<i>N</i> = 240	
Child age (<i>SD</i>)	10.5 (2.3)	10.5 (2.3)	10.5 (2.3)	0.91
Female sex (%)	268 (55.9%)	150 (62.8%)	118 (49.2%)	0.004
Access to improved water (%)	311 (64.9%)	163 (68.2%)	148 (61.7%)	0.16
Access to improved sanitation (%)	163 (34%)	72 (30.1%)	91 (37.9%)	0.09
Mean maternal age (<i>SD</i>)	36.3 (9)	35.6 (7.6)	36.9 (10.1)	0.11
Mother's educational attainment				
No formal education/some primary school (%)	133 (27.8%)	67 (28%)	66 (27.5%)	0.002
Completed primary school (%)	211 (44.1%)	90 (37.7%)	121 (50.4%)	
Some secondary school (%)	42 (8.8%)	23 (9.6%)	19 (7.9%)	
Completed secondary school (%)	83 (17.3%)	56 (23.4%)	27 (11.2%)	
Attended university (%)	10 (2.1%)	3 (1.3%)	7 (2.9%)	
Household income				
Lower 3rd (%)	121 (25.6%)	49 (20.9%)	72 (30.4%)	0.002
Middle 3rd (%)	176 (37.3%)	103 (43.8%)	73 (30.8%)	
Upper 3rd (%)	87 (18.4%)	49 (20.9%)	38 (16%)	
Income unknown (%)	88 (18.6%)	34 (14.5%)	54 (22.8%)	
Household assets				
Kitchen (%)	380 (79.3%)	187 (78.2%)	193 (80.4%)	0.64
Mattress (%)	444 (92.7%)	219 (91.6%)	225 (93.8%)	0.47
TV (%)	217 (45.3%)	114 (47.7%)	103 (42.9%)	0.34
Bank (%)	127 (26.5%)	65 (27.2%)	62 (25.8%)	0.81

Bolded values indicate statistical significance.

order correlations and examined the intercorrelations between the five fluid tests using Pearson correlations.

While convergent validity would ideally evaluate the same construct with a different measure, few tests are appropriate or have been translated into these African languages with any evaluation of psychometric properties. While a few Swahili version tests are available (NeuroScreen, Plus EF, among others) (McHenry et al., 2023), only Tangerine EF Touch has published its psychometric properties (Willoughby et al., 2010, 2012, 2019). Unfortunately, this measure is limited to 3–5 year olds and cannot be compared to the NIH Toolbox[®] African Languages. No psychometric data for cognitive tests have been published for use with the Dholuo language. As such, we chose to evaluate the relationship between these domains, as they all should represent fluid cognitive abilities. This is an imperfect evaluation of convergent validity, but highlights the critical assessment void that the NIH Toolbox[®] African Languages addresses.

Reliability testing

The internal consistency reliability of the tests were evaluated using either Cronbach's alpha (for fixed-length tests) or median split-half correlation with Spearman–Brown correction across 10,000 permutations (for variable-length tests). The test structure of PSM precluded calculation of an internal consistency estimate; scores should increase across trials due to learning – preventing a correlation of the two trials – and the sum of adjacent pairs is less conducive to permutation and random splits. For the Dholuo sample, test–retest reliability and practice effects were evaluated using linear mixed models with a person-specific random intercept. The intraclass correlation (ICC) from the random intercept indexes test–retest reliability, while the fixed effect for assessment occasion estimates practice effects. For the linear mixed effect models, alpha was set at 0.05. For all of these analyses, interpretation of the psychometric findings follow published rules for magnitude of effects (Cohen, 1988; Nunnally & Bernstein, 1994).

Results

Between the Swahili- and Dholuo-speaking cohorts, the former had a higher incomes and a higher proportion of females, as well as mothers who completed schooling beyond primary school (Table 2).

Ownership of a television was positively associated with scores of three tests: DCCS (p -value = 0.019), Flanker (p = 0.06), and PSM (p < 0.01). Higher income levels were also positively associated with scores of three tests: Flanker (p = 0.02), PC (p = 0.03), and PSM (p = 0.009). Other associations are noted within Table 3, with a sub-analysis by language featured in Appendix A.

The relationships between age and NIH Toolbox[®] scores indicated weak associations, with participants performing better as their age increased (p < 0.01), for both the Dholuo- and Swahili-language Flanker tests (Figure 1). For the DCCS test, scores were noted to cluster around 2.13 in both the Swahili and Dholuo samples. This score value is obtained with a DCCS accuracy score of 17, corresponding to correctly answering (or being old enough to skip) seven pre- and post-switch items on the 30 mixed trials, seven of which are color-matching mixed trials. Within the Swahili sample, 45 participants (19%) scored 2.13, with all but one correctly answering every color question and missing every shape question. Within the Dholuo group, 25% (n = 58) missed all shaped items and got all color questions correct. Participants more frequently scored 2.13 at younger ages, but this scoring occurred across the range of study participants in both cohorts. Additionally, a few participants were unable to correctly identify the color items, but the majority of those were 7 or 8 years old. Finally, participants who did not achieve adequate accuracy on practice items of the Flanker and DCCS tests did not receive scores for these domains.

