Journal of Pediatric Neuropsychology (2021) 7:113–138
https://doi.org/10.1007/s40817-021-00096-9

TEST REVIEW

Assessing Neuropsychological Functions in Middle Childhood: a Narrative Review of Measures and Their Psychometric Properties Across Context

Maina Rachel[1,2] · Van De Vijver J. R. Fons[2,3] · Abubakar Amina[4,5] · Miguel Perez-Garcia[6] · Kumar Manasi[7]

Received: 31 October 2020 / Revised: 4 January 2021 / Accepted: 11 January 2021 / Published online: 15 February 2021

Abstract

Background There is a significant number of neuropsychological measures for use among children aged 6–12 years. However, most of these tests have been developed in high-income contexts (HICs). To avoid or to at least to minimize bias in assessment, most researchers carry out cultural adaptations of these tools. In selecting sub-tests to adapt before using the entire test battery, researchers would benefit from having a reference source summarizing available tools and how easily they can be used in different context. This is where the paper makes a contribution. This narrative review has a twofold aim: first, to identify tools commonly used among 6–12-year-olds; second, to summarize the psychometric properties of these tools especially emphasizing their usage across different cultural contexts.

Methods We searched the literature from 1 January 1987 to 31 December 2017 for tools used among children aged 6 to 12 years. An extensive search of PubMed, Psych Info and Web of Science using the keywords (i) neuropsychological or neurocognitive with (ii) assessment or test was done.

Results A hundred and forty-five papers out of 306 reported on psychometric properties of different tools including Behavior Rating Inventory of Executive Functioning—BRIEF (count = 6), Visual-Motor Integration—VMI (count = 6), the Test of Memory Malingering—TOMM (count = 6), MSVT (count = 6) and Continuous Performance Tests—CPT (count = 6). Forty-six percent of the papers reported studies conducted in the USA. Most of these studies were based in high-income countries, which further highlights the need to validate these measures for use in lower- and middle-income countries. Psychometric check was adequate in most tests for measuring executive functioning such as BRIEF, although tests such as CPT that measure complex attention had mixed findings. Moreover, we found that these studies addressed certain aspects of validity and or reliability while leaving out others; thus, a comprehensive picture is lacking.

Conclusion We propose further studies to thoroughly investigate and report the psychometric properties of these measures, especially in lower- and middle-income countries.

Keywords Child neuropsychological assessments and tools · Psychometrics · Continuous performance · Executive functioning · Sensitivity and specificity norms

List of Abbreviations

BRIEF	Behaviour Rating Inventory of Executive Functioning
WISC	Wechsler Intelligence Scales
MSVT	Medical Symptom Validity Test
TOMM	Test of Memory Malingering
WMT	Word Memory Test
CPT	Continuous Performance Tests
SWM	Spatial Working Memory
CANTAB	Cambridge Neuropsychological Test Battery
TOVA	Test of Variables of Attention
KABC	Kaufman Assessment Battery for Children
DAS	Differential Ability Scales
NEPSY	Developmental Neuropsychological Assessment
WIAT	Wechsler Individual Achievement Test
SSA	Sub-Saharan Africa
SOPT	Self-Ordered Pointing Test
VLL	Verbal List Learning

✉ Maina Rachel
R.W.Maina@uvt.nl

Extended author information available on the last page of the article

CPM	Colored Progressive Matrices
CNT	Contingency Naming Test
DSM V	Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
LMIC	Lower- and middle-income countries
ANT	Attention Network Test
HRNB-C	Halstead-Reitan Neuropsychological Test Battery for Children
CANTAB	Cambridge Neuropsychological Test Automated Battery
BENCI	BaterÃa de EvaluaciÃn NeuropsicolÃgica Infantil
BADS-C	Behavioral Assessment of Dysexecutive Syndrome for Children.

Introduction

The ages 6–12 are known as the 'ages of reason' by the likes of Piaget's cognitive theorists. Children ages 6–7 years are likely to start developing reasoning abilities related to the concrete operational level of cognitive development where they can form complex representations and be able to solve complex problems. For example, a child at this age can understand that a parent can be a disciplinarian and at the same time be a provider while a teacher can also be a parent at their own home hence be a disciplinarian and provider to his/her own children. As these cognitive abilities develop, formal operations level of cognitive development quips in starting at ages 10–12 years. This is where the children can form generalizations across different instances and also have hypothetical reasoning ability. They can combine several shapes to form an overall pattern.

Performance on these cognitive abilities is founded on the physiological growth of the brain in terms of neurons whose plasticity or formation is a function of environmental factors/triggers. Performance is measured adequately by valid and reliable neuropsychological tools and the psychometric adequacy inquiry forms the objectives of this narrative review. This is particularly for children aged 6–12 years whose literature on psychometric properties of cognitive tools is marred by mixed findings (Llorente et al. 2009; Spironello et al. 2010). The mixed findings embedded in different literatures make it hard to find one tool for a certain cognitive function whose validity and reliability indicators are good for forming a hypothesis on the functionality of a child (Hubley and Zumbo 1996). Children aged 6–12 years are just starting school, and their ability to learn is embedded on cognitive functions such as those related to memory formation, problem solving, flexibility and judgement (Chen at al. 2009; Stad et al. 2019). Functions such as cognitive flexibility among these children have been found to be related to school performance (Stad et al. 2019). Culture-sensitive tools can be used to identify learning problems as well as inform an instruction plan that improves performance or a treatment plan that rehabilitates cognitive deficits. Tools for children aged 6–12 years are diverse and with mixed findings on their validity and reliability indicators (Ahonniska et al. 2000; Holding et al. 2004; Llorente et al. 2009; Spironello et al. 2010). Cultural diversity calls for development of tools that are sensitive and specific to the cultural constructs hence the diversity in tools formed and reported psychometric properties. This narrative review aims to summarize findings on psychometric properties of cognitive tools used among children aged 6–12 years.

