Measuring executive function skills in young children in Kenya

Michael T. Willoughby, Benjamin Piper, Dunston Kwayumba & Megan McCune

To cite this article: Michael T. Willoughby, Benjamin Piper, Dunston Kwayumba & Megan McCune (2018): Measuring executive function skills in young children in Kenya, Child Neuropsychology, DOI: 10.1080/09297049.2018.1486395

To link to this article: https://doi.org/10.1080/09297049.2018.1486395

Published online: 17 Jun 2018.
ARTICLE

Measuring executive function skills in young children in Kenya

Michael T. Willoughby\textsuperscript{a}, Benjamin Piper\textsuperscript{b}, Dunston Kwayumba\textsuperscript{b} and Megan McCune\textsuperscript{c}

\textsuperscript{a}Education & Workforce Development, RTI International, Research Triangle Park, North Carolina, USA; \textsuperscript{b}International Education, RTI International, Nairobi, Kenya; \textsuperscript{c}International Education, RTI International, Washington, DC, USA

ABSTRACT

Interest in measuring executive function skills in young children in low- and middle-income country contexts has been stymied by the lack of assessments that are both easy to deploy and scalable. This study reports on an initial effort to develop a tablet-based battery of executive function tasks, which were designed and extensively studied in the United States, for use in Kenya. Participants were 193 children, aged 3–6 years old, who attended early childhood development and education centers. The rates of individual task completion were high (65–100%), and 85% of children completed three or more tasks. Assessors indicated that 90% of all task administrations were of acceptable quality. An executive function composite score was approximately normally distributed, despite higher-than-expected floor and ceiling effects on inhibitory control tasks. Children’s simple reaction time ($\beta = -0.20, p = 0.004$), attention-related behaviors during testing ($\beta = 0.24, p = .0005$), and age ($\beta = -0.24, p = .0009$) were all uniquely related to performance on the executive function composite. Results are discussed as they inform efforts to develop valid and reliable measures of executive function skills among young children in developing country contexts.

Executive functions (EFs) are a set of cognitive skills that contribute to planning, problem solving, and goal-directed behavior (Diamond, 2013). In contrast to other more crystallized or automatized aspects of cognition, EFs are fluid cognitive processes that individuals draw upon in circumstances when automaticity is not possible. Three foundational components of EF include working memory (defined as the holding in mind and updating of information while performing some operation on it), inhibitory control (defined as the inhibition of automatized responding when engaged in task completion), and cognitive flexibility (defined as the ability to shift one’s cognitive set among distinct but related aspects of a given task). Working memory, inhibitory control, and cognitive flexibility are the components of EF that are most frequently studied in early childhood (Diamond, 2013; Garon, Bryson, & Smith, 2008).

EF skills reflect individual differences in the integrity and efficiency of neural networks that are mediated by the prefrontal cortex and the anterior cingulate cortex.
(Miller & Cohen, 2001; Petersen & Posner, 2012). Notably, both regions are extensively interconnected with subcortical structures that are involved in emotion processing, stress response, and sensorimotor inputs (Bush, Luu, & Posner, 2000; Koziol, Budding, & Chidekel, 2012). This structure provides the neurobiological basis for EF skills to exert “top-down” controls on other aspects of cognition and behavior but also to be influenced by “bottom-up” processes, including emotional reactivity and stress (Arnsten, 2015).

EF skills, as well as the neural substrates that support them, evidence life-course development (De Luca & Leventer, 2008). Early childhood, which is often defined as 3–6 years of age, is the developmental period in which rapid improvements in EF skills are first evident. These rapid improvements represent quantitative differences in the efficiency of neural processes that support EF skills (e.g., Bell & Wolfe, 2007; Moriguchi & Hiraki, 2011), as well as the qualitative strategies that are used during EF tasks (Chevalier, Huber, Wiebe, & Espy, 2013). As such, EF skills may be especially susceptible to environmental influences during early childhood (Fox, Levitt, & Nelson, 2010; Knudsen, 2004). Optimistically, this openness to environmental influence implies that efforts to enhance or intervene on EF may be especially effective in early childhood, which aligns with broader economic arguments regarding the strong returns on investment that result from early childhood interventions (Heckman & Mosso, 2014). Pessimistically, the openness to environmental influence implies that chronic exposure to risk factors across early childhood may have pernicious effects on EF, which is consistent with evidence that early life poverty impairs EF skills and the neural processes that support them (Hackman, Farah, & Meaney, 2010; Noble et al., 2015).

