Translation and cultural adaptation of NIH Toolbox cognitive tests into Swahili and Dholuo languages for use in children in western Kenya

Megan Marie Duffey, MD[1,2], **David Ayuku, PhD**[3,4], **George Ayodo, PhD**[5,6], **Emily Abuonji, BS**[5], **Mark Nyalumbe, BS**[3], **Amy Kovacs Giella, BA**[7], **Julie N Hook, PhD, MBA**[7], **Tuan M Tran, MD, PhD**[1,2], **Megan S McHenry, MD, MS**[1,4]

[1]Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN, USA

[2]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA

[3]Moi University School of Medicine, Department of Mental Health & Behavioral Sciences, Eldoret, Kenya

[4]Academic Model Providing Access to Healthcare (AMPATH), Eldoret, Kenya

[5]Kenya Medical Research Institute, Centre of Global Health Research, Kisumu, Kenya

[6]Jaramogi Oginga Odinga University of Science and Technology, Bondo, Kenya

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

Abstract

Objective: Performing high-quality and reliable cognitive testing requires significant resources and training. As a result, large-scale studies involving cognitive testing are difficult to perform in low- and middle-income settings, limiting access to critical knowledge to improve academic achievement and economic production in these populations. The NIH Toolbox® is a collection of cognitive, motor, sensory, and emotional tests that can be administered and scored using an iPad® tablet, reducing the need for training and quality monitoring; and thus, it is a potential solution to this problem.

Method: We describe our process for translation and cultural adaptation of the existing NIH Toolbox tests of fluid cognition into the Swahili and Dholuo languages for use in children aged 3–14 in western Kenya. Through serial forward and back-translations, cognitive interviews, group consensus, outside feedback, and support from the NIH Toolbox team, we produced translated tests that have both face validity and linguistic validation.

Results: During our cognitive interviews, we found that the five chosen tests (one each of attention, cognitive flexibility, working memory, episodic memory, and processing speed) were

Corresponding Author Megan S. McHenry, MD, MS, FAAP, Assistant Professor of Pediatrics, Division of Pediatric Infectious Diseases and Global Health, Indiana University School of Medicine, 705 Riley Hospital Drive, Room 5853, Indianapolis, IN 46202, 317.274.4143 tel, 317.948.0860 fax, msuhl@iu.edu.

Declaration of Conflicting Interests
The authors declare that there is no conflict of interest.

generally well-understood by children aged 7–14 in our chosen populations. The cognitive interviews informed alterations in translation as well as slight changes in some images to culturally adapt the tests.

Conclusions: This study describes the process by which we translated five fluid cognition tests from the NIH Toolbox into the Swahili and Dholuo languages. The finished testing application will be available for future studies, including a pilot study for assessment of psychometric properties.

Keywords

Academic Success; Child; Cognition; Developing countries; Language; Linguistics

Introduction

Cognitive development in childhood lays the foundation for academic achievement and economic production for a lifetime (Blair & Razza, 2007; Eigsti et al., 2006; Moffitt et al., 2011; Riggins, Miller, Bauer, Georgieff, & Nelson, 2009; Rose, Feldman, Jankowski, & Van Rossem, 2008). Worldwide, nearly 250 million children are at risk for not meeting their full developmental potential (Black et al., 2017). As we seek to reduce risk factors and develop interventions for children globally who are at risk for poor development, the measurement of cognition becomes critically important (Anguera et al., 2017; Bei, Oiberman, Teisseire, & Barres, 2018; Lambez, Harwood-Gross, Golumbic, & Rassovsky, 2020; Sherr, Croome, Bradshaw, & Parra Castaneda, 2014; Tusing & Ford, 2004; Vanderwood, McGrew, Flanagan, & Keith, 2002).

Multiple challenges exist when measuring cognitive development within global settings. Performing valid, reliable, and clinically relevant cognitive assessments often requires substantial training, prerequisite credentials, and monitoring to ensure consistent, high-quality administration (Miles, Fulbrook, & Mainwaring-Mägi, 2016). Additionally, while some tests are moving towards electronic scores, most are still scored manually which introduces human error. Furthermore, substantial heterogeneity exists regarding the specific types of assessments, which results in challenges when interpreting results across studies, especially when the quality of each assessment is unknown. In addition, many testing items are not contextually relevant within cross-cultural settings. Nearly all assessments used in resource-limited settings are ones that were developed in resource-rich settings, such as the Wechsler Adult Intelligence Scale-Third Edition (Wechsler, 1997), the Wechsler Intelligence Scale for Children-Fifth Edition (Wechsler, 2014), the Bayley Scales of Infant and Toddler Development-Third Edition (Bayley, 2006), and the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983). In some cases, these assessments are appropriately translated and adapted to the culture in which they are used (McHenry et al., 2021; Pendergast et al., 2018), and in some cases they have been used in countries such as South Africa in which English is an acquired language in school (Cockcroft, Alloway, Copello, & Milligan, 2015; Skuy, Taylor, O'Carroll, Fridjhon, & Rosenthal, 2000). However, in many cases, there is no evidence that the assessments or translations were determined to be appropriate for the setting, limiting the validity of the results (McHenry et al., 2018).

The National Institutes of Health Toolbox for Assessment of Neurological and Behavioral Function (NIH Toolbox®) is a potential solution to the many challenges faced when measuring neurodevelopment, such as cognition, emotional, motor, and sensory domains, in resource-limited settings. The NIH Toolbox was developed by more than 250 scientists over six years and was sponsored by the 15 institutes of the National Institutes of Health that comprise the NIH Neuroscience Blueprint (Hodes, Insel, & Landis, 2013). A major advantage of the NIH Toolbox is that all measures are administered through an iPad® tablet-based application, requiring light training to administer the evaluations and digital scoring. This opens a myriad of possibilities for cognitive testing. However, as is true in conventional cognitive testing, the items in the NIH Toolbox may not be culturally relevant everywhere. This study is one of the first to describe the process by which some of the NIH Toolbox tests were culturally adapted for administration to children outside the United States, specifically for those children in Kenya. While the NIH Toolbox includes over 100 measures of cognitive, emotional, sensory, and motor domains, our study will focus on five fluid cognitive measures which test attention, cognitive flexibility, working memory, episodic memory, and processing speed.