Neuropsychological Tools

Neuropsychological tools are measures used to assess the brain-behaviour relationship (Casaletto and Heaton 2017). Executive function, memory, visuomotor coordination, processing speed, language and attention are basic cognitive domains measured using these tools (Fasfous et al. 2015). Intrusive tests such as spinal tap were used before the advent of neuropsychological tools which have over the years evolved from paper-based tools to computerized ones. Neuropsychological tools have not only made it less intrusive to assess cognitive functions, they have over the years become more comprehensive and easier to administer with some of the tools needing no training to administer and score. This has made it possible to diagnose neurocognitive disorders as well as monitor dysfunction progression and recovery thereby better informing interventions.

Good neuropsychological tools have to be standardized, reliable and valid. When a test measures what it is purported to measure, then the test is said to be valid and it is reliable when it accurately measures what it is supposed to measure (Hubley and Zumbo 1996; Kelley 1927). A test is said to have sensitivity when it is able to identify those with disease and have specificity when it is able to identify those without disease (Parikh et al. 2008). Testing of validity and reliability of a test is construed in different forms. Construct validity is assumed whenever there is good correlation between constructs and responses from the measures (Teglasi et al. 2012). A tool is said to have construct validity whenever it is able to show response variations in relation to real life and the measured phenomenon. Discriminant and convergent validity are used to establish construct validity. Discriminant validity is established when two tools that are supposed to measure different phenomena demonstrate this difference. Convergent validity is established whenever two tests that are supposed to measure the same phenomenon show this similarity. Factor analysis also establishes construct validity by showing

whether a cluster of items that are supposed to be caused by the target constructs. As a note, in this review, when the form of construct validity is not specified as none of the three discussed above, it will be identified as just construct validity.

Studies that have previously reviewed neurocognitive tests have either reviewed tests relevant to specific diseases or other age groups with partial relevance to early schoolers (Bradley-Johnson 2001; Ezeamama et al. 2018; Stadskleiv 2020; Williams et al. 2014). The test specific reviews have published findings on psychometric properties and cultural relevance of different neurocognitive tests(Bradley-Johnson 2001). The current study furthers these findings and gives prominence to the early schoolers.

Study Objective

This narrative review looks at neurocognitive tools developed and standardized from 1987 to 2017 specifically for children ages 6–12 years. A narrative review is recommended for a critical discussion of knowledge on a topic of interest with the aim of collating and summarizing study findings on the topic as well as identifying research gaps (Ferrari 2015). The aims of this review are to identify and summarize commonly used neuropsychological tools among 6–12 years globally and their psychometric properties across different contexts. Specifically, the review aims at answering the following research questions:

1. Which standardized neurocognitive tools are commonly used among 6–12-year-olds?
2. Which cultural adaptations have been made to these tools?
3. What is the reliability, validity, sensitivity, and specificity of these tools?

Methods

We identified studies conducted between 1987 and 2017 through a thorough search of PubMed, Psych Info, and Web of Science using the keywords (i) neuropsychological or neurocognitive with (ii) assessment or test.

Following this search, we included original studies that examined any form of psychometric properties using neuropsychological tests among children aged 6–12 years globally. RM examined each study against the exclusion and inclusion criteria and determined whether it should be included in the review. Inclusion criteria: written in English language, use of Neuropsychological measures, children 6–12 years, all peer-reviewed published journal articles, publications between 1997 and 2017 and human subjects' research. Studies that partially covered the age criteria were also included in the review. Exclusion criteria: not in English language, neurophysiological measure, grey literature, full text missing and animals. Information concerning the type of neuropsychological assessment, cognitive domain measured (executive functions, perceptual motor, complex attention, language, learning and memory), study setting and type of standardization conducted was extracted. She developed a template of key findings on a spreadsheet and shared with other mentors. She received feedback from FV, AA, MPG and KM. There were 12 papers that lacked clarity in their psychometric findings where all the other authors reviewed these papers one by one. Out of these papers, three were selected on the basis that they did have results showing the tools' validity. Figure 1 shows the data extraction flow chart. Information from the papers was coded in terms of authors, country where the study was done, population of interest, tool examined and domains it covers, as well as the reliability and validity outcomes. This information was entered into an online Excel sheet that was accessible by all the authors. Cognitive domains and sub-domains were classified according to the *Diagnostic and Statistical Manual of Mental Disorders Fifth Edition* (DSM-5) (Sachdev et al. 2014) as shown in Fig. 2.

Results

The narrative review identified 306 papers, in which 145 papers met the inclusion criteria as indicated in Fig. 1. Figure 1 provides a data extraction flow chart (also see Appendix 1 in Supplementary Information). Most of the papers used

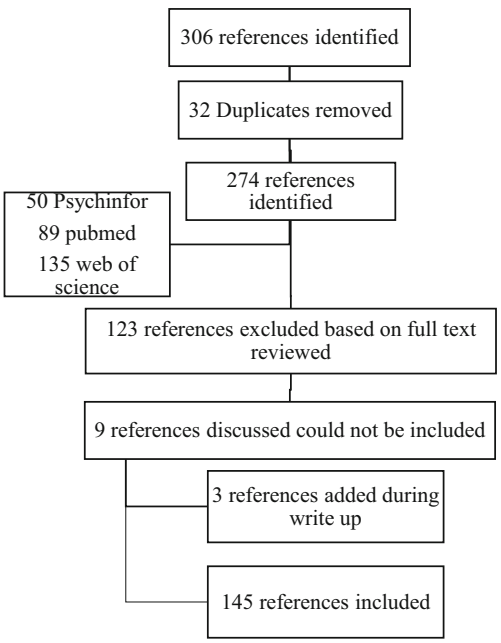

Fig. 1 Data extraction flow chart

Fig. 2 Classification of cognitive domains and sub-domains

Cognitive Domain	Sub-Domains
Executive Functions	Planning, decision making, working memory, responding to feedback, inhibition & flexibility.
Perceptual Motor	Visual perception, visuo-constructional, reasoning, perceptual-motor & coordination.
Complex attention	Divided attention, sustained attention, processing speed and selective attention.
Social Cognition	Recognition of emotions, insight & theory of mind.
Language	Object naming, word finding, fluency, grammar & syntax, & receptive language.
Learning and Memory	Free recall, cued recall, recognition memory, semantic & autobiographical, long term memory, implicit learning.