Convergent validity

Among the Swahili-language tests, although all of the test intercorrelations were statistically significant (all p < 0.05), only

Table 3. Univariate association of study participant characteristics with NIH Toolbox® scores

Covariate	DCCS	Flanker	List Sorting	Pattern Comparison	Picture Sequence Memory
Mother's age	0.025 (0.003–0.047), P = 0.029	0 (0–0), P = 0.79	–0.001 (–0.036 to 0.034), P = 0.958	0.099 (0.003–0.196), P = 0.044	0.015 (–0.038 to 0.068), P = 0.584
Improved water	0.321 (–0.095 to 0.736), P = 0.132	0.03 (–0.02 to 0.07), P = 0.23	–0.001 (–0.665 to 0.662), P = 0.997	1.299 (–0.524 to 3.122), P = 0.163	0.507 (–0.485 to 1.499), P = 0.317
Improved sanitation	–0.255 (–0.675 to 0.164), P = 0.233	–0.06 (–0.11 to –0.02), P = 0.01	–0.417 (–1.086 to 0.251), P = 0.222	–1.251 (–3.087 to 0.586), P = 0.183	0.062 (–0.938 to 1.062), P = 0.903
Child's age	0.485 (0.41–0.56), P < 0.01	0.02 (0.01–0.03), P < 0.01	0.688 (0.564–0.812), P < 0.01	1.975 (1.638–2.312), P < 0.01	0.829 (0.636–1.022), P < 0.01
Male gender	–0.176 (–0.576 to 0.224), P = 0.389	0 (–0.05 to 0.04), P = 0.86	0.293 (–0.345 to 0.93), P = 0.369	–0.999 (–2.753 to 0.754), P = 0.265	–0.225 (–1.179 to 0.729), P = 0.645
Educational attainment					
No formal education/some primary school	–0.146 (–0.625 to 0.333), P = 0.55	–0.03 (–0.08 to 0.03), P = 0.35	0.162 (–0.605 to 0.93), P = 0.678	–0.628 (–2.717 to 1.461), P = 0.556	0.969 (–0.177 to 2.114), P = 0.098
Completed primary school	Reference	Reference	Reference	Reference	Reference
Some secondary school	–0.548 (–1.279 to 0.183), P = 0.143	0.03 (–0.05 to 0.11), P = 0.42	1.045 (–0.13 to 2.219), P = 0.082	1.036 (–2.152 to 4.224), P = 0.524	0.96 (–0.788 to 2.708), P = 0.282
Completed secondary school	0.516 (–0.045 to 1.076), P = 0.072	0.05 (–0.01 to 0.11), P = 0.11	0.252 (–0.645 to 1.148), P = 0.582	2.851 (0.406–5.296), P = 0.023	1.352 (0.012–2.692), P = 0.049
Attended university	–0.833 (–2.233 to 0.567), P = 0.244	–0.04 (–0.19 to 0.11), P = 0.64	–1.15 (–3.374 to 1.073), P = 0.311	–7.183 (–13.289 to –1.076), P = 0.022	0.812 (–2.536 to 4.16), P = 0.635
Income level					
Lower 3rd	Reference	Reference	Reference	Reference	Reference
Middle 3rd	0.131 (–0.381 to 0.642), P = 0.616	0.07 (0.01–0.12), P = 0.02	0.236 (–0.582 to 1.054), P = 0.572	2.49 (0.245–4.734), P = 0.03	0.965 (–0.249 to 2.179), P = 0.12
Top 3rd	0.482 (–0.127 to 1.091), P = 0.121	0.03 (–0.04 to 0.09), P = 0.42	0.495 (–0.47 to 1.461), P = 0.315	1.582 (–1.089 to 4.254), P = 0.246	1.923 (0.479–3.368), P = 0.009
Income unknown	–0.219 (–0.826 to 0.387), P = 0.479	0.04 (–0.03 to 0.1), P = 0.27	–0.562 (–1.528 to 0.403), P = 0.254	–0.181 (–2.843 to 2.482), P = 0.894	0.09 (–1.35 to 1.53), P = 0.903
Assets					
Kitchen	–0.704 (–1.191 to –0.216), P = 0.005	–0.04 (–0.09 to 0.01), P = 0.12	–0.51 (–1.293 to 0.272), P = 0.202	–1.229 (–3.378 to 0.921), P = 0.263	–1.006 (–2.172 to 0.161), P = 0.092
Mattress	–0.295 (–1.059 to 0.469), P = 0.449	–0.08 (–0.16 to 0), P = 0.06	0.053 (–1.155 to 1.262), P = 0.931	–1.803 (–5.148 to 1.543), P = 0.291	–0.718 (–2.537 to 1.102), P = 0.44
Television	0.478 (0.081–0.876), P = 0.019	0.04 (0–0.08), P = 0.06	0.45 (–0.186 to 1.086), P = 0.166	1.634 (–0.111 to 3.379), P = 0.067	1.806 (0.868–2.744), P < 0.01
Bank	–0.019 (–0.47 to 0.431), P = 0.933	–0.03 (–0.08 to 0.02), P = 0.27	–0.069 (–0.787 to 0.649), P = 0.851	–1.041 (–3.013 to 0.932), P = 0.302	–0.24 (–1.313 to 0.833), P = 0.661