Most of what is known about EF in early childhood, as well as across the life course more generally, is based on studies of participants who reside in high-income countries. However, researchers and policymakers have increasingly emphasized the importance of developing and adapting neurodevelopmental assessments (of which EF is a subset) for use with children in low- and middle-income countries (LMICs) (Fernald, Kariger, Engle, & Raikes, 2009; Fernandes et al., 2014; Kutlesic, Brewinski Isaacs, Freund, Hazra, & Raiten, 2017; Prado et al., 2010). The reason for interest in neurodevelopmental assessments for use in LMICs is twofold. First, progress in the treatment of specific pediatric diseases that directly impact brain and behavior development, such as HIV and malaria, has shifted the focus away from concerns of mortality toward concerns of morbidity. Efforts to improve the neurodevelopmental outcomes of children with specific pediatric diseases require improved measurement tools (Sherr, Croome, Parra Castaneda, Bradshaw, & Herrero Romero, 2014). Beyond disease-specific research, researchers and policymakers have emphasized that children who reside in LMICs face numerous health and nutritional risk factors that collectively undermine their long-term developmental potential (Grantham-McGregor et al., 2007), and these risks are inequitably distributed, further disadvantaging the already vulnerable and poor. This growing consensus has underscored the importance of developing and adapting neurodevelopmental assessments that can be used to evaluate intervention-related efforts. A special emphasis has been placed on tools that are appropriate for use in early childhood, given the assumption that early interventions have the greatest potential impacts (Fernald et al., 2009; Suchdev et al., 2017).
Although EF skills represent one facet of neurodevelopmental functioning, most neurodevelopmental assessments that have been used in LMICs have not measured EF skills; this is especially true for assessments that are intended for use in early childhood (Fischer, Morris, & Martines, 2014; Semrud-Clikeman et al., 2017). Similarly, although numerous early learning assessments have been developed for use with young children in LMIC contexts, with two exceptions, they have tended to measure social-emotional competence, emergent literacy, language, and numeracy skills, and fine and gross motor development, but not EF skills (Brinkman et al., 2017; Gladstone et al., 2010; McCoy et al., 2017; Pisani et al., 2017). The exceptions are the International Development and Early Learning Assessment and the Measuring Early Learning and Quality Outcomes (MELQO) batteries, which included three performance-based tasks of EF in addition to the more commonly measured early learning constructs (UNESCO, UNICEF, Brookings Institution, & the World Bank, 2017; Wolf et al., 2017). At the time of this writing, we are unaware of any published EF data that originated from these batteries. However, a broader search of the research literature identified five studies that administered performance-based EF assessments to 3- to 6-year-old children in LMIC contexts. These studies took place in Albania (Von Suchodoletz, Uka, & Larsen, 2015), Indonesia (Prado et al., 2012), Costa Rica and Cameroon (Chasiotis, Kiesling, Hofer, & Campos, 2006), Zambia (McCoy, Zuilkowski, & Fink, 2015), and Pakistan (Tarullo et al., 2017). These pioneering studies provide the strongest demonstration to date of the feasibility of using EF tasks that were developed in high-income countries with young children in LMICs. The studies by Chasiotis et al. (2006) and Tarullo et al. (2017) were especially noteworthy because they involved multi-task batteries that were used to create more reliable composites of EF skills.

An implicit assumption in these studies is that tasks that were used to measure EF skills in Western contexts are generalizable to LMICs. This assumption is consistent with the characterization of EF skills as a domain-general construct that is appropriate for assessment globally (Boivin & Giordani, 2009). In their systematic review of the use of the Kaufman batteries in sub-Saharan Africa, Van Wyhe, Van De Water, Boivin, Cotton, and Thomas (2017) noted that previous efforts to improve the “cultural fairness” of the Kaufman battery have included a reliance on local norms to define impairment, the use of local translations and data collectors, and avoidance of knowledge and achievement subtests that rely heavily on crystallized (cf. fluid) intelligence that is highly contextually dependent. Many commonly used EF assessments now accommodate these issues (i.e., the results are EF testing are often based on within-sample comparisons and the content of tasks do not rely on crystallized knowledge). Most of what has been written regarding cultural influences on EF has focused on performance differences (1) that are observed from children who are mono- versus multilingual (Ross & Melinger, 2017; Von Bastian, Souza, & Gade, 2016) or (2) that result from comparisons of children in Eastern versus Western cultures (Imada, Carlson, & Itakura, 2013; Moriguchi, Evans, Hiraki, Itakura, & Lee, 2012). Neither of these issues was directly relevant to this study. However, we did make concerted efforts to ensure that task stimuli, verbal instructions, and general assessment approach for EF tasks were culturally appropriate for Kenya. We also note that previous studies have successfully measured EF skills in school-aged children and adults in Kenya, which increased our expectation that the same could be accomplished with young children (Esopo et al., 2018; Kariuki, Abubakar, Newton, & Kihara, 2014; Kitsao-Wekulo, Holding, Taylor, Abubakar, & Connolly, 2013).
All studies that have measured EF skills among young children in LMICs have been limited by their use of paper-and-pencil methods to administer tasks and to record child responses. Paper-and-pencil tasks require time-intensive staff training and certification, create opportunities for errors in administration and scoring, often require task-specific materials (e.g., cards, blocks), and involve data entry expense. The use of touch-screen-enabled computerized data collection tools has the potential to address these limitations and is essential if EF assessments are to be incorporated into large-scale data collection efforts. Bangirana and colleagues made a similar point in their investigation of CogState, which is a rapid computerized neurocognitive assessment that was used with 5- to 13-year-old children in Uganda (Bangirana, Sikorskii, Giordani, Nakasujja, & Boivin, 2015).