multiple tools, with the total frequency of different tools used amounting to 142. Twenty-three of the tools had a total frequency of ≥2. The majority of the studies were conducted in the clinical population ($N = 102$). The cognitive domain distribution of studies included 77 on executive function tests, 75 on complex attention tests, 49 on perceptual motor and motor tests, 21 on learning tests, 28 on language tests and 62 on memory tests, as well as 14 on tests distributed across arithmetic, social cognition, cognitive reserve, intelligence, representational competence and academic achievement domains. The validity and reliability terms as well as the statistical and adaptation criteria described are those referenced by the original study authors. Almost half of the studies were conducted in the USA as shown in Table 1.

Adaptation Processes in the Reviewed Studies

There were eleven papers that reported on formation of completely new tools compared to a hundred and thirty-four that adapted and/or tested the psychometric properties of already-existing tools.

Assembly

There are eleven studies that chose to develop new tools. (van Nieuwenhuijzen et al. 2011) developed a social information processing measure because there wasn't a tool that measured this cognitive domain. This measure involved using vignettes in combination with cartoons, picture and video which depicted different social situations, and the child was required to respond to different questions like what was happening and how they would respond in a similar situation. Scores were developed that evaluated the responses information processing trajectory within a linear scale. Shorter versions of tools were also created for screening purposes.

Adoption

Adoption involved translation and making iterations to the items. Tools whose adaptation included translation had to be keen in ensuring the new versions did not lose the characteristics of the original tool. This is because the respondent's language background tends to exert some effect on the tools.

Table 1 Country distribution of the extracted studies

	Total number of papers	Detailed description	*N* (%)
Countries	145	USA	65 (46.4)
		Canada	11 (7.9)
		Netherlands	6 (4.3)
		Brazil	5 (3.6)
		Australia, Kenya	4 (5.7)
		Finland, Italy, Spain	3 each (6.4)
		Taiwan, Colombia, France, Germany, Mexico, Hong Kong, Israel, Korea, Sweden, Uganda	2 each (15.7)
		Argentina, Austria, Bangalore, Belgium, Cyprus, Denmark, Faroe Islands, Japan, Morocco, Portugal, Romania, Iran, Thailand, UK	1 each (10.0)

It is for this reason that most cultural adaptations took language into account (Rohitrattana et al. 2014; Siqueira et al. 2016). Some of the tools did not rely heavily on language; hence, the new versions had to translate the instructions only (Bangirana et al. 2015). Bilingual translators were preferred in five studies and a back-translation design adapted for the translation (Bangirana et al. 2015; Fasfous et al. 2015; Hwang et al. 2006; Siqueira et al. 2016). Where two translators would not agree on instruction or stimuli translation, a third one would be integrated as a tie breaker.

Translation was only done after permission was sought from the original authors. This, at times, faced challenges where authors were not willing to give permission for development of a different version, or in situations where they did, permission was partial in that the developers gave access to for example the tool's stimuli (Siqueira et al. 2016).

Once translation was done, substitution was pursued for certain items that were unfamiliar to the respondents with items that were familiar (Kitsao-Wekulo et al. 2013b). In adapting a neurobehavioural test battery among Thai children, the authors substituted envelopes with paper as well as hair brush with hair clip (Rohitrattana et al. 2014). The later substitution was interestingly because of similar pronunciations to a toothbrush. An adaptation of the Child Hayling Test (CHT) among Brazilian children included the exclusive use of nouns instead of a mixture of nouns, adverbs and adjectives that were used in the adult version of the test (Siqueira et al. 2016). This was done to meet the linguistic preference of the Brazilian children population. These forms of changes are integrated in the stimuli and instructions. Usually, mental health practitioners, such as psychologists at postgraduate level, judged whether each item is representative of the cognitive domains for which the tool is supposed to measure, and whether they would be easily comprehended.

Practice effects were determined in one study by doing a paired *T*-test analysis (Kitsao-Wekulo et al. 2013a), while in another, percentage change and reliability change indexes were calculated (Llorente et al. 2009). Reliability change indexes take into account performance that is likely to be because of measurement unreliability. To reduce practice effects in test–retest reliability measurements, adaptation also involved creating alternative forms of the same tests. Creating alternate forms may not always be the best practice as a study among Thai children showed low test–retest reliability in tests with alternate forms (Rohitrattana et al. 2014). Comparability of alternate forms may need to be improved to reduce such effects. Sub-measures, as opposed to a full neuropsychological battery, have been targets for adaptation based on the objectives of the study (Reitan and Wolfson 2004; Thomas et al. 2016). Sadeh, Burns and Sullivan (2012) investigated the predictive power of the EF screener within Behavior Assessment System for Children-Teacher Report (BASC). An EF screener with strong predictive power would be useful in screening for behavioural problems early enough for preventive and intervention purposes.

Pilot

Ten pilot studies evaluated the linguistic, semantic and syntax complexities of the tools. P. K. Kitsao-Wekulo et al. (2013a) did a pilot study for the Kilifi Toolkit to check translation comprehension, familiarity of the items and ceiling and floor effects of the modifications, as well as ease of administration and scoring. Pilot studies exuded vital information such as the impact of examples in helping children understand the guidelines (Hwang et al. 2006).