Bolded values indicate statistical significance.
 β (95% CI) from univariate linear regression model.

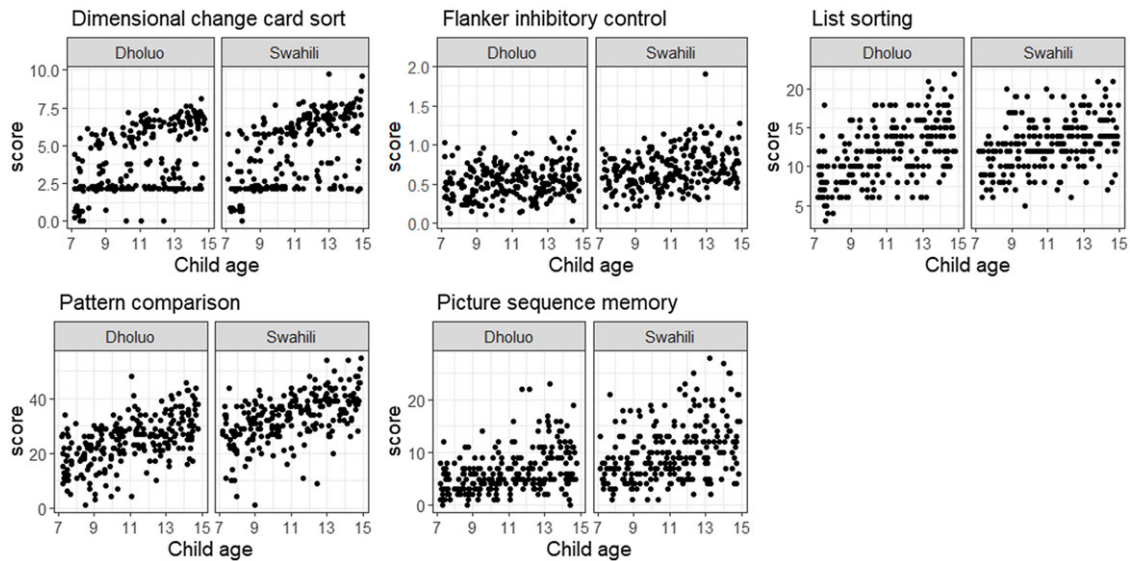


Figure 1. Scatterplot of child's ages and NIH Toolbox® score, by test and language. The Spearman rank-order correlation coefficient is reported for each language group.

PC as related to DCCS and Flanker reached a moderate correlation (Figure 2a). The correlation between PC and either LS or PSM was less moderate; although the sample size allowed for even small correlations to be detected as statistically significant, the magnitude of the PC correlations with LS or PSM did not reach a notable level. Several tests had low levels of correlation. In particular, the Flanker test correlated poorly with PSM and LS.

Further, the Dholuo-language NIH Toolbox® tests also exhibited lower inter-test correlations (Figure 2b). Only PC with DCCS and LS met moderate correlations, though Flanker was only slightly lower. LS with either DCCS OR PSM was also a poor correlation. Flanker continued to show much lower correlations than expected, especially with PSM – the only correlation which was statistically nonsignificant from zero ($r = 0.09$, $p = 0.17$).

Reliability testing

Internal consistency estimates (Cronbach's alphas or split-half correlations) for both the Dholuo- and Swahili-language tests were high, indicating strong reliability for the tests (Table 4). Practice effects within the repeated testing in the Dholuo-speaking population were present across all five tests. On all tests, the participants exhibited significant improvement on the second assessment occasion ($p < 0.001$). For DCCS, participants performed 0.5 points higher at follow-up, which may mean (1) improving accuracy; (2) improving speed; or (3) improving both.

We examined the pattern of results and found that participants who were able to match at least one shape item on the first administration but still had low accuracy were driving this improvement; on the retest, they improved their accuracy by 10 trials on average, sometimes even sufficiently enough to get credit for reaction time/speed. Most participants who failed to match on shape initially also failed on retest. Individuals with high enough accuracy to receive credit for their speed on the initial response had largely unchanged scores on retest. For Flanker, participants increased their number correct per second by 0.08 points on the second administration. For LS, PC, and PSM, participants had 1–2, 4–5, and 2–3 more correct questions on the second administration, respectively (Table 5).