Method

Setting/Participants

This study was performed in two study sites in western Kenya from September 2019-June 2020: Eldoret (an urban setting) and Ajigo (a rural setting). Participants aged 3–14 were recruited for cognitive interviews using refined translations of the five fluid cognitive measures in the NIH Toolbox Cognitive battery. Participants were recruited from primary schools in the area where the Swahili and Dholuo languages are primarily spoken; the municipality of Eldoret, Kenya was chosen for Swahili and the Ajigo ward in the county of Siaya, Kenya was chosen for Dholuo. Inclusion criteria were as follows: (1) between 3–14 years of age; (2) fluently spoke the specific language of interest for that study site (Swahili for Eldoret, Dholuo for Ajigo); (3) and a primary caregiver who spoke either Swahili or Dholuo who was available for consent and completion of a questionnaire. Three participants for each age in years were recruited and enrolled in the study for each language. Gender and handedness demographic data were collected on each participant.

Overview of the NIH Toolbox for Assessment of Neurological and Behavioral Function

The NIH Toolbox is a set of standardized, validated, 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. Tests have been validated for persons aged 3–85 years (Mungas et al., 2013; Weintraub et al., 2013) and have been translated into Spanish (Gershon et al., 2020; Victorson et al., 2013). The NIH Toolbox Cognition Battery was normed using a U.S. 2010 Census-matched normative sample of 1,020 typically-developing children and adolescents aged 3–20 years (Akshoomoff et al., 2014).

Five of the seven core cognitive tests available on the NIH Toolbox were chosen to use with our population: Flanker Inhibitory Control and Attention Test (Flanker), Dimensional

Change Card Sort Test (DCCS), Picture Sequence Memory Test (PSM), Pattern Comparison Processing Speed Test (PC), and List Sorting Working Memory Test (LSWM), which are depicted in Figure 1. The selected tests measure fluid cognitive abilities which assess the individual's ability to problem solve and process new information. These cognitive abilities are more subject to change with biological function. In contrast, the remaining two core cognition tests, Picture Vocabulary and Oral Reading, measure crystallized cognitive abilities which rely more heavily on previous learning. During the original validation studies for PC and LSWM, 18 children aged 3–5 years and nine children aged 3–6 years, respectively, out of 120 total children could not complete the task due to lack of attention and noncompliance (Carlozzi, Tulsky, Kail, & Beaumont, 2013; Tulsky et al., 2013; Weintraub et al., 2013). As such, in this study all children completed the Flanker, DCCS, and PSM tests, and only children aged 7 and above completed the PC and LSWM tests.

Selected NIH Toolbox Cognitive Tests

The Flanker and DCCS tests measure different aspects of executive function. The Flanker test measures attention allocation, inhibitory control, and mental flexibility. It requires participants to choose the direction of a target picture that is surrounded or flanked by similar images which may be facing the same or different direction. In the NIH Toolbox Flanker test, directional fish and then arrows are used for children ages 3–7 years, and arrows for children 8 years and older. For this test, participants start with their dominant index finger on "home base." Home base is a printed circle with iPad positioning markers and is provided by the test publisher. It is laid in front of the iPad and participants are asked to return their dominant index finger to the home base after answering each item. Because timing is used as a factor in scoring, it standardizes the distance for every participant when touching the screen.

DCCS assesses cognitive flexibility and set-shifting. Participants are presented with three images: one is the target and the other two provide sorting options. The participant then selects the appropriate sorting option, in this case color or shape. To standardize response times used in scoring, home base is also used between items.

PSM measures episodic memory. During this task, the participant is presented with a series of images related to a story or action. Each image is presented in sequential order with an auditory cue. Next, they are scrambled, and the participant has to order them appropriately. LSWM measures the participant's ability to remember a series of objects presented visually with verbal cues. The participant then needs to mentally re-arrange these objects following specific criteria and repeat them back to the examiner.

PC examines processing speed by having participants indicate whether two side-by-side images are the same or different. Each of these items is timed, which factors into the scoring of the test.

Translation and Cultural Adaptation

Figure 2 illustrates our method of translation and cultural adaptation. Bilingual members of the Kenyan study team provided the first forward translations of the tests, and a back-

translation was performed by a separate bilingual individual who did not have access to the original English text. Several iterations of forward and back-translations were performed until all site teams and the NIH Toolbox team, from Northwestern University, agreed on the language and a preliminary finalized version of the text was generated. Cognitive interviews were then performed in Eldoret and Ajigo, the results of which were qualitatively discussed in a workshop held with both the Swahili and Dholuo teams, led by each Kenyan site Principal Investigator (both PhD-level researchers, one of whom is an academic clinical psychologist), the NIH Toolbox team, and the study Principal Investigator (MSM). During this workshop, participant feedback gathered from the cognitive interviews was discussed with the NIH Toolbox team and used to improve the forward translations. The NIH Toolbox team also consulted with neuropsychologists familiar with the cognitive tests regarding proposed alterations in test images. Three more iterations of forward and back-translations were performed, and then a preliminary version of the iPad application was created with translated text and audio. Fine tuning and harmonization were guided by a trilingual (i.e., fluent in Swahili, Dholuo, and English) member of the study team (EA). Two additional reviewers from central and coastal Kenya (Nairobi and Kilifi) also provided feedback on understandability of the proposed Swahili translations in their respective regions prior to their finalization in the iPad application.

Cognitive Interviews

Cognitive interviewing was used to better understand how particular words and questions were interpreted with our first set of finalized translations. Two common methods of cognitive interviewing include "think-aloud interviewing," where the participant is encouraged to speak in a stream of consciousness, and "verbal probing," where the participant is asked discrete follow-up questions (Willis & Artino, 2013). Frequently, these modalities are used in conjunction. However, due to the young age of the study participants, verbal probing was the primary technique used. The objectives of our cognitive interviews were to improve linguistic validation and face validity. The Eldoret study team received face-to-face training sessions on cognitive interviewing and the Ajigo study team received similar training via video conferencing by a research assistant (MMD). Content of the training sessions included familiarity with the tests and cognitive interview guides, objectives of cognitive interviewing, and building rapport and interacting with a child during a cognitive interview. Interviewers used mostly proactive with some reactive probes, an approach best suited for our pediatric population. Proactive probes were read from a cognitive interview guide, for example, participants were asked to name the shape of several objects in the DCCS test. If a child did not report the intended answer, the examiner could rephrase the question to ask, "what is this," or "what object is this," in order to understand if the child was unable to identify the object, or if the word chosen to represent "shape" did not make sense.