Standardization

Validity and reliability estimates were evaluated for the tests in one hundred and forty-one papers depending on the objectives of the study in relation to the tool. Four papers sought to extract age-related test norms. Test–retest reliability was assessed using intraclass correlation (ICC) while internal consistency (extent to which items hang together) was evaluated using Cronbach alpha. Confirmatory factor analysis has been used in the studies to assess the tests' construct validity or assess how well the factor structure fits the test items. A good fit is one with a non-significant p value, a root mean square error of approximation (RMSEA) of less than .01 and a comparative fit index (CFI) of more than .90 (Rose et al 2011). Construct validity has also been assessed by identifying group difference between diseased and healthy samples based on their cognitive outcomes in the tests (Spironello et al. 2010). Discriminant (a tool's ability to differentiate those with cognitive impairment from those without) and convergent validity (two tools' ability to identify those with cognitive impairment) is part of construct validity. Another way used to look at the internal structure of a test is through factor analysis with eigenvalues among other calculations being done to evaluate the number of factors (Stinnett et al 2002). Concurrent validity, where the level of agreement between two tools is evaluated, was measured using Pearson's correlation coefficient (Spironello et al. 2010). Receiver operating characteristics (ROC) have been used to assess the sensitivity and specificity of tests i.e. the tests' scoring ability in differentiating those with cognitive impairment from those without (Thaler et al. 2010). Area under the curves (AUCs) have also been used with ROC to assess for group differences. An AUC of .80 and above indicates good classification which is synonymous with support for predictive discrimination. Sensitivity has also been assessed using univariate analyses of variance (ANOVAs). ANOVA has also been used in studies creating norms for tests where the effects of age and gender are

evaluated (Reynolds et al. 2016). Multiple regression analysis (MANOVA) gives a clearer picture of associations by removing confounding effects and measurement errors among other factors that influence outcomes. (Konstantopoulos et al 2015). chose to use MANOVA when creating normative data for CCTT where the relationship between completion time and age and gender was investigated. Structural equations do the same as they have been used to give an overall accurate estimation of associations (Budtz-Jorgensen et al. 2002). Higher sensitivity and specificity are predictive of the best cut-off points/scores when assessing for impairment in children. Test of Memory and Learning (TOMAL) evaluation indicated that a cut-off point of .80 indicated the best sensitivity and specificity combination (sensitivity .70, specificity .62) (Thaler et al. 2010). Criterion validity has been used to evaluate further the internal structure of a tool by elucidating the test's ability to denote the severity of the cognitive impairment (Woodward and Donders 1998).

The population chosen to test the tools' psychometric properties is based on the objectives of the study and the population most likely to exhibit cognitive impairment. Thirty-seven papers chose an entirely healthy population to study, while thirty-six chose a population with a healthy control and seventy-two chose an entirely diseased population depending on the cognitive deficit of interest to the study. Thirty-seven papers had populations with attention deficit hyperactive disorder representing the most (26%) preferred population in the studies.

There are studies which chose to adapt the test among males only ($n = 2$) and another on female only ($n = 1$) instead of both genders (Carone 2014; Termine et al. 2016). As much as gender is highlighted as a confounder in research, in neurocognitive adaptation studies, gender effect on cognitive measurements has been found to be insignificant. (Roy et al. 2015). found that gender was insignificant in executive function measurements.

Cognitive Domains Psychometric Checks

The psychometric results of different measures are outlined, and they are organized into the neurocognitive domains the tools measure. The description of the results as either poor/weak, moderate and good/high is according to the original study authors' classification of findings. In the main text, a summary is provided, and detailed information on the countries where the studies were conducted as well as the specific psychometric outcomes with actual numbers (including the presence and absence of specific psychometric checks and the reported statistics) are in Appendix 1 of the supplementary materials.

Executive Function Tests Standardization Outcomes

The Behaviour Rating Inventory of Executive Functioning (BRIEF) had the highest number of standardization studies ($N = 7$). It passed validation indicators though reliability studies were yet to be done. The WISC III and IV reported good validity though reliability indicators varied with regard to subsets under study. The Digit Span subtest of the WISC III had low test–retest reliability (Table 2).

Memory Tests Standardization Outcomes

The Medical Symptom Validity Test (MSVT) was the most heavily researched on ($N = 6$) closely followed by the Test of Memory Malingering (TOMM and TOMM 2) and Word Memory Test (WMT) each having five studies looking at their psychometric properties. TOMM had varying studies indicating different findings with regard to validity, specificity and sensitivity. The other two tests showed high validity (Table 3).

Complex Attention Standardization Outcomes

The Continuous Performance Test (CPT) and its revisions had the highest number of studies ($N = 6$) looking into its psychometric properties. Different studies found differing standardization outcomes as indicated in Table 3. CANTAB came in second and its general validity was established though its subtests, spatial working memory (SWM), had low construct validity. CANTAB's test–retest reliability was also found to be low (Tables 4).

Motor and Perceptual Motor Standardization Outcomes

Six studies looked at the Developmental Test of Visuo-Motor Integration psychometric indicators. The studies had differing findings when it came to discriminant validity and test–retest reliability. General validity was established, but two studies could not agree on the discriminant validity of the tool as one reported the validity to be poor (Table 5).

Learning Standardization Outcomes

Cogstate Battery, WISC IV, Differential Ability Scales (DAS) and NEPSY were the most frequently studied tests (count ≥2). Cogstate validity was not questionable, but two studies found its test–retest reliability to range from weak/low to strong. NEPSY had similar reliability outcomes (Table 6).