In summary, the ICCs were good for DCCS (ICC = 0.816), moderate for Flanker (0.686), LS (0.587), and PC (0.788), and poor for PSM (0.480) (Appendix B).

Discussion

This study examines the psychometric properties of the NIH Toolbox® African Languages in a cohort of children from western Kenya. The NIH Toolbox® African Languages offers culturally adapted tests of fluid cognition that can provide measurements of development within Swahili- and Dholuo-speaking regions. The tools examined had generally acceptable psychometric properties for research use within Kenya. Given the limited number of cognitive tools psychometrically examined within this context, our study provides valuable insights for the future use of these research tools in Kenya. As the validity and reliability of this tool continues to be explored, there is an opportunity to further refine its acceptability within a diverse setting through shifting focus to further pilot studies and norming data among these Kenya regions.

Our study found that scoring for the Flanker test had to be modified due to general differences in the prioritization of speed and accuracy within this context, given that the children were accurate in their judgments. We also found that issues related to shape identification limited the utility of the DCCS test for many participants, with approximately 25% of participants aged 7–14 years unable to match based on shape.

In our study, we found that altering rate-based scoring for certain timed tests, specifically the Flanker test, ensures appropriate measurement of the targeted construct of inhibitory control and attention. Within the case of the Flanker test, a number of participants did not receive credit for their accurate scores due to the extended time it took to select the answer. Although the administrator of the test prompted participants to answer items as quickly as possible, the average time taken to answer was much slower compared to their North American peers, for whom the original scoring method was developed (Weintraub et al., 2013). When items must be answered within timed limits to receive credit, cultures that prioritize accuracy over speed will be at an unfair disadvantage. The emphasis on speed and timing, which is

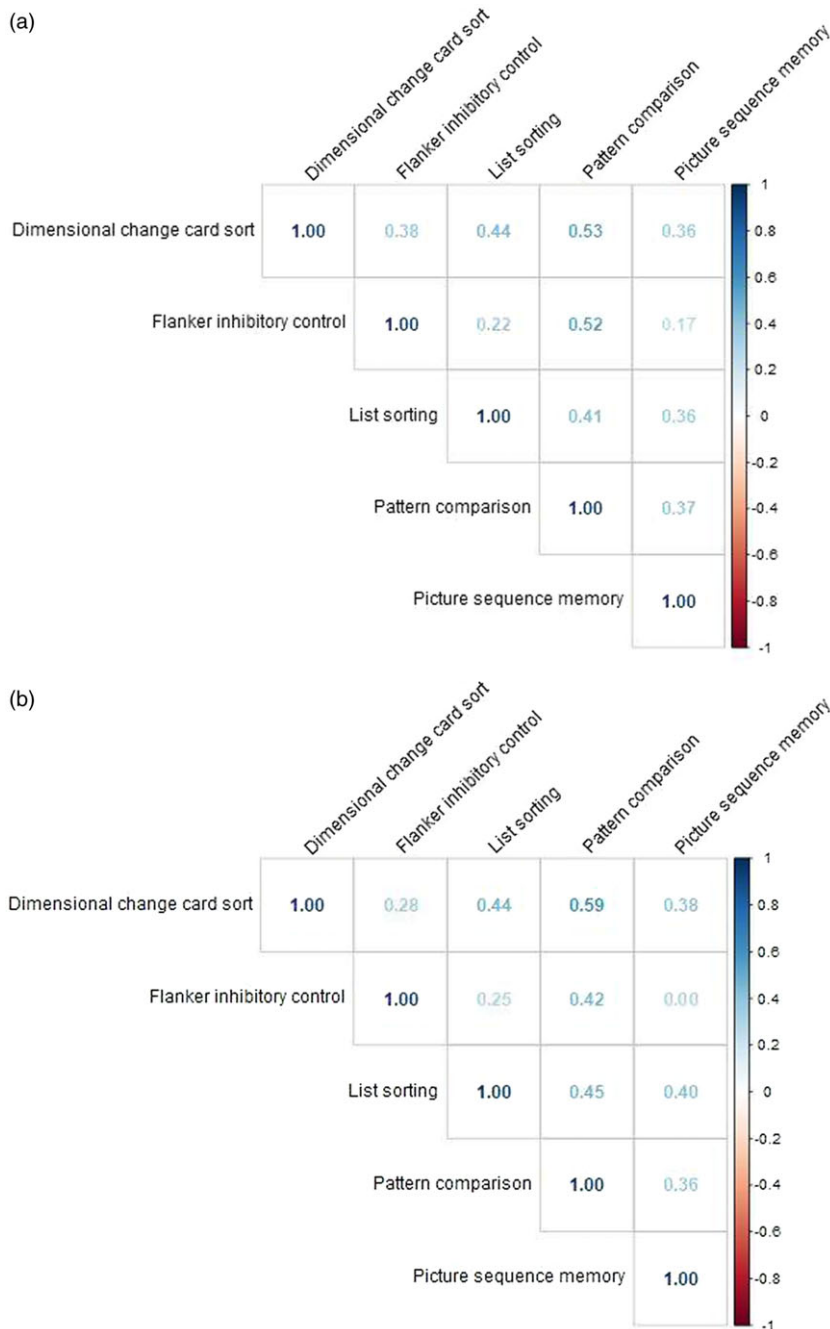


Figure 2. (a) Swahili correlation matrix. (b) Dholuo correlation matrix.