The cognitive interviews were administered onsite in each school by members of the study teams. In order to minimize variability, all cognitive interviews were performed by one person in Ajigo and two people in Eldoret. An additional member of the study team was also present to help fill out the cognitive interview guides and organize the iPad and testing papers.

Abridged versions of the tests were used in order to streamline administration time while still answering our questions about whether the translated language and images were understandable and the ability of the child to understand and perform the test's intended function. For ease of use, some tests were comprised of screenshots of the test printed on paper, while others were acceptable to administer with the iPad. Administration of a limited number of tests using the iPad was performed to determine whether children could understand the iPad's touch screen function without prior familiarity. The tests were administered in the same order for every child, and the cognitive interviews lasted approximately 40–60 minutes. Of note, each test on the iPad is designed to be given in ten minutes or fewer, and this extra time was comprised of the actual cognitive interviewing, as well as breaks upon request.

Analysis

The qualitative results of the cognitive interviews were discussed amongst members of the research group, including members of the NIH Toolbox team, multiple times as revisions were made by iterative process. Adjustments to language translations were made by group consensus comprised of members of the study team at each site, each site Principal Investigator, the NIH Toolbox team, and the study Principal Investigator. While the emphasis was placed on preserving as much of the original tests as possible, some changes were required and were at the ultimate discretion of the NIH Toolbox Team.

Ethical Approval

This study was completed in accordance with the Helsinki declaration and was approved as an expedited study by the Indiana University Institutional Review Board (Protocol 1904295419) in the United States. In Kenya, it was approved by both Moi University in Eldoret (IREC/2019/15) and the Kenya Medical Research Institute (KEMRI) at the Ajigo site (KEMRI/SERU/CGHR/177/3768). Prior to cognitive interviews, written consent was obtained from parents and verbal assent was obtained from participants 13 years and older.

Results

All 72 children completed the cognitive interviews. Demographic data for the study participants are depicted in Table 1. About half of the participants, 46% (n=33), were female and the majority, 99% (n=71), were right-handed.

Home Base

For purposes of this study, the English phrase "home base" was translated literally to "shelter" or "home" in Swahili, and "starting point" in Dholuo. These translations were chosen to convey the same concept. Flanker was the first test administered that used the home base. A few (n=6) children aged 3–12 in both the Swahili and Dholuo versions did not initially understand the purpose of home base, and the directions had to be repeated several times.

Flanker Inhibitory Control and Attention Test (Flanker)

Overall, the test was generally well-understood by children aged 7 and older. All of the children aged 7 were able to transition from fish to arrows without difficulty, and the test for children aged 8 and above only contained arrows. Children in the 3–6 age range (n=5) had more issues identifying the middle fish/arrow itself, as well as determining which direction it was pointing. Some (n=5) children aged 7–14 in the Dholuo version had trouble identifying the middle arrow as well.

To ease administration, the test for ages 8+ was given on the iPad, and the test for ages 3–7 was given with printed screenshots of the test on paper. According to two interviewers, the test was more difficult to administer on the iPad for the older children because many children had not seen one before and it was distracting.

Most children required 0–1 reminders to return their finger to home base. Additionally, one Swahili-speaking child used the non-dominant hand frequently while other children used their dominant index finger as directed.

Dimensional Change Card Sort Test (DCCS)

Overall, in both the Swahili and Dholuo versions, matching cards by shape was more difficult for children to understand than matching by color, particularly for children under age 7. Some children (n=14) interpreted "shape" as a geometric shape such as a triangle and did not extend that definition to include the shape/form of objects such as a boat. However, despite this different interpretation, most children aged 7 and older were still able to complete the test. Some children (n=9) were unable to match by shape at all, and instead matched by color for all items. Some children (n=9) were able to match by shape and color during the practice items but were unable to perform "task switching" at the end with mixed items and instead matched by color for all items. A few children (n=4) identified colors incorrectly but were still able to match by color.

Most children aged 7 and older needed 0–2 reminders to return their finger to home base with the exception of one child aged 12 and one aged 7, who needed a reminder for every item. Younger children on average needed more reminders, and five children aged 3–4 needed a reminder for every item. One Swahili-speaking child aged 5 named colors in English and one Dholuo-speaking child aged 5 named some shapes in English. In the Dholuo version, it was very difficult to find a Dholuo word for "shape" that conveyed the construct that we link to this term in English-speaking cultures. The English term "shape" is generally known in Dholuo-speaking populations, so ultimately the English word was retained.

Picture Sequence Memory Test (PSM)

Overall, PSM was generally well-understood in participants aged 7 and older. The cognitive interviews did not reveal any major challenges in administration in either language, thus altering the images was deemed unnecessary for accurate administration of the test. However, some adjustments of the translations were needed. For instance, in both languages we replaced some English words with more descriptive words or phrases, such

as exchanging "clown" with "entertainer," and "ride" with "merry-go-round." Occasionally, the translations changed the meaning of what was happening in the picture, for example, the English "watch the tractor pull" was ultimately translated to the Swahili "watch the tractor being pulled," and the Dholuo "someone is watching the tractor pulling." For the Dholuo version in particular, all of the participants (n=39) interpreted the descriptions as commands and reported it would make more sense if they were phrased as "someone is [doing this action]" instead.

Pattern Comparison Processing Speed Test (PC)

Per the validation data described in the methods section, this test was only administered to children aged 7 and older. PC items were shown to children with printed screenshots on paper for the cognitive interviews. Children occasionally tried to use their non-dominant finger to answer. However, instructions were generally well-understood and no major changes to the translations were made.