One of us has been involved in the development of a battery of EF tasks (i.e., EF Touch) that were intended for use in early childhood in the United States (summarized in Willoughby & Blair, 2016). EF Touch, which includes tasks that measure inhibitory control, working memory, and cognitive flexibility, is appropriate for use in large-scale studies involving young children. EF Touch is portable, requires no task-specific materials, and involves highly scripted instructions to facilitate consistent administration by end users with minimal training and no technical knowledge of EF. The tasks were specifically designed with young children in mind. This included making the instructional language “child friendly,” and limiting individual tasks to 3–7 min in length to minimize attentional demands and to create opportunities for breaks as needed. A common structure is used for each task (item demonstration, followed by training, followed by test items). Children complete test items only if they pass training items ("passing" is automatically determined). The content of the tasks in EF Touch is similar to that of other tasks that are in wide use with young children. What has differentiated EF Touch from other efforts is the extensive psychometric evaluation that has focused on the measurement properties of individual tasks, the battery-wide score, and the iterative approach to task development and refinement (Willoughby, Blair, Wirth, Greenberg, & The Family Life Project Key Investigators, 2010, 2012; Willoughby, Wirth, & Blair, 2011; Willoughby, Wirth, Blair, & The Family Life Project Key Investigators, 2012; Willoughby, Pek, & Blair, 2013; Willoughby, Blair, & Family Life Project Key Investigators, 2016).

EF Touch was initially developed in a paper-and-pencil format, transitioned to a Microsoft Windows environment, and has recently been expanded to an Android tablet environment using Tangerine™ software. The transition into Tangerine™ is notable because this electronic platform already had been used for data collection in more than 60 LMICs (http://www.tangerinecentral.org). Migrating EF Touch into Tangerine™ has the potential to expedite the global measurement of EF skills in young children in LMICs, and the ability to adapt and update tools to different LMIC contexts. Here, we report on a proof-of-concept study that was designed to inform basic questions about the feasibility and validity of direct measurement of EF skills with young children in an LMIC context.

The goals of this study were threefold. First, we pilot tested an approach for task review and adaptation, translation, and data collector training. Second, we evaluated the feasibility of computerized (tablet-based) assessments of EF skills in young children living in a developing-country context. Consistent with our previous work (Kuhn, Willoughby, Blair, & McKinnon, 2017), feasibility was defined with respect to rates of task completion, the length of
administration, and assessor impressions of data quality. Evidence indicating high rates of task completion, developmentally appropriate time expectations for task completion, and high assessor confidence in data quality would all support the feasibility of this work. Third, we tested the validity of the EF tasks by correlating children’s performance with other factors known to affect measures of EF skills, including child age, processing speed, and on-task behavior. If EF Touch is performing as expected, age differences in overall performance should be evident, children with shorter (faster) reaction time should perform better on individual tasks, and children’s task engagement should be related to better performance. Collectively, this evidence would demonstrate the feasibility and initial validity of Tangerine™ EF Touch and would provide the basis for larger-scale and more rigorous investigations of this battery in LMIC contexts.

Methods

Setting

Nairobi is the capital city of Kenya. It is located on the southern edge of the rich agricultural area of central Kenya. The 2009 Kenyan census placed the population of Nairobi at 2.7 million residents (Kenya National Bureau of Statistics, 2010), but projected a 2017 population closer to 3.6 million. As a capital city, Nairobi is diverse ethnically, with representation from across the country. The ethnic groups with the largest populations in Nairobi County include the Kikuyu, Luo, Luhya, Kisii, Kamba, and Somali. In addition to Kenyan ethnic groups, there is a sizable population of Asians and Europeans. Nairobi also hosts citizens from neighboring countries disturbed by open conflict or insecurity, including South Sudan, Uganda, the Democratic Republic of Congo, Ethiopia, and Somalia.

Nairobi’s early childhood development and education (ECDE) provisions vary. While Nairobi has more ECDE centers than any other jurisdiction in the country, most (90%) are private rather than public ECDE centers supported by government funds (Ministry of Education, Science and Technology, 2014). The private school population is also diverse, with a large majority of private schools located in urban settlements; the Ministry of Education refers to them as Alternative Provision of Basic Education and Training (APBET) institutions (Zuilkowski, Piper, Ong’ele, & Kiminza, In press). APBET institutions typically are small, with less than 200 students at each center, and are often located in temporary physical structures. The Ministry of Education’s language-of-instruction policy requires that the language of instruction in ECDE and lower primary education be the “language of the catchment area,” which is either English or Kiswahili given Nairobi’s ethnic and language diversity (Kenya Institute of Education, 1992). The language used by most private schools and APBET institutions in Nairobi’s informal settlements is English, although many children are more comfortable with Kiswahili (Piper, Zuilkowski, & Ong’ele, 2016).

Sample and participants

This study was conducted in cooperation with the Kenyan Ministry of Education and was approved by Kenya’s National Commission for Science, Technology and Innovation and the Kenya Medical Research Institute, the two bodies responsible for
approving human subjects research in Kenya. At the ECDE institution level, consent was first sought from the school head teacher and teachers, followed by oral assent from each of the children who participated in the study. We sampled 16 schools from Nairobi for the *EF Touch* pilot assessment. Six public and 10 APBET institutions were randomly selected using Stata statistical software. From the 16 sampled schools, the final 10 schools (4 public and 6 APBET) were selected because they had at least 10 children in the pre-primary 1 and 2 classrooms. Pre-primary 1 serves 4 year olds and pre-primary 2 is for 5 year olds, although generally there are some children older and younger in both types of classrooms. We used systematic random sampling at the school level, balancing by gender within each ECDE level.