common within Western settings, is not universal across the globe (Ardila, 2005). We developed an R function, provided within Appendix C, which will calculate rate-based scores using the data exports from the NIH Toolbox® African Languages. While our team considered using rate-based scoring for DCCS as well, it was determined to not make a significant difference in scoring for this study, due to overall lower accuracy with that test (primarily due to the inability to match on shapes). While this scoring method was not utilized within this manuscript, we include the ability to calculate rate-based scores for DCCS with the code in Appendix C for settings where accuracy is higher. This code can be used for both DCCS and Flanker data.

Cultural and language differences also presented some challenges when using the DCCS test within Dholuo and

Swahili-speaking populations. When the DCCS test was developed for English-speaking children, it was found to lead to perseverative errors, with children having difficulty switching from shape items to color items (Doebel & Zelazo, 2015; Zelazo, 2006). However, we found the opposite issue within this population. During the cultural adaptation and translation of the DCCS test, the word “shape” was noted to be difficult to translate (Duffey et al., 2022). The word “Umbo” was used for Swahili, but after trialling multiple Dholuo words that approximated the construct of a shape, none were well understood by study participants. The Dholuo-speaking advisors on the project recommended to retain the English word, “Shape,” as English is taught in primary school in Kenya and this word was well known (Duffey et al., 2022). Despite this thoughtful process, both Swahili and Dholuo-speaking participants appeared

Table 4. Internal consistency, by test and language

Test	Internal consistency	
	Dholuo	Swahili
Dimensional Change Card Sort ^a	0.96	0.94
Flanker Inhibitory Control ^a	0.93	0.95
List Sorting ^a	0.88	0.85
Pattern Comparison ^b	0.96	0.97

^aSplit-half reliability was calculated.

^bCronbach's alpha was calculated.

Table 5. Intraclass correlations (ICC) for tests in Dholuo language

	Adjusted ICC	Practice effects ^a	<i>p</i> -Value	Scoring scales
Dimensional Change Card Sort	0.816	0.509	<0.001	0–10
Flanker Inhibitory Control	0.686	0.085	<0.001	0–3
List Sorting	0.587	1.697	<0.001	0–26
Pattern Comparison	0.788	4.502	<0.001	0–130
Picture Sequence Memory	0.480	2.702	<0.001	0–31

^aImprovement in cognitive test performance due to repeated evaluation with the same test materials.

to have difficulty answering items related to shapes. For those who successfully completed the practice items related to shape, they appeared to progress with the remaining test without issue. However, 20–25% of participants were unable to progress past the practice items for shape, demonstrating a critical limitation of the test. Without appropriate understanding of the instructions and key aspects of the test, it will not provide a valid measure for the construct of cognitive flexibility. Thus, we feel that pilot testing of the DCCS test prior to administration for research purposes is particularly critical.

As a result of these cultural challenges in administered DCCS, our outcome data for this test were not normally distributed. This contributed to the weak correlations seen between DCCS and the Flanker tasks. Similar to DCCS, some participants were also unable to progress past the practice items within the Flanker test (four Luo, one Swahili). Due to the challenges in administering both tests, the correlations of scores between them are likely to be negatively impacted. Additionally, considering the rapid brain development that occurs within the first 8 years of a child's life (Center on the Developing Child, 2007), we generally expect to see improvement of performance as age increases. However, it may not be surprising that the associations were not strong, given that our current study excluded children younger than 7. Children have consistently demonstrated improvements in their executive functioning skill in early childhood, but around 7–8 years, the evidence is mixed (Best & Miller, 2010). Although the 7-year-old participants generally performed well on the tests, use of the NIH Toolbox[®] African Languages tests is not recommended in children younger than 7 years due to challenges with language and exposure (Duffey et al., 2022). As children become adolescents and adults, we anticipate scores plateauing, reducing the need to adjust for age within statistical modeling.

Despite the limitations of some individual tests, specifically the Flanker and DCCS, the psychometric properties of the NIH Toolbox[®] African Languages remain adequate for use among Dholou- and Swahili-speaking children. Furthermore, a validation study on the cognitive tests of the original NIH Toolbox (in

English) in Zambian youth showed promise (Kabundula et al., 2022). However, more research and discussion are needed to determine whether further development and refinement of the NIH Toolbox[®] African Languages would be a fruitful pursuit in advancing the methods needed to study cognition in settings of sub-Saharan Africa.