List Sorting Working Memory Test (LSWM)

This test was generally well-understood, but there were some issues with item recognition. Of the 36 total items (20 animals and 16 foods) shown in the test, eight were ultimately altered (two animals and six foods). One image's color was changed in order to better reflect the color of this fruit normally found in Kenya (a red apple became a green apple), five images were replaced with something of a similar size that is more familiar to Kenya (e.g., "bear" was replaced with "hippo"), and two images were retained but renamed (e.g., the image for "peach" was renamed "mango"). Figure 3 depicts two practice items that were substituted for something of a similar size that is more familiar to Kenya. The criteria in choosing an acceptable substitution included (1) something familiar to Kenya, and (2) retained its placement in each series in which it was included. For instance, we were able to substitute a cheetah for a tiger although a cheetah is smaller, because its placement in all of the testing series was preserved. We also allowed for extra terms to be used to identify some images (e.g., "burger" or "cake" was also accepted for "hamburger.") Upon discussion with the group, we also accepted English terms.

Discussion

This study is the first to describe the process for translating cognitive tests of the NIH-Toolbox beyond its English and Spanish versions. Through multiple rounds of forward and back-translations and cognitive interviews, we learned that some language had to be adjusted from its original English version in order to convey the intended purpose of the test. We also discovered that some concepts, such as colors, shapes, and identifying the "middle" item out of a group of items were more challenging for children aged 3–6 years old. Additionally, it was more challenging for them to return their finger to the home base. Thus, for this population, we recommend using these tests for children 7 years and older. We believe that these tests will be valuable in future research studies, including a pilot study to assess psychometric properties of these tests when formally administered in this population.

We feel that the linguistic validation and face validity of this translation of select NIH Toolbox tests were strengthened by our use of cognitive interviews. This allowed us to determine in more detail the child's line of thinking when choosing an answer, and where it broke down if the wrong answer was chosen. Our cognitive interviews employed mostly proactive verbal probing over think-aloud interviewing due to the age of the study participants (Willis & Artino, 2013). At times, a child's unclear answer could be subject to reactive probing questions in order for the interviewer to better understand the child. This method has some disadvantages; for instance, verbal probing can potentially "lead the subject" depending on how the question is worded or asked. Furthermore, an interviewer asking questions to a child may lead to unclear answers or lead the child to answer in the affirmative when asked if something makes sense, in order to preserve social hierarchy. For example, during item identification portions of our interviews, some younger children responded to every question in the affirmative, even when the examiner ultimately felt that the child did not actually understand. In this situation, it is important to build rapport with the participant before and during testing, and probe more purposefully when there are concerns regarding the participant's response. Regardless of these issues, the cognitive interviews did bring to light several issues in translation and unfamiliar items that were addressed to optimize the tests.

Many prior studies have also utilized cognitive interviews to further refine translations (Fregnani et al., 2017; Huang et al., 2012; Marangu et al., 2017; Masquillier, Wouters, Loos, & Nöstlinger, 2012). Within East Africa, cognitive interviews have aided in Swahili translation and interpretation of questions on a tuberculosis-related stigma questionnaire (Marangu et al., 2017) and helped find contextual differences in the Dholuo, Luganda, and Eastern African English translations of a health-related quality of life questionnaire, which necessitated revision (Masquillier et al., 2012). Within the latter study, cognitive interviews revealed that adolescents did not understand the concept of "free time," and many interpreted the word "satisfied" to be in relation to food and fullness. These important cultural differences may have been missed if cognitive interviews had not been conducted, and cognitive interviewing should be an integral part of high-quality translation of cognitive assessments.

In addition to cognitive interviews, many prior studies include a combination of forward and back-translations and committee consensus (Bonomi et al., 1996; Cella et al., 1998; Smit, Van den Berg, Bekker, Seedat, & Stein, 2006). We believe that our method produced a test with both linguistic validation and face validity by placing emphasis on functional equivalence, allowing us to convey the same message across languages while keeping the original English text intact. Other studies have used the "decentering" method in order to strengthen their translations (Cella et al., 1998; Lent, Hahn, Eremenco, Webster, & Cella, 1999; Smit et al., 2006). This method can only be performed on a source test or questionnaire that has not been finalized, as it involves adjusting the language of the source text in order to align both the actual text as well as the conveyed meaning between the source and target languages (Sechrest, Fay, & Zaidi, 1972). Because the NIH Toolbox has already been validated and normed in English and Spanish in the United States, this was not an appropriate option for our study. Decentering can be a viable option for an instrument that will only be available in two languages, but with expansion of available languages

comes increased complexity (Harkness & Schoua-Glusberg, 1998). For the NIH Toolbox, it was not practical to employ this method since the eventual goal is the availability of this application for a multitude of languages. Rather, English was chosen as the carrier language, from which all current and future translations will stem.

Cultural Considerations

While processing feedback from our cognitive interviews, it was clear that each test required different skillsets, and thus also required differing levels of adaptation. Ideally, every aspect of each test would be culturally adapted to the population in question, but each successive adaptation increases the difference from the validated and normed test in the source language. For example, the PSM test, used to assess episodic memory, requires a participant to recall in the same order a series of visual and audio stimuli, while LSWM, used to assess working memory, requires the participant to process visual and audio stimuli and reorder them mentally. Working memory is used for complex cognitive tasks and often requires manipulation of the memories for daily activities (Tulsky et al., 2013), whereas episodic memory does not (Bauer et al., 2013). Thus, it is much easier to utilize one's working memory with information or stimuli that are more familiar. Having psychologists involved in the adaptation of items is essential to inform the degree to which cultural context may impact the construct being measured, and a considerable amount of time was spent in correspondence with the NIH Toolbox team on this subject.

It is also important to keep in mind the cultural differences that exist in the construct of learning. For instance, mothers in Japan and the United States interact with their 5-month old infants differently according to their cultural constructs (Bornstein, Miyake, & Tamis-Lemonda, 1987), and mothers who are Puerto Rican teach their infants differently from mothers who are middle-class white Americans of non-Hispanic European ancestry according to their cultural importance on different tasks (Harwood, Schoelmerich, Schulze, & Gonzalez, 1999).