Language Standardization Outcomes

The language tests had equal variance on frequency of studies done. The WSC IV vocabulary test was found to have no

Table 2 Executive function standardized tests among 6–12-year-olds

Executive function tests	Frequency of studies	Validity	Reliability	Normative data
Tower of Hanoi Test	2	Construct validity high	Test–retest reliability high in one study and low in another	–
Tower of London	1	General validity significant	–	–
Storytelling performance measure of EF	1	–	Reliability—intraclass correlation (ICC) and internal consistency reliability excellent	–
Self-Ordered Pointing (SOP)	1	–	Test–retest reliability moderate	Normative data for 7–12 years
A standard Stroop (Golden Version); Sun-Moon Stroop and Fruit Stroop	1	–	Test–retest reliability strong for Sun-Moon Stroop and Fruit Stroop	Normative data for 7–12 years
CogState battery	2	Construct validity good (3 factor structure), concurrent and convergent validity partially significant; general validity partially significant	Test–retest reliability moderate in one study and moderate to high in another	–
Children's Kitchen Task Assessment (CKTA)	2	Discriminant partial significant, concurrent low to moderate	Interrater high and internal consistency moderate; interclass correlation and internal consistency high	–
Five to Fifteen parent questionnaire (FTF)	1	Criterion and discriminant partial significance, internal consistency high	–	–
Wisconsin Card Sorting Test (categories, failure to maintain set, total errors)	1	–	Test–retest reliability low	–
Delis-Kaplan Executive Function System (D-KEFS) (Trail Making—visual scanning, number sequencing, motor speed, total errors; Verbal Fluency—set loss errors, repetition errors; Tower Test—rule violation/item ratio)	1	–	Test–retest reliability low	–
Children's Cooking Task (CCT)	1	Discriminant high and concurrent significant for some tests	Internal consistency and test–retest reliability high	–
The ecological 'cooking task'	1	Discriminant validity significant	Inter-rater high	–
Trail-Making Test (TMT).	1	Discriminant validity partially significant	–	
Digit span	2	Discriminant validity partially significant; general validity poor	–	–
Korean Educational Development Institute-Wechsler Intelligence Scales (KEDI-WISC) (subtests include Continuous Performance Test (CPT), Children's Colour Trails Test (CCTT) and Stroop Colour-Word Test (SCWT))	1	General validity partial significance	–	–
Amsterdam Neuropsychological Tasks (ANT) subtests: baseline speed, focused attention four letters, shifting attentional set–visual (measures vigilance, inhibition and cognitive flexibility) and sustained attention	1	Discriminant validity partial, sensitivity moderate, specificity moderate	–	–
Behaviour Rating Inventory of Executive Functioning (BRIEF)	6	Concurrent no significance; convergent significant; general validity partially significant; concurrent validity partial significance and discriminant validity significant; general validity partially significant; ecological validity partially significant		
Luria-Nebraska Test for Children (TLN-C, in Portuguese)	1	General validity high	Internal consistency high	–

Table 2 (continued)

Executive function tests	Frequency of studies	Validity	Reliability	Normative data
FAS Verbal Fluency Test	1	General validity partial significance	–	–
Arizona Cognitive Test Battery (ACTB)	1	–	Test–retest partial	–
Cattell–Horn–Carroll (CHC)	1	General validity partial significance	–	–
BaterÃa de EvaluaciÃ³n NeuropsicolÃ³gica Infantil (BENCI)	1	Discriminant validity high	Test–retest reliability moderate to high	–
The Cambridge Neuropsychological Test Automated Battery (CANTAB)—subsets include pattern recognition memory (PMR), spatial recognition memory (SRM), spatial span (SSP), Stockings of Cambridge (SOC), intra–extra dimensional set shift (IED), reaction time (RTI), rapid visual information processing (RVP)	1	Construct validity good	Internal consistency poor to high	–
n-back	1	Criterion validity good. Factorial structure	Internal consistency high	–
Wechsler Intelligence Scale for Children v 3 (WISC–III)	1	General validity significant	–	–
Wechsler Intelligence Scale for Children v 3 (WISC–III) Symbol Search subtest	2	Convergent validity partial	Test–retest reliability poor to good and in another study moderate to high	–
Wechsler Intelligence Scale for Children v 3 (WISC–III) Coding subtest	2	Convergent validity partial; general validity not significant	Test–retest reliability poor to good	–
Wechsler Intelligence Scale for Children v 3 (WISC–III) Digit Span subtest	1	–	Test–retest reliability moderate to high	–
Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV)	1	–	Test–retest reliability low to excellent	–
Wechsler Intelligence Scale for Children Fourth Edition (WISC IV)–General Ability Index (GAI), Full Scale IQ (FSIQ) and Cognitive Proficiency Index (CPI)	2	Sensitivity high; general validity partial—VcSiMrBd subtests highest accuracy estimate for GAI	–	–
Children's Category Test – Level 2 (CCT-2)	1	Criterion partial, discriminant partially significant, sensitivity partial	–	–
Japanese short form of the Swanson Cognitive Processing Test	1	Concurrent validity moderate	Test–retest reliability high	–
Reynolds Intellectual Assessment Scale (RIAS)—subtests include Composite Intelligence Index (CIX), Nonverbal Intelligence Index (NIX) and Verbal Intelligence Index (VIX)	1	Construct validity partial	–	–
The Children's Executive Functions (CEFS)	1	General validity partially significant	–	–
Behavioural screener for the assessment of executive functions version 2 (BASC-2-EF) screener	1	General reliability—adequate to strong; construct good	Internal consistency high	–
EF scale from the Behaviour Assessment System for Children-Teacher Report	1	Construct validity good, predictive validity weak and partially significant	Reliability high	–
Testbatterie zur Aufmerksamkeitsprufung fur Kinder (KITAP)	1	Discriminant validity partial		–
Clock test (clock drawing test, clock face test)	1	Discriminant validity partial	Interrater reliability high	–
Brief neurocognitive screener (DIVERGT)—subtests Digit Span Test, The Verbal Fluency Test, The Grooved Pegboard Test and The Trail Making Test	1	Sensitivity and specificity—moderate to high. Predictive validity significant, discriminant validity significant	Test–retest reliability good	–
Autism/Tics, AD/HD and other Comorbidities (A&TAC) inventory	1	–	Intrarater and interrater moderate to high	–
Korean Computerized Neurobehavioral Tests (KCNT)—subtests include Simple Reaction Time (response speed), Choice Reaction Time (psychomotor speed),	1	Test–retest moderate	–	–