As the NIH Toolbox[®] African Languages is further developed, it will need to address current shortcomings preventing its use as a clinical neuropsychological assessment, such as challenges similar to the English-language version of the NIH Toolbox[®]. While the original validation studies showed strong psychometric properties, there are some populations for which it has had poor-to-adequate construct validity in measuring attention and executive function, episodic memory, and processing speed in adults, for whom it was initially designed (Ott et al., 2022). Within a US sample of children, reliability of the NIH Toolbox[®] was generally low-to-moderate reliability over time, with very few tests having adequate stability to meet research standards, and none having the stability needed for clinical applications (Taylor et al., 2022). A new version of the Toolbox was recently released which aims to address some of these shortcomings, yet more research is necessary to demonstrate whether the modifications were successful. Just as the English-language version is being refined, ongoing development is necessary to maximize the cultural appropriateness and validity of the NIH Toolbox[®] African Languages. In sum, it is not yet clear whether this tool will have clinical utility in the future or whether populations could be compared across countries. However, the NIH Toolbox[®] African Languages shows promise for utility when used to compare similar groups within the same population.

Limitations

This study has a few additional limitations to consider. Each language of the NIH Toolbox[®] was evaluated in different regions of the country and, thus, the sample may not be generalizable to all children and adolescents in Kenya. Furthermore, due to the scope of the study and resources available, we were unable to test other psychometric properties, such as how the tool – considered a brief assessment – would perform compared to a gold standard assessment of cognition or the test–retest reliability within the Swahili-speaking population.

The NIH Toolbox[®] African Languages remains an emerging tool. The original English-version tool continues to garner discretion in its construct validity and reliability among researchers, especially in its tablet-based and newer mobile formats (Brearly et al., 2019; Gershon et al., 2022). Therefore, there are still critical considerations to be made before ascertaining its utility for diverse populations, especially in any clinical or diagnostic contexts. As with any cognitive tool, the quality of a child's education and exposure to multiple languages may influence the results and should be considered, when possible, as a potentially confounding variable within research. This study provides valuable insights on basic psychometric evaluations of a culturally adapted NIH Toolbox[®], which is often lacking in many tools used outside of the USA. As piloting is underway across other diverse settings, such as Nairobi, there may be more opportunities to evaluate psychometric properties of the NIH Toolbox[®] in the future.

The NIH Toolbox[®] African Languages shows unique potential in its applicability as a cognitive tool for underserved populations in Kenya. This pilot study took an important first step in investigating the current Toolbox, but further pilot studies and

modifications to the current version are needed to determine the validity of this tool for clinical assessment. This will require norming data for using a representative sample of Kenyan individuals, as well as critical investigation of the current tool's reliability and long-term utility. The undertaking of these future directions in research will allow the opportunity to apply the NIH Toolbox[®] African Languages for children at risk of not achieving their full developmental potential.

In conclusion, the NIH Toolbox[®] African Languages performs an exploratory, psychometrically valid assessment of fluid cognition and can be used in similar settings as urban or local schools in Kenya. Scoring alterations and pilot testing are recommended to optimally measure outcomes. While the NIH Toolbox[®] African Languages program only utilizes two languages of the 70 living languages spoken in Kenya, it is an important start to expand the science of cognition for African researchers and those working within these settings. It also provides insights on some of the challenges and possibilities for adaptation or creation of future psychometric tools within sub-Saharan Africa. More research is necessary, including replicating these results in additional samples and further evaluations on how demographic and socioeconomic factors impact the test scores.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S1355617723000632>

Acknowledgments. We would like to thank the study participants and their families for engaging in the study. We would like to thank the head teachers, other teachers and staff at the schools with whom we engaged, and Melissa R. Thomas for her careful copy editing.

Funding statement. This project was funded, in part, from the Indiana Clinical and Translational Sciences Institute, and in part by Grant Number UL1TR002529 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. Additionally, Dr. Megan McHenry's salary while working on this project was supported by a career development award funded through the National Institutes of Mental Health [K23MH116808, PI: MSM.] The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests. All eight authors do not have any conflicts of interest or financial conflicts to disclose.