In total, 200 children were recruited into the study. In the process of establishing initial rapport with each child, data collectors determined the child’s language preference (English or Kiswahili) and were instructed to administer all tasks in this language. At the data processing phase, we determined that seven students had tasks administered in more than one language. We dropped these data from consideration to eliminate language administration as a source of potential error. Among the 193 participating children, most had tasks administered in Kiswahili (74%), with the remaining in English. Participating children were 51% female and 3–6 years old (2% age 3; 35% age 4; 40% age 5; 23% age 6) at the time of the assessment.

**Procedures**

*EF Touch adaptation*

The initial step in this work involved expanding the technical functionality of the Tangerine™ platform to allow the delivery of audio and visual images as stimuli and to record child responses in a touch screen environment. Once these technical developments were completed, we conducted a four-step task adaptation process to prepare *EF Touch* tasks for use in Kenya. First, a content review of all tasks (in English) was conducted by multiple PhD-level colleagues with extensive experience in early childhood assessment in East Africa. Second, the English version of each task was translated into Kiswahili by a professional in-country translator and the Kiswahili version was independently compared back to the English version by an individual working with the translator. Third, a 2-day adaptation workshop was conducted in Nairobi that involved the research staff (including all coauthors) and Kenyan Ministry of Education officials, all of whom had expertise in early childhood development in Kenya and various assessment batteries for early childhood. Some of them also had experience in adapting MELQO (UNESCO et al., 2017) to the Kenyan context (RTI International, 2015). Each task was reviewed in detail – the intent of the task as it relates to the local cultural context, as well as the appropriateness of the translated text to the Kenyan context, taking into consideration child-friendly language and local languages. Workshop participants collected data on 32 children, and then attended a group debriefing to finalize task translations and study procedures based on the applicability of the tasks and the language used to Kenya’s context.

Research leads and Ministry of Education officials who participated in the adaptation workshop subsequently conducted a 3-day training of 20 data collectors. Data collectors were selected from a pool of 157 individuals who had experience in undertaking
assessments at the pre-primary level, plus geographical knowledge of the study sites. Like the adaptation workshop, the facilitators made a didactic presentation about EF, then reviewed each EF Touch task in depth. Pilot data from an additional 40 children were collected during the data collection training workshop, to familiarize data collectors with the process of using EF Touch and to provide opportunities to generate questions. Debriefing of data collection experience and continued practice ensured a high degree of data collector confidence. On the final day of training, each data collector conducted a mock administration of the entire battery before one of the research leads. The leads completed a data collector certification checklist that was reviewed with data collectors before the start of formal data collection.

**Data collection**
Five teams of four data collectors collected EF Touch data from 200 children over a 2-day period. All data collection occurred in the mornings at participating ECDE centers. Equal numbers of boys and girls were sampled from each classroom, and individual testing occurred in spaces provided by school staff (e.g., outdoors on benches in a common space; in a separate building that adjoined the school). Children worked one-on-one with data collectors. Data collectors determined the language of task administration (English or Kiswahili) based on initial, rapport-building conversations with children (e.g., asking simple questions of the child and noticing which language was used in her response). Data were collected using Android tablets (9.4 cm × 15.2 cm). Tablets were placed in an upright position in tablet stands (approximately US$7 each) in landscape orientation, to create a more uniform testing experience. Assessment data, devoid of student identifiers, were uploaded from tablets to a Tangerine™ server at the end of each data collection day.

**Measures**

**EF Touch: Bubbles**
This 30-item task measured simple reaction time. A series of blue bubbles was presented, one at a time, and children were instructed to “pop” each bubble as fast as they could. Items were presented for up to 5,000 ms, and the time that transpired between stimuli onset and child touch was recorded by the tablet. Item responses that were faster than 400 ms were considered too fast to be plausible (i.e., likely reflected a response recorded from the previous screen) and were set to missing. If a child failed to touch any bubble, the item was considered inaccurate and reaction time was not recorded. The mean reaction time across all correctly answered items was used to index simple reaction time.

**EF Touch: Silly Sounds Stroop**
This 17-item Stroop-like task measured inhibitory control. Each item displayed pictures of a dog and a cat (the left–right placement on the screen varied across trials) and presented the sound of either a dog barking or cat meowing. After the assessor and child established the sounds that each animal typically made, the child was instructed to play a silly game that involved touching the picture of the animal that did not make the sound (e.g., touching the cat when hearing a dog bark). Each item was presented for 3,000 ms, and the accuracy and reaction time of responses were recorded.
responses that were faster than 400 ms were considered too fast to be plausible and were set to missing. If an item was omitted, the child was given an accuracy score of zero, and reaction time was not recorded. Mean accuracy across all items was used to represent performance.

**EF Touch: Animal Go/No-Go**

This 40-item go/no-go task measured inhibitory control. Individual pictures of animals were presented, and children were instructed to touch a centrally located button on their screen every time that they saw an animal (the “go” response), except when that animal was a pig (the “no-go” response). Each item was presented for 3,000 ms, and the accuracy and reaction time of responses were recorded. Item responses that were faster than 400 ms were considered too fast to be plausible and were set to missing. If an item was omitted, children were given an accuracy score of zero, and reaction time was not recorded. Mean accuracy across the eight no-go items was used to represent performance.