It is vital to keep this cultural framework in mind, as Kenyan cultural values and system of learning may differ from elsewhere. For instance, the emphasis on speed in testing in the United States is not shared in many other cultures (Ardila, 2005), and this may affect tests whose scores rely partly on speed. In addition, the cultures in Kenya are viewed to be more collectivistic and less individualistic compared to American culture (Ma & Schoeneman, 1997; Oyserman, Coon, & Kemmelmeier, 2002). East Asian cultures are also more collectivistic and engage in more holistic processing, and it has previously been demonstrated that East Asians are likely to focus more holistically on an image as opposed to Americans who focus more on the central main object (Park & Huang, 2010). This difference in processing could affect tests such as Flanker, which require one to focus on a central object. Furthermore, many of the children tested in our cognitive interviews interpreted the word "shape" in the DCCS test as a word to describe a geometric shape such as a triangle instead of accepting a broader definition, as to describe a boat. Upon much deliberation and discussion with lead study investigators fluent in both languages, we feel that this was not a translation issue, but instead a difference of cultures, in which this particular concept is taught more concretely in Kenya. Because of these integral differences

between cultures, this adaptation should only be used to compare participants within the same region with one another, such as within a case-control or randomized controlled study. As such, it is not possible to diagnose a delay in any of the tested domains at this time. Large-scale validation studies are needed to obtain appropriate norming data for this population.

Since its conception, the NIH Toolbox team has recognized that allowing their tests to be accessible to a wide range of ages, ethnicities, and cultures would require the input of experts, and thus several committees were formed during the development phase. The Cultural Working Group and Spanish Language Working Group were formed to make improvement recommendations in their areas of expertise according to literature in the field and expert opinion (Victorson et al., 2013). During the initial Spanish translation, some of the cognitive tests available on the NIH Toolbox, such as the Picture Vocabulary Test and Oral Reading Recognition Test, required a complete overhaul due to the dependence on language. However, for the fluid cognition tests, translations were proposed and reviewed by the Spanish Language Working Group and recommendations for improvement were made as needed (Gershon et al., 2020).

Limitations

This study had several limitations, one being our use of feedback from only Kenyan Swahili speakers to refine our translations. The Swahili dialect spoken on the coasts of Kenya and Tanzania where the Swahili language originated is referred to as "standard Swahili," while different dialects of Swahili are spoken elsewhere (Duran, 1979). Thus, it is unclear whether these tests could be used in other Swahili-speaking areas such as Tanzania or Uganda, as these different dialects do not necessarily adhere to all of the grammatical rules of standard Swahili. Within Kenya, we addressed these concerns by having the translations reviewed by Swahili speakers living in central and coastal Kenya. However, the use of these tests may be challenging in other countries where Swahili is spoken. In addition, another limitation is the absence of a pilot study in order to verify the tests' psychometric properties. The scope of this study was limited to a detailed description of the methods for adaptation of the NIH Toolbox cognitive tests. We are currently conducting pilot studies to determine the psychometric properties of these adapted tests and determine if our translations are functional and lead to an accurate assessment of these domains. As such, we realize that findings from our ongoing study may necessitate further revisions to the content or administration of our translated tests. Lastly, to better acclimate participants with the iPad prior to testing, we have also translated the touch screen tutorial, which we recommend study participants complete prior to test initiation.

Conclusions

This study reports the process of culturally adapting and translating some of the first translations of NIH Toolbox cognitive tests for use outside of the United States. Our study utilized serial forward and back-translations, cognitive interviews, group and expert consensus, and feedback from sources living in outside areas in order to produce tests of fluid cognition that had face validity and linguistic validation. Because of this work, five

culturally-adapted fluid cognition tests in Dholuo and Swahili now exist for use in future studies.

Acknowledgments

References

Akshoomoff N, Newman E, Thompson WK, McCabe C, Bloss CS, Chang L, … Jernigan TL (2014). The NIH toolbox cognition battery: Results from a large normative developmental sample (PING). In (Vol. 28, pp. 1–10).

Anguera JA, Brandes-Aitken AN, Antovich AD, Rolle CE, Desai SS, & Marco EJ (2017). A pilot study to determine the feasibility of enhancing cognitive abilities in children with sensory processing dysfunction. PLOS ONE, 12(4), e0172616. doi:10.1371/journal.pone.0172616 [PubMed: 28380008]

Ardila A (2005). Cultural Values Underlying Psychometric Cognitive Testing. Neuropsychology Review, 15(4), 185. doi:10.1007/s11065-005-9180-y [PubMed: 16395623]

Bauer PJ, Dikmen SS, Heaton RK, Mungas D, Slotkin J, & Beaumont JL (2013). III. NIH Toolbox Cognition Battery (CB): measuring episodic memory. Monogr Soc Res Child Dev, 78(4), 34–48. doi:10.1111/mono.12033 [PubMed: 23952201]

Bayley N (2006). Bayley Scales of Infant and Toddler Development—Third Edition. In. San Antonio, TX: Harcourt Assessment.

Bei EI, Oiberman A, Teisseire D, & Barres J (2018). Strategies of blind children to achieve cognitive development. A qualitative study. Arch Argent Pediatr, 116(3), e378–e384. doi:10.5546/aap.2018.eng.e378 [PubMed: 29756709]

Black MM, Walker SP, Fernald LCH, Andersen CT, DiGirolamo AM, Lu C, … Lancet Early Childhood Development Series Steering, C. (2017). Early childhood development coming of age: science through the life course. Lancet (London, England), 389(10064), 77–90. doi:10.1016/S0140-6736(16)31389-7

Blair C, & Razza RP (2007). Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten. Child Development, 78(2), 647–663. Retrieved from www.jstor.org/stable/4139250 [PubMed: 17381795]

Bonomi AE, Cella DF, Hahn EA, Bjordal K, Sperner-Unterweger B, Gangeri L, … Zittoun R (1996). Multilingual Translation of the Functional Assessment of Cancer Therapy (FACT) Quality of Life Measurement System. Quality of Life Research, 5(3), 309–320. Retrieved from http://www.jstor.org/stable/4034377 [PubMed: 8763799]

Bornstein MH, Miyake K, & Tamis-Lemonda C (1987). A cross-national study of mother and infant activities and interactions: Some preliminary comparisons between Japan and the United States.