References

- Alcock, K. J., Holding, P. A., Mung'ala-Odera, V., & Newton, C. R. J. C. (2008). Constructing tests of cognitive abilities for schooled and unschooled children. *Journal of Cross-Cultural Psychology*, 39(5), 529–551. <https://doi.org/10.1177/0022022108321176>
- Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review*, 15(4), 185–195. <https://doi.org/10.1007/s11065-005-9180-y>
- Bauer, P. J., Dikmen, S. S., Heaton, R. K., Mungas, D., Slotkin, J., & Beaumont, J. L. (2013). III. NIH Toolbox Cognition Battery (CB): Measuring episodic memory. *Monographs of the Society for Research in Child Development*, 78(4), 34–48. <https://doi.org/10.1111/mono.12033>
- Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, 81(6), 1641–1660. <https://doi.org/10.1111/j.1467-8624.2010.01499.x>
- Brearly, T. W., Rowland, J. A., Martindale, S. L., Shura, R. D., Curry, D., & Taber, K. H. (2019). Comparability of iPad and web-based NIH Toolbox Cognitive Battery administration in veterans. *Archives of Clinical Neuropsychology*, 34(4), 524–530. <https://doi.org/10.1093/arclin/acy070>
- Casaletto, K. B., Umlauf, A., Beaumont, J., Gershon, R., Slotkin, J., Akshoomoff, N., & Heaton, R. K. (2015). Demographically corrected normative standards for the English version of the NIH Toolbox Cognition Battery. *Journal of the International Neuropsychological Society*, 21(5), 378–391. <https://doi.org/10.1017/S1355617715000351>
- Center on the Developing Child (2007). *The Science of Early Childhood Development (INBrief)*. Harvard University. Retrieved 7 April, 2022, from <https://developingchild.harvard.edu/resources/inbrief-science-of-ecd/>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Dikmen, S. S., Bauer, P. J., Weintraub, S., Mungas, D., Slotkin, J., Beaumont, J. L., Gershon, R., Temkin, N. R., & Heaton, R. K. (2014). Measuring episodic memory across the lifespan: NIH Toolbox Picture Sequence Memory Test. *Journal of the International Neuropsychological Society*, 20(6), 611–619. <https://doi.org/10.1017/s1355617714000460>
- Doebel, S., & Zelazo, P. D. (2015). A meta-analysis of the Dimensional Change Card Sort: Implications for developmental theories and the measurement of executive function in children. *Developmental Review*, 38, 241–268. <https://doi.org/10.1016/j.dr.2015.09.001>
- Duffey, M. M., Ayuku, D., Ayodo, G., Abuonji, E., Nyalumbe, M., Giella, A. K., Hook, J. N., Tran, T. M., & McHenry, M. S. (2022). Translation and cultural adaptation of NIH Toolbox cognitive tests into Swahili and Dholuo languages for use in children in western Kenya. *Journal of the International Neuropsychological Society*, 28(4), 414–423. <https://doi.org/10.1017/S1355617721000497>
- Ferrer, E., O'Hare, E. D., & Bunge, S. A. (2009). Fluid reasoning and the developing brain. *Frontiers in Neuroscience*, 3(1), 46–51. <https://doi.org/10.3389/neuro.01.003.2009>
- Gershon, R. C., Fox, R. S., Manly, J. J., Mungas, D. M., Nowinski, C. J., Roney, E. M., & Slotkin, J. (2020). The NIH Toolbox: Overview of development for use with Hispanic populations. *Journal of the International Neuropsychological Society*, 26(6), 567–575. <https://doi.org/10.1017/S1355617720000028>
- Gershon, R., Sliwinski, M., Mangravite, L., King, J., Kaat, A., Weiner, M., & Rentz, D. (2022). The Mobile Toolbox for monitoring cognitive function. *The Lancet Neurology*, 21(7), 589–590. [https://doi.org/10.1016/S1474-4422\(22\)00225-3](https://doi.org/10.1016/S1474-4422(22)00225-3)
- Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH toolbox for assessment of neurological and behavioral function. *Neurology*, 80(11 Suppl 3), S2–S6. <https://doi.org/10.1212/WNL.0b013e3182872e5f>
- Gladstone, M., Lancaster, G. A., Umar, E., Nyirenda, M., Kayira, E., van den Broek, N. R., Smyth, R. L., & Osrin, D. (2010). The Malawi Developmental Assessment Tool (MDAT): The creation, validation, and reliability of a tool to assess child development in rural African settings. *PLOS Medicine*, 7(5), e1000273. <https://doi.org/10.1371/journal.pmed.1000273>
- Hodes, R. J., Insel, T. R., Landis, S. C., & NIH Blueprint for Neuroscience Research (2013). The NIH toolbox: Setting a standard for biomedical research. *Neurology*, 80(11 Suppl 3), S1. <https://doi.org/10.1212/WNL.0b013e3182872e90>
- Holding, P. A., Taylor, H. G., Kazungu, S. D., Mkala, T., Gona, J., Mwamuye, B., Mbonani, L., & Stevenson, J. (2004). Assessing cognitive outcomes in a rural African population: Development of a neuropsychological battery in Kilifi District, Kenya. *Journal of the International Neuropsychological Society*, 10(2), 246–260. <https://doi.org/10.1017/s1355617704102166>
- Kabundula, P. P., Mbewe, E. G., Mwanza-Kabaghe, S., Birbeck, G. L., Mweemba, M., Wang, B., Menon, J. A., Bearden, D. R., & Adams, H. R. (2022). Validation of the National Institute of Health Toolbox Cognition Battery (NIHTB-CB) in children and adolescents with and without HIV infection in Lusaka. *Zambia Aids and Behavior*, 26(10), 3436–3449. <https://doi.org/10.1007/s10461-022-03669-7>
- Liesfeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods*, 51(1), 40–60. <https://doi.org/10.3758/s13428-018-1076-x>
- Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberg, U., & Tucker-Drob, E. M. (2020). Education and cognitive functioning across the life span. *Psychological Science in the Public Interest*, 21(1), 6–41. <https://doi.org/10.1177/1529100620920576>
- McHenry, M. S., Mukherjee, D., Bhavnani, S., Kirolos, A., Piper, J. D., Crespo-Llado, M. M., Gladstone, M. J., & Pant Pai, N. (2023). The current landscape and future of tablet-based cognitive assessments for children in low-