**EF Touch: Spatial Conflict Arrows**

This 36-item spatial conflict task measured inhibitory control and cognitive flexibility. Two soft buttons appeared on the left- and right-most sides of the tablet screen. Children were instructed to touch the button to which an arrow was pointing. Three blocks of 12 arrows were depicted in which arrows appeared above the button to which they were pointing (congruent condition), above the opposite button to which they were pointing (incongruent condition), or in mixed locations. Each item was presented for 3,000 ms, and the accuracy and reaction time of responses were recorded. Item responses that were faster than 400 ms were considered too fast to be plausible and were set to missing. If an item was omitted, children were given an accuracy score of zero, and reaction time was not recorded. Mean accuracy for the number of incongruent items (from incongruent and mixed conditions) was used to represent performance.

**EF Touch: Pick the Picture**

This 32-item self-ordered pointing task measured working memory. Children were presented with arrays of pictures that varied in length (i.e., two, three, four, or six pictures per set). For each set, children initially were instructed to touch any picture of their choice. On subsequent trials within that set, the pictures were presented in different locations, and children were instructed to pick a picture that had not yet been touched. The mean accuracy of responses in each picture set (except for the first picture, which did require working memory) was used to represent task performance.

**EF Touch: Something’s the Same**

This 30-item two-part task was intended to measure attention shifting and flexible thinking. In part 1 (20 items), children were presented with two pictures (e.g., two animals, two flowers) that were similar in one dimension (i.e., color, shape, or size) and described as such. A third picture was then presented alongside the original two, and the child was asked to select which of the original pictures was similar to the new picture along some other dimension (e.g., blue flower and red flower are similar
because they are flowers; the third new picture of blue chair is similar to the blue flower because they are both blue). In part 2 (10 items), a new set of instructions and demonstrations was provided that involved presenting the child with all three pictures and asking that the child identify one pair of pictures that were similar along one dimension (i.e., color, shape, size) and then a second pair among the same three pictures that were similar in some other dimension. After deployment, we determined a programming error in part 2 of the task and thus do not report on those items here (we mention part 2 because it contributed to our estimate of task time). The mean accuracy of responses in part 1, which corresponded to our previous work, was used to represent task performance.

**EF Touch: Task quality ratings**

At the end of each task, data collectors completed a one-item three-category rating of the perceived quality of the collected data (i.e., 1 = I have SERIOUS CONCERNS about these data [e.g., the child did not seem to understand the task, was frequently distracted/interrupted, was unable to pay attention, was not engaged in task, etc.]; 2 = I feel OK about these data [e.g., the child seemed to understand the task pretty well, only experienced a few distractions or interruptions, paid attention reasonably well, was somewhat engaged in task, etc.]; 3 = I feel GOOD about these data [e.g., the child clearly understood task, had very few [if any] distractions/interruptions, paid attention to task, was engaged, etc.]). These ratings reflected the data collectors’ subjective impressions of children’s comprehension of task instructions, engagement with the task, and the suitability of the testing environment.

**Preschool self-regulation assessment (PSRA)**

Data collectors rated child behavior during testing using the PSRA assessor report (Smith-Donald, Raver, Hayes, & Richardson, 2007). The PSRA includes items that reflect children’s attention (e.g., pays attention during instructions; sustains attention), behavior (e.g., remains seated; engages with person administering task), or affect (e.g., shows pleasure in accomplishment; shows feelings) during the EF Touch battery. Examiners rated each item on a three-point Likert rating scale, with each scale point consisting of clear, behavioral descriptors. Four items were omitted, based on feedback from data collectors and stakeholders during the training workshop (those items referred to children grabbing test items or showing various forms of aggression, none of which was considered relevant). In line with prior research (Smith-Donald et al., 2007), two subscales, which reflected attention-relative behaviors (14 items; Cronbach’s $\alpha = 0.87$) and positive affect (9 items; Cronbach’s $\alpha = 0.75$) during testing, were derived from factor analysis of these items in this sample. Due to a technical problem, PSRA reports were available for only 155 students.

**Results**

**Feasibility**

Descriptive analyses were used to characterize the feasibility of using Tangerine™ EF Touch with young children who attended center-based care in Nairobi, Kenya. The first feasibility metric was the proportion of children who completed a task, where
completion was defined as passing training items. Between 62% and 100% of children completed each EF task (see Table 1). Children completed an average of four of five possible EF tasks (English $M = 3.9$ tasks; Kiswahili $M = 4.3$ tasks), with 85% of children completing three or more tasks (84% of English, 96% Kiswahili).

Children who passed the training items for a given task typically completed all of the resulting test items ($M = 85–100\%$ and Mode = 100% across tasks). Tasks took an average of 3–5 min to complete (see Table 1; note that while “Something’s the Same” took an average of 9 min to complete, it was a two-part task and is better considered two adjoined tasks for purposes of time estimates). Task times in Table 1 include the time associated with the description of the task and the completion of demonstration items by the data collector, as well as the completion of training and (when appropriate) test items from children. These times do not include any breaks that were offered in between the tasks, which were given at the discretion of the data collector. In general, the rates of task and item completion, as well as task lengths, were similar irrespective of the language of administration.