Table 2 (continued)

Executive function tests	Frequency of studies	Validity	Reliability	Normative data
Colour Word Vigilance (attention), Addition (executive functions), Symbol Digit (executive functions) and Finger Tapping Speed (manual dexterity)				
Halstead–Reitan Neuropsychological Test Battery for Older Children (HRNB-C)	1	Construct small to large correlations	Reliability high	–
Halstead–Reitan Neuropsychological Test Battery for Children (HRNB-C)	1	Construct validity significant	–	–
Halstead–Reitan Neuropsychological Test Battery—Trail Making Test	1	Sensitivity high, discriminant validity significant		
Kaufman Assessment Battery for Children, second edition (KABC-II)	3	Construct validity high, predictive validity low to high; reliability good, construct validity good; construct good: yielded five factors (sequential processing, Simultaneous processing, planning and learning)	Test–retest reliability moderate to high	–
Online version of IMPACT	1	–	Test–retest reliability poor to good	–
Pediatric ImPACT	1	Convergent significant moderate to high correlations, discriminant significant moderate		–
Immediate Post concussion Assessment and Cognitive Testing (ImPACT)	1	–	–	Norms established for children aged 10–– 12 years
Omnibus test of cognitive functioning; Trail Making A (attention), Continuous Performance Task (CPT) (attention)]; Trail Making B (Executive Function); Cog Set Shifting (Executive Function), Controlled Oral Word Association Test (COWAT) (Executive Function); Digit Span (Working Memory), Spatial Span (Working Memory), and California Verbal Learning Test (CVLT)(Verbal Memory)	1	General validity significantly strong	Interrater reliability high	–
Timo's Adventure	1	Discriminant validity high, sensitivity high, specificity high	–	–
Combination of Kaufman Hand Movements Scale; The Stroop Colour-Word Association Test (Stroop); The Controlled Oral Word Association Test (COWAT); Trail Making Test; Arithmetic and Digit Span subtests of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III); Conners' Continuous Performance Test (CPT)	1	Discriminant validity not significant, sensitivity and specificity low to high	–	–
Neuropsychological Battery: subtests Mental Control; Target Detection Cancellation Test; Visual-Verbal Learning Curve; Rey-Osterrieth Complex Figure Test; Language Comprehension and Working Memory test; Language Fluency test; Wisconsin Card Sorting Test-Abbreviated Version (WCST-A)	1	Construct validity good, discriminant validity poor, sensitivity and specificity poor to high	–	–
Lebby-Asbell Neurocognitive Screening Examination—Children and Adolescent versions (LANSE-C/A)	1	Discriminant validity not significant, convergent validity partial significance	Internal consistency low to high	–

Table 2 (continued)

Executive function tests	Frequency of studies	Validity	Reliability	Normative data
Pediatric Attention Disorders Diagnostic Screener (PADDS)	1	Concurrent validity strong	–	–
Swanson, Nolan and Pelham Questionnaire (SNAP-IV scale)	1	General validity poor	–	–
Behavioural Assessment of Dysexecutive Syndrome for Children (BADS-C) (subtests: Playing Cards test, Water test, Key search test, Zoo map tests, Six parts test)	2	Convergent validity weak and non-significant; ecological validity partial significance	Interrater moderate to high	–
Developmental Neuropsychological Assessment (NEPSY)	2	Discriminant validity significant; construct validity partial correlations, specificity low to high	Reliability moderate to high	–
Groton Maze Learning Task (GMLT)	1	Construct validity partially significant	–	–
Child Hayling Test (CHT)	1	Content validity high, sensitivity partial	–	–
The Corsi test	1	General validity significant	–	–
A Maze task	1	Discriminant validity significant, convergent validity partially significant	–	–
	77			

validity, and the Seashore Rhythm Test had low internal consistency. Most of the language tools had no validity indicators (Table 7).

Other Tests' Standardization Outcomes

There was no validity for the cognitive reserve subtest within WIAT-II. Tests used for social cognition were found to be valid including interesting tools such as cartoons, pictures and video vignettes (Table 8).

Tools Tested in LMIC Including Sub-Saharan Africa (SSA)

Six studies from SSA four in Kenya and two in Uganda were included. In Uganda, the authors tested construct, concurrent and convergent validity, as well as test–retest reliability for the computerized, self-administered CogState battery and construct validity for the KABC-II (Bangirana et al. 2009; Bangirana et al. 2015). Moderate test–retest coefficient correlations were found while concurrent and convergent validity correlations were found with tools such as KABC-II and TOVA. In Kenya, on the other hand, internal consistency was tested for Tower Test (planning), Self-Ordered Pointing Test (SOPT; verbal/visual selective reminding), Verbal List Learning (VLL; working memory), Colored Progressive Matrices (CPM; reasoning), Dots (nonverbal memory), Contingency Naming Test (CNT; attention and attention shift, Score (auditory sustained and selective attention), as well as People Search (visual sustained and selective attention) (Kitsao-Wekulo et al. 2013a). Test–retest reliability for immediate memory span and CNT was found to be below acceptable levels while the other subtests had marginally to acceptable reliability. Internal consistent results ranged from .70 to .84. The sensitivity, specificity and test–retest reliability of the Ten Questions Questionnaire, which measures perceptual motor and memory domains, was also tested among 6–9-year-old Kenyan children (Mung'ala-Odera et al 2004). Test–retest reliability was found to be excellent for motor, vision, speech and four cognition questions while specificity and sensitivity rates were greater than 70% and 96% respectively.