- resourced settings. *PLOS Digital Health*, 2(2), e0000196. <https://doi.org/10.1371/journal.pdig.0000196>
- Mungas, D., Widaman, K., Zelazo, P. D., Tulsy, D., Heaton, R. K., Slotkin, J., Blitz, D. L., & Gershon, R. C. (2013). VII. NIH Toolbox Cognition Battery (CB): Factor structure for 3 to 15 year olds. *Monographs of the Society for Research in Child Development*, 78(4), 103–118. <https://doi.org/10.1111/mono.12037>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Ott, L. R., Schantell, M., Willett, M. P., Johnson, H. J., Eastman, J. A., Okelberry, H. J., Wilson, T. W., Taylor, B. K., & May, P. E. (2022). Construct validity of the NIH toolbox cognitive domains: A comparison with conventional neuropsychological assessments. *Neuropsychology*, 36(5), 468–481. <https://doi.org/10.1037/neu0000813>
- Ozawa, S., Laing, S. K., Higgins, C. R., Yemeke, T. T., Park, C. C., Carlson, R., Ko, Y. E., Guterman, L. B., & Omer, S. B. (2022). Educational and economic returns to cognitive ability in low- and middle-income countries: A systematic review. *World Development*, 149, 105668. <https://doi.org/10.1016/j.worlddev.2021.105668>
- Peng, P., & Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives*, 14(1), 15–20. <https://doi.org/10.1111/cdep.12352>
- Shields, R. H., Kaat, A. J., McKenzie, F. J., Drayton, A., Sansone, S. M., Coleman, J., Michalak, C., Riley, K., Berry-Kravis, E., Gershon, R. C., Widaman, K. F., & Hessel, D. (2020). Validation of the NIH Toolbox Cognitive Battery in intellectual disability. *Neurology*, 94(12), e1229–e1240. <https://doi.org/10.1212/WNL.00000000000009131>
- Taylor, B. K., Frenzel, M. R., Eastman, J. A., Wiesman, A. I., Wang, Y.-P., Calhoun, V. D., Stephen, J. M., & Wilson, T. W. (2022). Reliability of the NIH toolbox cognitive battery in children and adolescents: A 3-year longitudinal examination. *Psychological Medicine*, 52(9), 1718–1727. <https://doi.org/10.1017/s0033291720003487>
- UNICEF (2015). *Children in Africa: Key statistics on child survival, protection and development*. Author.
- Victorson, D., Manly, J., Wallner-Allen, K., Fox, N., Purnell, C., Hendrie, H., Havlik, R., Harniss, M., Magasi, S., Correia, 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(11 Suppl 3), S13–S19. <https://doi.org/10.1212/WNL.0b013e3182872e26>
- Weintraub, S., Bauer, P. J., Zelazo, P. D., Wallner-Allen, K., Dikmen, S. S., Heaton, R. K., Tulsy, D. S., Slotkin, J., Blitz, D. L., Carlozzi, N. E., Havlik, R. J., Beaumont, J. L., Mungas, D., Manly, J. J., Borosh, B. G., Nowinski, C. J., & Gershon, R. C. (2013). I. NIH Toolbox Cognition Battery (CB): Introduction and pediatric data. *Monographs of the Society for Research in Child Development*, 78(4), 1–15. <https://doi.org/10.1111/mono.12031>
- Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2010). The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment*, 22(2), 306–317. <https://doi.org/10.1037/a0018708>
- Willoughby, M. T., Blair, C. B., Wirth, R. J., & Greenberg, M. (2012). The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment*, 24(1), 226–239. <https://doi.org/10.1037/a0025361>
- Willoughby, M. T., Piper, B., Oyanga, A., & Merseth King, K. (2019). Measuring executive function skills in young children in Kenya: Associations with school readiness. *Developmental Science*, 22(5), e12818. <https://doi.org/10.1111/desc.12818>
- Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U., & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology*, 33(3), 292–296. <https://doi.org/10.1037/a0030727>
- Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, 1(1), 297–301. <https://doi.org/10.1038/nprot.2006.46>