On average, 90% of all administrations were deemed to be of acceptable (“OK”) or good quality, with the remaining 10% of administrations raising some concern (see Figure 1). Data quality ratings were similarly high irrespective of the language in which the task was administered.

### Task and battery performance

Children’s performance on individual tasks is summarized in Table 2. Children completed between 61% and 82% of test items correctly on each task. Floor effects were evident for two tasks. Specifically, 11% of children overall (9% of Kiswahili and 19% of English) who passed the training items on the Animal Go/No-Go task did not answer a single test item (i.e., no-go trials) correctly. Moreover, 10% of children overall (11% of Kiswahili and 9% of English) who passed the training items on the Spatial Conflict Arrows task did not answer a single test item (i.e., incongruent arrows) correctly. Ceiling effects were evident for all three of the inhibitory control tasks and were particularly pronounced for the Animal Go/No-Go (47%) and Silly

### Table 1. EF task feasibility metrics.

<table>
<thead>
<tr>
<th>Task</th>
<th>Tasks attempted (%) completion</th>
<th>% task items completed</th>
<th>Task length (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Kiswahili</td>
<td>English</td>
</tr>
<tr>
<td>Animal Go/No-Go</td>
<td>193 (92)</td>
<td>142 (100)</td>
<td>51 (71)</td>
</tr>
<tr>
<td>Silly Sounds Stroop</td>
<td>188 (88)</td>
<td>141 (89)</td>
<td>47 (85)</td>
</tr>
<tr>
<td>Spatial Conflict Arrows</td>
<td>192 (62)</td>
<td>141 (60)</td>
<td>51 (67)</td>
</tr>
<tr>
<td>Something’s the Same</td>
<td>192 (85)</td>
<td>142 (86)</td>
<td>50 (82)</td>
</tr>
<tr>
<td>Pick the Picture</td>
<td>191 (100)</td>
<td>141 (100)</td>
<td>50 (100)</td>
</tr>
</tbody>
</table>

Note. N = sample size; M = mean; Mdn = median. For Something’s the Same, time data inadvertently were not collected in the English version of the tasks. Moreover, the total time reflects parts 1 and 2 of the task, although only part 1 scores are used here.
Sounds Stroop (30%) tasks. Descriptively, the mean levels of task performance were comparable irrespective of the language in which the task was administered.

Bivariate correlations among the EF task scores, as well as the simple reaction time task (Bubbles), are summarized in Table 3. Three points are noteworthy. First, children with slower simple reaction times (i.e., longer mean responses on the Bubbles task) performed somewhat more poorly on each of the EF tasks ($r_s = -0.08$ to $-0.29$). Second, the magnitude of the correlations among EF tasks was modest ($r_s = -0.06$ to $0.25$). Third, the Animal Go/No-Go task was unrelated to performance on all other tasks, which was likely a function of limited task variation resulting from floor and ceiling effects.

Consistent with our previous work (Willoughby, Blair, & The Family Life Project Key Investigators, 2016; Willoughby, Kuhn, Blair, Samek, & List, 2017), we created a battery-wide composite by standardizing each task score ($M = 0$, $SD = 1$) and averaging children’s responses across tasks. This battery-wide composite reflected a child’s average standing on all the tasks that s/he completed. As shown in Figure 2, the EF composite
was approximately normally distributed and exhibited expected age-related improvements (3- and 4-year-old children were combined into a single group because of the small number of 3s in the sample). Consistent with individual task scores (see Table 3), the EF composite was negatively correlated with simple reaction time from the Bubbles task \((r = -0.29, p < .001)\). The EF composite was positively related to data collector impressions of children’s attention-related behaviors \((r = 0.36, p < .001)\) – not their positive affect during testing \((r = 0.15, p = .07)\). This finding may be related to cultural differences in how all children interact with adults in Kenya.

An ordinary least-squares regression model was estimated to investigate the unique contributions of simple reaction time (Bubbles task), demographic factors (child age, gender, language of administration), and data collector ratings of children’s behavior during testing (mean replacement was used to accommodate missing values for the PSRA scales). This set of predictors was significantly associated with the EF composite score, \(F(7, 185) = 8.0, p < .001\), \(adjusted R^2 = 0.20\). Individual differences in simple reaction time \((\beta = -0.20, p = .004)\) and attention-related behaviors during testing \((\beta = 0.24, p = .0005)\) were both uniquely related to the EF composite. Specifically,
children who had a faster simple reaction time and who demonstrated greater attention and engagement during task completion had higher EF composite scores. Relative to 5 year olds (reference group), 3- and 4 year olds performed more poorly on the EF composite ($\beta = -0.24, p = .0009$). None of the other demographic factors (language of administration, $p = .78$; sex, $p = .53$; 6 year olds, $p = .62$) were uniquely related to the EF composite, nor were data collector ratings of positive affect during testing ($p = .26$).