Discussion

This narrative review covered studies on adaption and standardization of neurocognitive tools that were done in between 1987 and 2017 among children aged 6–12 years old. The narrative review investigated the standardized tools that are commonly used and the cultural adaptations made to these tools, as well as the reliability, validity, sensitivity and specificity of these tools.

Commonly Used Tools and Psychometric Outcomes

The cognitive domains covered were exhaustive of the DSM-5 classification though tools that covered executive functions, complex attention and memory domains were the most researched on tools. The child neuropsychological test

Table 3 Memory standardized tests among 6–12-year-olds

Memory tests	Frequency of studies	Validity	Reliability	Normative data
California Verbal Learning Test, Children's Version CVLT-C	2	Sensitivity and specificity ranged from moderate to high; construct validity good—yielded a 4-factor model consisting of Attention Span, Learning Efficiency, Delayed Recall, and Inaccurate Recall	Reliability good	–
QS4-G: Parent Questionnaire for the Developmental Evaluation of 4-Year-Old	1	Sensitivity moderate to high, specificity high, predictive high apart from academic difficulties	–	–
Test of Memory and Learning (TOMAL)	1	Criterion good, discriminant significant, convergent partial significance, factorial analysis produced 5 factors; sensitivity and specificity low to high	–	–
Word Completion Memory Test (WCMT)	1	Specificity high; validity partial significance	–	–
The Test of Memory Malingering (TOMM); TOMM 2	6	General validity high, specificity high, sensitivity high; specificity high—vary according to disorder; TOMM 2 sensitivity and specificity highly accurate, general validity significant; TOMM 2 performance validity established; TOMM predictive validity partially significant; TOMM general validity partially significant, sensitivity low, specificity good; TOMM specificity high, general validity partially significant	–	–
Medical Symptom Validity Test (MSVT)	6	General validity high, sensitivity high, specificity high; performance validity not significant and specificity high; general validity good; performance validity good	–	–
Fifteen Item Test (FIT)	1	General validity high	–	–
Word Memory Test (WMT)	4	General validity moderate to high; performance validity not significant, specificity high; performance validity good; general validity partially significant, specificity high	–	–
Nonverbal Medical Symptom (NV-MSVT).	1	Performance validity not significant, specificity high	–	–
Five to Fifteen parent questionnaire (FTF)	1	Criterion and discriminant partial significance	Internal consistency high	–
Memory Screening Index (MSI) from the WRAML (Wide Range Assessment of Memory and Learning)	1	Factor structure good, criterion significant	–	–
Rey's Auditory-Verbal Learning Test (AVLT).	1	–	Test–retest reliability low to high	–
Children's Memory Scale	1	–	Test–retest reliability low	–
Word List Delayed Recognition	1	–	Test–retest reliability low	–
Trail-Making Test (TMT).	1	Discriminant validity partially significant	–	–
Amsterdam Short-Term Memory (ASTM)	1	Specificity high for 9 years and above, general validity partially significant	–	–
Luria-Nebraska Test for Children (TLN-C, in Portuguese)	1	General validity high	Internal consistency high	–
Arizona Cognitive Test Battery (ACTB)	1	–	Test–retest partial	–
Cattell-Horn-Carroll (CHC)	1	General validity partial significance	–	–
BaterÃa de EvaluaciÃ³n NeuropsicolÃ³gica Infantil (BENCI)	1	Discriminant validity high	Test–retest reliability moderate to high	–
Cambridge Neuropsychological Test Battery (CANTAB)	3	General validity partially significant; general validity partial significant, construct good	Test–retest reliability low	–
The Cambridge Neuropsychological Test Automated Battery (CANTAB)—subsets	1	Construct validity good	Internal consistency poor to high	–

Table 3 (continued)

Memory tests	Frequency of studies	Validity	Reliability	Normative data
include Pattern recognition memory (PMR), Spatial recognition memory (SRM), Spatial span (SSP), Stockings of Cambridge (SOC), Intra-extra dimensional set shift (IED), Reaction time (RTI), Rapid visual information processing (RVP)				
WISC-IV Digit Span subtest	1	Specificity high; sensitivity high	–	–
Differential Ability Scales (DAS). Differential Ability Scales - Second Edition (DAS II)	2	Discriminant validity good; predictive validity for DASII high	–	–
CNS Vital Signs (CNSVS)—subtests: verbal and visual memory, finger tapping, symbol digit coding, the Stroop Test, a test of shifting attention and the continuous performance test	1	Concurrent validity moderate and discriminant validity good	Test–retest reliability moderate to high	–
Children's Category Test – Level 2 (CCT-2)	1	Criterion partial, discriminant partially significant, sensitivity partial	–	–
Kilifi Toolkit—subtests include Tower Test, Self-Ordered Pointing Test, Verbal List Learning, Coloured Progressive Matrices, Dots, Contingency Naming Test, Score, People Search	1	Predictive validity partially significant	Internal consistency moderate, test–retest low to moderate	–
Brief neurocognitive screener (DIVERGT)—subtests Digit Span Test, The Verbal Fluency Test, The Grooved Pegboard Test and The Trail Making Test	1	Sensitivity and specificity moderate to high; predictive validity significant, discriminant validity significant	Test–retest reliability good	–
Perceived cognitive function (PCF)	1	Discriminant validity significant	–	–
Autism/Tics, AD/HD, and other Comorbidities (A&TAC) inventory	1	–	Intrarater and interrater moderate to high	–
Kaufman Assessment Battery for Children, second edition (KABC-II)	3	Construct validity high, predictive validity low to high; construct validity good; construct good: yielded five factors (sequential processing, simultaneous processing, planning and learning)	Test–retest reliability moderate to high; reliability good in another study	–
Standardised Assessment of Concussion (SAC)	1	Convergent validity partial	Test–retest poor to good	–
Ten Questions' Questionnaire (TQQ)	1	Sensitivity high, specificity high	Test–retest fair to excellent, interrater good to excellent	–
Pediatric ImPACT	1	Convergent significant moderate to high correlations, discriminant significant moderate	–	–
Immediate Post concussion Assessment and Cognitive Testing (ImPACT)	1	–	–	Norms established for children aged 10–12 years
CMS Delayed Verbal Recall>Delayed Recognition memory subtests	1	Specificity high, sensitivity high	–	–
Neuropsychological Battery: subtests Mental Control; Target Detection Cancellation Test; Visual-Verbal Learning Curve; Rey-Osterrieth Complex Figure Test; Language Comprehension and Working Memory test; Language Fluency test; Wisconsin Card Sorting Test-Abbreviated Version (WCST-A)	1	Construct validity good, discriminant validity poor, sensitivity and specificity poor to high	–	–
Lebby-Asbell Neurocognitive Screening Examination—Children and Adolescent versions (LANSE-C/A)	1	Discriminant validity not significant, convergent validity partial significance	Internal consistency low to high	–
	1	General validity partial		–