RESEARCH AND CLINICAL CENTER FOR CHILD DEVELOPMENT Annual Report, 9, 1–12.

Carlozzi NE, Tulsky DS, Kail RV, & Beaumont JL (2013). VI. NIH Toolbox Cognition Battery (CB): measuring processing speed. Monogr Soc Res Child Dev, 78(4), 88–102. doi:10.1111/mono.12036 [PubMed: 23952204]

Cella D, Hernandez L, Bonomi AE, Corona M, Vaquero M, Shiomoto G, & Baez L (1998). Spanish Language Translation and Initial Validation of the Functional Assessment of Cancer Therapy Quality-of-Life Instrument. Medical Care, 36(9), 1407–1418. Retrieved from http://www.jstor.org/stable/3767502 [PubMed: 9749663]

Cockcroft K, Alloway T, Copello E, & Milligan R (2015). A cross-cultural comparison between South African and British students on the Wechsler Adult Intelligence Scales Third Edition (WAIS-III). Frontiers in psychology, 6, 297–297. doi:10.3389/fpsyg.2015.00297 [PubMed: 25821443]

Duran JJ (1979). Non-standard forms of Swahili in west-central Kenya. Readings in creole studies, 2.

Eigsti I-M, Zayas V, Mischel W, Shoda Y, Ayduk O, Dadlani MB, … Casey BJ (2006). Predicting Cognitive Control from Preschool to Late Adolescence and Young Adulthood. Psychological Science, 17(6), 478–484. Retrieved from www.jstor.org/stable/40064397 [PubMed: 16771797]

Fregnani CMS, Fregnani JHTG, Paiva CE, Barroso EM, Camargos M. G. d., Tsunoda AT, … Paiva BSR (2017). Translation and cultural adaptation of the Functional Assessment of Chronic Illness Therapy - Cervical Dysplasia (FACIT-CD) to evaluate quality of life in women with cervical intraepithelial neoplasia. Einstein (Sao Paulo, Brazil), 15(2), 155–161. doi:10.1590/S1679-45082017AO3910

Gershon RC, Fox RS, Manly JJ, Mungas DM, Nowinski CJ, Roney EM, & Slotkin J (2020). The NIH Toolbox: Overview of Development for Use with Hispanic Populations. J Int Neuropsychol Soc, 26(6), 567–575. doi:10.1017/s1355617720000028 [PubMed: 32063249]

Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, & Nowinski CJ (2013). NIH Toolbox for Assessment of Neurological and Behavioral Function. Neurology, 80(11 Supplement 3), S2 LP–S6. doi:10.1212/WNL.0b013e3182872e5f [PubMed: 23479538]

Harkness J, & Schoua-Glusberg A (1998). Questionnaires in translation. In: DEU.

Harwood RL, Schoelmerich A, Schulze PA, & Gonzalez Z (1999). Cultural differences in maternal beliefs and behaviors: A study of middle-class Anglo and Puerto Rican mother-infant pairs in four everyday situations. Child Development, 70(4), 1005–1016. [PubMed: 10446732]

Hodes RJ, Insel TR, & Landis SC (2013). The NIH Toolbox. Neurology, 80(11 Supplement 3), S1 LP–S1. doi:10.1212/WNL.0b013e3182872e90

Huang KT, Owino C, Vreeman RC, Hagembe M, Njuguna F, Strother RM, & Gramelspacher GP (2012). Assessment of the face validity of two pain scales in Kenya: A validation study using cognitive interviewing. BMC Palliative Care, 11, 1–9. doi:10.1186/1472-684X-11-5 [PubMed: 22221932]

Kaufman A, & Kaufman N (1983). K-ABC administration and scoring manual. In: Circle Pines, MN: American Guidance Service.

Lambez B, Harwood-Gross A, Golumbic EZ, & Rassovsky Y (2020). Non-pharmacological interventions for cognitive difficulties in ADHD: A systematic review and meta-analysis. Journal of Psychiatric Research, 120, 40–55. doi:10.1016/j.jpsychires.2019.10.007 [PubMed: 31629998]

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

Ma V, & Schoeneman TJ (1997). Individualism versus collectivism: A comparison of Kenyan and American self-concepts. Basic and Applied Social Psychology, 19(2), 261–273.

Marangu D, Mwaniki H, Nduku S, Maleche-Obimbo E, Jaoko W, Babigumira J, … Rao D (2017). Adapting a stigma scale for assessment of tuberculosis-related stigma among English/Swahili-speaking patients in an African setting. Stigma and Health, 2(4), 307–326. doi:10.1037/sah0000056

Masquillier C, Wouters E, Loos J, & Nöstlinger C (2012). Measuring Health-Related Quality of Life of HIV-Positive Adolescents in Resource-Constrained Settings. PLOS ONE, 7(7), e40628. doi:10.1371/journal.pone.0040628 [PubMed: 22815776]

McHenry MS, McAteer CI, Oyungu E, McDonald BC, Bosma CB, Mpofu PB, … Vreeman RC (2018). Neurodevelopment in Young Children Born to HIV-Infected Mothers: A Meta-analysis. Pediatrics, 141(2), e20172888. doi:10.1542/peds.2017-2888 [PubMed: 29374109]

McHenry MS, Oyungu E, Yang Z, Hines AC, Ombitsa AR, Vreeman RC, … Monahan PO (2021). Cultural adaptation of the Bayley Scales of Infant and Toddler Development, 3rd Edition for use in Kenyan children aged 18–36 months: A psychometric study. Research in Developmental Disabilities, 110, 103837. doi:10.1016/j.ridd.2020.103837 [PubMed: 33453695]

Miles S, Fulbrook P, & Mainwaring-Mägi D (2016). Evaluation of Standardized Instruments for Use in Universal Screening of Very Early School-Age Children: Suitability, Technical Adequacy, and Usability. Journal of Psychoeducational Assessment, 36(2), 99–119. doi:10.1177/0734282916669246

Moffitt TE, Arseneault L, Belsky D, Dickson N, Hancox RJ, Harrington H, … Heckman JJ (2011). A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences of the United States of America, 108(7), 2693–2698. Retrieved from www.jstor.org/stable/41002200 [PubMed: 21262822]