**Discussion**

This study represented an initial test of the feasibility of adapting and deploying a battery of performance-based EF tasks – all of which had been administered extensively among 3- to 5-year-old children in the United States – for use in LMIC contexts, with Kenya serving as the initial exemplar. This effort involved technical changes in the method in which tasks were administered (from a Windows-based, dual-monitor setup to an Android-based, single-screen tablet). It also involved the development of an overall approach for task review, translation, data collector training, and task evaluation. With the adoption of a computerized format, this study extends previous studies that have successfully used paper-and-pencil methods to measure EF in young children in developing country contexts. Most children completed most tasks, assessment time was relatively short, and the face validity of tasks was evident. Below, we elaborate specific findings, identify task limitations that will require future attention, and consider the broader implications of this work.

We used a multistep process to review and adapt tasks for use in Kenya. This included a review of task content by in-country experts in early childhood assessment, the use of a professional in-country translator, didactic presentation of EF skills and tasks to stakeholders (research colleagues and Ministry of Education officials), and stakeholder participation in the pilot testing of tasks. Through this process, we learned that the translator needs to clearly understand the intent of each task – not simply the (in this case, English) words to be translated. Providing additional instructions to the translator might have streamlined the stakeholders’ debates about whether the Kiswahili and English versions of tasks were communicating identical task objectives. We also learned the importance of having key technical stakeholders participate in pilot data collection. This arrangement minimized speculation about how young children might perceive and respond to tasks and gave the stakeholders a shared set of experiences with which to discuss tasks. Engaging key research officers in this process in advance of their conducting training for data collectors was also beneficial because they could directly relate to and address questions that arose when data collectors undertook pilot data collection.

In terms of feasibility, with one exception (Spatial Conflict Arrows, described below), 85% or more of children completed each of the EF tasks. The common structure of all tasks (demonstration items, followed by training items, followed by test items for those children that passed training) helped to ensure that the children clearly understood the task objectives. Moreover, automating the determination of whether training items were “passed” standardized the assessment experience. Specifically, data collectors were not required to decide whether the child passed the training items, nor how many rounds of training were appropriate (children were given up to two opportunities to complete training items before the task was automatically discontinued), as the programmed
tablets responded to the child’s performance. Finally, only 62% of children successfully passed the training items for the Spatial Conflict Arrows task. In our previous work, this task was not viewed as more difficult, which suggests that task instructions and/or criteria for whether children could be considered to have passed training items might require revision in LMIC contexts such as Kenya.

Additional support for the feasibility of direct assessments came from the quality ratings that were completed by data collectors at the end of each assessment. The quality ratings indicated that data collectors had confidence in the data that they were collecting, and this was evident for all tasks in both languages. The ratings were a brief but structured method for data collectors to share their subjective impressions of data quality with data analysts. We found this approach more useful than open-form text comments, which are often difficult to analyze and interpret systematically.

In addition to questions of feasibility, we considered children’s performance on each EF task. On average, children completed 61–82% of the test items on each task correctly, with the full range of variation present on each task. However, the inhibitory control tasks suffered from unacceptably high rates of floor and/or ceiling effects. There are two likely explanations for the relatively poor performance of these tasks. First, EF Touch has primarily been used with 3- to 5-year-old children in the United States and was designed to ensure that individual differences could be detected even among 3 year olds from low-income households. The current study included older children (63% of children were 5 or 6 years old) from relatively high-performing preschools in the relatively wealthier capital city of Nairobi. These inhibitory control tasks may not have been sufficiently challenging for children with high ability. Second, the stimulus presentation times for the inhibitory control tasks in EF Touch (3,000 ms) were longer than in other studies that used similar tasks with children of similar ages (1,500–2,000 ms; Simpson & Riggs, 2006; Wiebe, Sheffield, & Andrews Espy, 2012). Shortening the presentation time for individual items would likely reduce ceiling effects. An open question is whether the impact of item presentation time would vary as a function of children’s familiarity with tablets, which is an experience that likely differs both within and across LMIC contexts.

On average, children completed four of the five EF tasks. However, consistent with the broader literature, children’s performances across tasks were weakly \((rs ≈ 0.1–0.3)\) associated (see Willoughby, Holochwost, Blanton, & Blair, 2014). Given modest correlations, we have advocated for creating an overall EF composite score that reflects the summation of performance across tasks rather than defining EF as reflecting only the shared variation across tasks (Willoughby et al., 2016, 2017). The EF composite score in this study was approximately normally distributed and was associated with expected age-graded improvements. Notably, child age, individual differences in simple reaction time, and child testing behavior were all uniquely associated with the EF composite score. In contrast, child gender, the language of task administration, and positive affect observed during task completion were not. These results provide initial face validity for Tangerine™ EF Touch tasks in this new format and setting.

Ardila (2005) identified several cultural-dependent values that threatened the use in LMICs of cognitive assessments that were developed in high-income countries; three such assessments are germane to this study. The first concerns the cultural appropriateness of one-on-one (adult to child) assessment. For better or worse, individualized testing of preprimary and primary-aged youth is increasingly common in LMICs in
general (Gove & Dubecck, 2016). For example, near the time of this study, extensive nationally representative literacy and numeracy assessments of this type were being conducted in Kenya (Freudenberger & Davis, 2017; Kenya National Examinations Council, 2016). Although adult–child assessments are not considered atypical in Kenya, this is an important consideration in other contexts where investments in education and assessments are less common. The second concern is uncertainty in children’s perceptions of the importance of testing, including what degree of effort should be expended. We would never consider EF Touch to be used for anything other than low-stakes decisions and have never described the tasks to children as “tests.” Instead, they are presented as “fun games.” Children are not given any indication of their performance (including when tasks are discontinued if training items are not completed), although they are praised for effort and offered a small reward at the completion of testing (e.g., children were given a pencil after task completion in this study). The third potential concern is cultural variation in time use, including time constraints associated with EF task completion. As we noted, the inhibitory control tasks that involved a speeded component appeared to be too easy. We expect that this was more a characteristic of the participating children than cultural variation in the experience of time. Fortunately, EF Touch allows manipulating both the stimulus presentation and inter-stimulus speeds of speeded tasks to ensure that this is not an issue in other settings.

This study had at least five limitations. First, it involved a small sample of children who were drawn from a convenience sample of preschools in Nairobi. Although preschools were selected to represent a range of settings (i.e., from publicly funded schools supported by the Ministry of Education, to privately funded, informal schools), they tended to be high-performing schools, given Nairobi’s wealth relative to the rest of the country. The results therefore may not represent the typical range of student performance in Kenya. Second, most children in this study ranged from 4 to 6 years of age, which is consistent with the target age of children served in pre-primary classrooms in Kenya. Although a small number of 3 year olds participated, the number of children was too small to make strong inferences about the feasibility of using EF Touch with children this young in LMIC contexts, though we have had good success administering these tasks to 3 year olds in the United States (Willoughby et al., 2010). Third, elsewhere we have emphasized the value of using modern psychometric methods to investigate EF task performance and to test for equivalence in item parameters across subgroups (Willoughby et al., 2011). Although the current study was not sufficiently large to ask these questions, an important direction for future research will involve formal tests of task equivalence across English vs. Kiswahili languages of administration. Fourth, data collectors determined the language of administration based on child responses to rapport-building questions. Given that English is the de facto language of instruction in many schools, some children may have responded to data collectors in English despite have greater proficiency in Kiswahali. This decision may have impacted their performance. Fifth, although we collected item-level reaction time, all EF task scores were based solely on accuracy data. Due to the speeded nature of inhibitory control tasks, the joint use of item-level reaction time and accuracy data might enhance task scoring by reducing instances of floor and ceiling effects (Magnus, Willoughby, Blair, & Kuhn, In press), both of which were a problem for inhibitory control tasks in this study. This is another direction for the next iteration of this work that involves a larger sample size.
Recent national assessments in Kenya have demonstrated a wide range of performance on educational outcomes at grade 2 (Freudenberger & Davis, 2017), grade 3 (KNEC, 2016), and grade 6 (Hungi & Thuku, 2010), and particularly in household-based assessments covering children from ages 6 to 16 (Uwezo, 2016). These performance variations are only partially explained by geographic and socioeconomic differences. Large-scale educational interventions in Kenya have recently begun, including the Tusome program in literacy (2014–2019) and the Primary Education Development program in numeracy (2015–2018). Initial impact evaluations have shown substantial improvements in learning outcomes in literacy (Freudenberger & Davis, 2017), yet massive variations in learning outcomes persist, and in some cases have expanded as a result of these interventions (Oketch, Ngware, & Ezeh, 2010; Piper, Jepkemei, & Kibukho, 2015). These widening gaps suggest that there are missing variables that can explain the differences in performance and that EF skills might contribute to some of that explanation. Moreover, Kenya has only recently begun, through the Tayari program, to support its counties to implement large-scale educational interventions in early childhood development (Piper, Merseth, & Ngaruiya, 2017); and before the advent of this intervention, little had been systematically done to improve the quality of pre-primary education in Kenya at scale within government structures. Understanding how EF skills interact with the range of necessary skills for pre-primary and primary education is a clear gap in the literature in Kenya and likely across other LMICs. Further development of EF Touch and other tools like it are essential for this work to proceed.

In summary, this study demonstrated the feasibility and initial face validity of using Tangerine™ EF Touch to measure EF skills in young children in an LMIC context. The widespread use of the Tangerine™ platform for data collection in LMICs has the potential to greatly accelerate our ability to deploy globally scalable measures of EF at young ages and in locations where standard assessment tools for these children do not yet exist. The ability to reliably and validly measure EF skills in young children in developing country contexts has the potential to make fundamental contributions to the scientific literature, particularly given the dramatic variations in early life experience of children in LMICs relative to high-income countries. Moreover, the availability of reliable, valid, and culturally relevant performance-based assessment of EF in young children in LMICs has the potential to inform policy-relevant questions related to early childhood programming and pedagogy.

Acknowledgments

This work was funded through internally supported research at RTI International and the Children’s Investment Fund Foundation. The authors are grateful for the ongoing collaboration with the Kenyan Ministry of Education, and the ministry’s county education officers. The authors especially appreciate the support of Dr. Hellen Kimathi, the Director of Early Childhood Development for the Ministry of Education.

Funding

This work was supported by the RTI International and the Children’s Investment Fund Foundation.
References


Magnus, B. E., Willoughby, M. T., Blair, C. B., & Kuhn, L. J. (In press). Integrating item accuracy and reaction time to improve the measurement of executive function abilities in early childhood. Assessment.