Mungas D, Widaman K, Zelazo PD, Tulsky D, Heaton RK, Slotkin J, … Gershon RC (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. [PubMed: 23952205]

Oyserman D, Coon HM, & Kemmelmeier M (2002). Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. Psychological bulletin, 128(1), 3. [PubMed: 11843547]

Park DC, & Huang C-M (2010). Culture Wires the Brain: A Cognitive Neuroscience Perspective. Perspectives on psychological science: a journal of the Association for Psychological Science, 5(4), 391–400. doi:10.1177/1745691610374591 [PubMed: 22866061]

Pendergast LL, Schaefer BA, Murray-Kolb LE, Svensen E, Shrestha R, Rasheed MA, … Seidman JC (2018). Assessing development across cultures: Invariance of the Bayley-III Scales Across Seven International MAL-ED sites. Sch Psychol Q, 33(4), 604–614. doi:10.1037/spq0000264 [PubMed: 30507236]

Riggins T, Miller N, Bauer P, Georgieff M, & Nelson C (2009). Consequences of Low Neonatal Iron Status Due to Maternal Diabetes Mellitus on Explicit Memory Performance in Childhood. Developmental Neuropsychology, 34(6), 762–779. doi:10.1080/87565640903265145 [PubMed: 20183732]

Rose SA, Feldman JF, Jankowski JJ, & Van Rossem R (2008). A Cognitive Cascade in Infancy: Pathways from Prematurity to Later Mental Development. Intelligence, 36(4), 367–378. doi:10.1016/j.intell.2007.07.003 [PubMed: 19122757]

Sechrest L, Fay TL, & Zaidi SH (1972). Problems of translation in cross-cultural research. Journal of cross-cultural psychology, 3(1), 41–56.

Sherr L, Croome N, Bradshaw K, & Parra Castaneda K (2014). A systematic review examining whether interventions are effective in reducing cognitive delay in children infected and affected with HIV. AIDS Care, 26, S70–S77. doi:10.1080/09540121.2014.906560 [PubMed: 24720704]

Skuy M, Taylor M, O'Carroll S, Fridjhon P, & Rosenthal L (2000). Performance of black and white South African children on the Wechsler Intelligence Scale for Children—revised and the Kaufman Assessment Battery. Psychological Reports, 86(3), 727–737. [PubMed: 10876320]

Smit J, Van den Berg C, Bekker L, Seedat S, & Stein D (2006). Translation and cross-cultural adaptation of a mental health battery in an African setting. African health sciences, 6(4).

Tulsky DS, Carlozzi NE, Chevalier N, Espy KA, Beaumont JL, & Mungas D (2013). V. NIH Toolbox Cognition Battery (CB): measuring working memory. Monogr Soc Res Child Dev, 78(4), 70–87. doi:10.1111/mono.12035 [PubMed: 23952203]

Tusing ME, & Ford L (2004). Examining Preschool Cognitive Abilities Using a CHC Framework. International Journal of Testing, 4(2), 91–114. doi:10.1207/s15327574ijt0402_1

Vanderwood ML, McGrew KS, Flanagan DP, & Keith TZ (2002). The contribution of general and specific cognitive abilities to reading achievement. Learning and Individual Differences, 13(2), 159–188. doi:10.1016/s1041-6080(02)00077-8

Victorson D, Manly J, Wallner-Allen K, Fox N, Purnell C, Hendrie H, … Gershon R (2013). Using the NIH Toolbox in special populations. Neurology, 80(11 Supplement 3), S13 LP–S19. doi:10.1212/ WNL.0b013e3182872e26

Wechsler D (1997). WAIS-3., WMS-3: Wechsler adult intelligence scale, Wechsler memory scale: Technical manual: Psychological Corporation.

Wechsler D (2014). WISC-V: Technical and Interpretive Manual. In. Bloomington, MN: Pearson.

Weintraub S, Bauer PJ, Zelazo PD, Wallner-Allen K, Dikmen SS, Heaton RK, … Gershon RC (2013). NIH TOOLBOX COGNITION BATTERY (CB): INTRODUCTION AND PEDIATRIC DATA. Monographs of the Society for Research in Child Development, 78(4), 1–15. Retrieved from www.jstor.org/stable/43772787

Willis GB, & Artino AR (2013). What Do Our Respondents Think We're Asking? Using Cognitive Interviewing to Improve Medical Education Surveys. Journal of Graduate Medical Education, 5(3), 353–356. doi:10.4300/jgme-d-13-00154.1 [PubMed: 24404294]
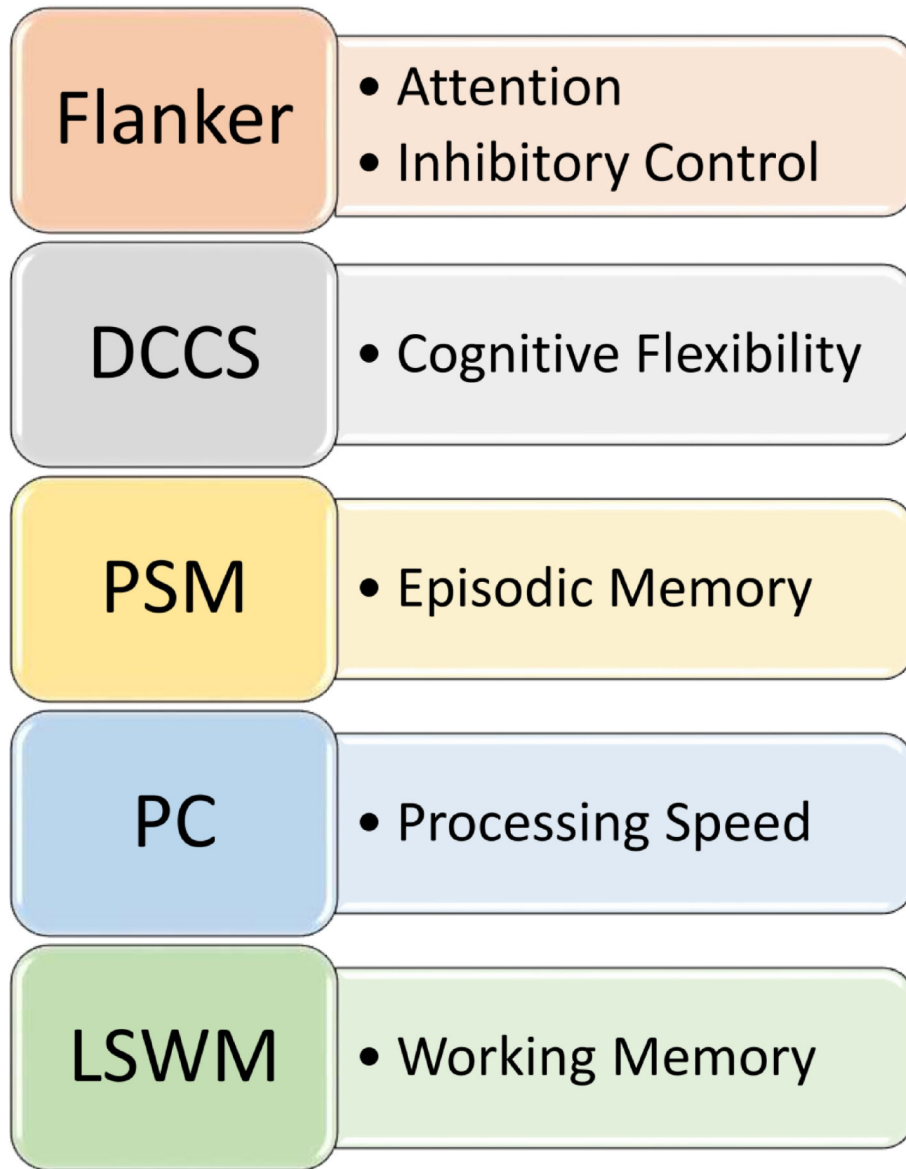
Figure 1.
Selected NIH Toolbox® TestsNote. Flanker, Flanker Inhibitory Control and Attention Test; DCCS, Dimensional Change Card Sort Test; PSM, Picture Sequence Memory Test; PC, Pattern Comparison Processing Speed Test; LSWM, List Sorting Working Memory Test.
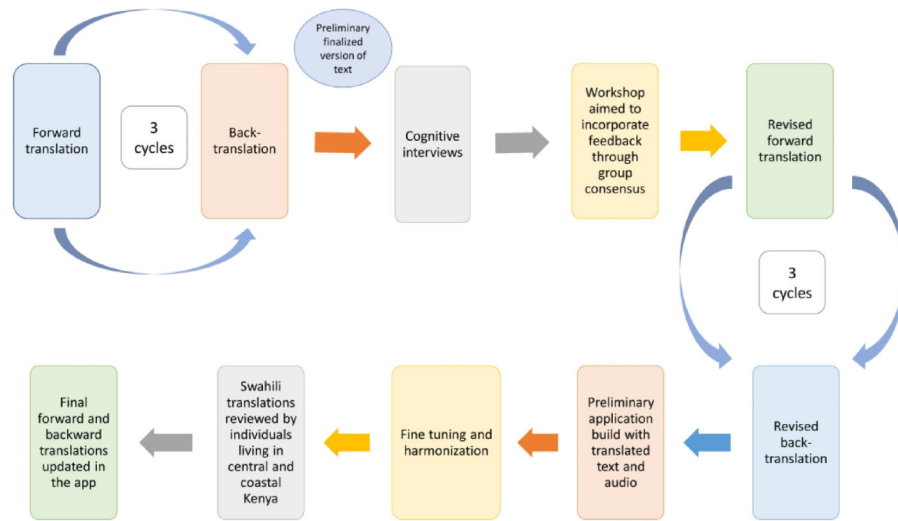
Figure 2.
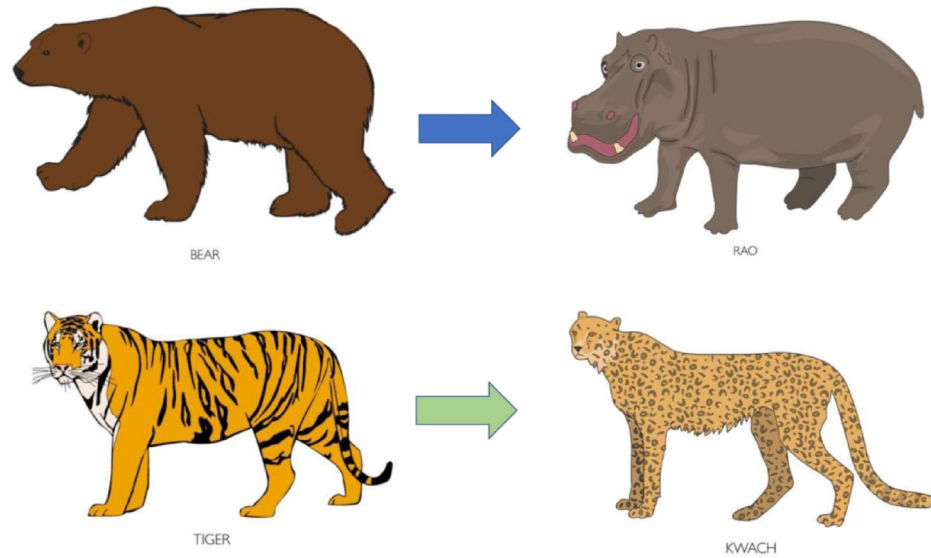Flowchart of the translation and cultural adaptation process

Figure 3.
Select items from the NIH Toolbox List Sorting Working Memory Test (Working Memory), used with permission NIH Toolbox © 2021, National Institutes of Health, and Northwestern University. The practice images "bear" and "tiger" were replaced with "hippo" and "cheetah," respectively. Pictured are the Dholuo words for hippo and cheetah. These were also translated into Swahili.

Table 1.

Demographic data of cognitive interview population

	Ajigo site (%)	Eldoret site (%)	Combined sites (%)
Female	17 (47)	16 (44)	33 (46)
Male	19 (53)	20 (56)	39 (54)
Right-handed	35 (97)	36 (100)	71 (99)
Left-handed	1 (3)	0 (0)	1 (1)
Total	36	36	72