Table 3 (continued)

Memory tests	Frequency of studies	Validity	Reliability	Normative data
Behavioural Assessment and Research System (BARS) (included tests of motor speed and dexterity, attention, memory and visuospatial coordination)			Test–retest low (for tests with alternate forms) to high (for tests without alternate forms)	
Swanson, Nolan and Pelham Questionnaire (SNAP-IV scale)	1	General validity poor	–	–
Developmental Neuropsychological Assessment (NEPSY)	2	Discriminant validity significant; construct validity partial correlations, specificity low to high	Reliability moderate to high	–
Groton Maze Learning Task (GMLT)	1	Construct validity partially significant	–	–
	62			

findings reviewed in this paper reported on mostly executive functioning standardization outcomes where BRIEF was the most researched on tool ($N = 6$) followed by KABC-II ($N = 3$). Validity indicators for the BRIEF showed partial-to-low correlation outcomes with only discriminant validity being wholly significant when it came to its three composite scores or scale scores, as well as comparison of its teacher-rated to parent-rated versions. BRIEF may have been a common tool due to the ease of administration through the parents (Vriezen and Pigott 2002). KABC-II construct validity was supported in all the studies though its predictive validity and reliability findings were rated as low to moderate. KABC-II was among the few executive function tools to be standardized in LMIC despite its complexity in administration (Bangirana et al. 2009).

Complex attention standardization outcomes were mainly reported for the CPT ($N = 6$) and Attention Network Test (ANT) ($N = 5$). The later had low reliability outcomes with only one study reporting moderate to high test–retest reliability findings. The validity outcomes were, however, high proving that the tool has good internal validity. CPT was also popularly studied, and this could have led to the very many developed versions of it which continue to be updated. Moreover, it has good discriminant validity indicators with moderate test–retest reliability. However, the specificity and sensitivity indicators range from moderate to high and the general validity was found to be partially significant.

Medical Symptom Validity Test (MSVT) ($N = 6$), The Test of Memory Malingering (TOMM) (N = 6) and Word Memory Test (WMT) (N-4) were commonly studied under the memory domain. WMT showed mixed results when it came to validity outcomes, but specificity was endorsed as high in two studies. This trend was not seen in MSVT which showed good validity and specificity outcomes while TOMM had mixed findings where validity was indicated as partially significant in some studies, specificity high and sensitivity as low. In some cases, insufficient effort could have affected the variability in validity and sensitivity outcomes.

Visuo-Motor Integration was the only perceptual motor prevalently studied test ($N = 6$) with mixed discriminant, validity findings but good convergent, construct, concurrent and criterion validity. Test–retest reliability ranged from low to high in varied studies while inter-rater reliability was ranked as high in one study. The popularity of this tool could be attributed to ease of administration (Ahonniska et al. 2001) especially due to the age of our population of interest or it could also be due to being among the very few tests that are available for the perceptual motor domain.

Neuropsychological batteries, tests that have several subtests within them, may not have been attributed as common as they were broken down into their respective subtests cognisant to the cognitive domain covered. They were, however, also widely studied. The tests include the Wechsler Intelligence Scale for Children (WISC), Halstead-Reitan Neuropsychological Test Battery for Children (HRNB-C) and Cambridge Neuropsychological Test Automated Battery (CANTAB). HRNB-C was found to have good discriminant and construct validity while reliability and sensitivity were found to be high. CANTAB as well was found to have good construct validity though internal consistency ranged from poor to high in between the subtests (Syvaoja et al. 2015). WISC III and IV subsets were commonly studied with reliability findings ranging from poor to high depending on the subtest while validity outcomes showed the same partial trend.

CogState battery along with other few tests have been validated in Africa (Bangirana et al. 2015; Holding et al. 2004; Mung'ala-Odera et al. 2004). In as much as only six studies have been conducted across Kenya and Uganda, the number of tests covered is nearly exhaustive of the cognitive domains identified as vital in DSM-5. Executive functions covered include planning, working memory and reasoning; complex attention subdomains covered include attention and attention shift/ selective attention; memory subdomains include non-

Table 4 Complex attention standardized tests among 6–12-year-olds

Complex attention measures	Frequency of studies	Validity	Reliability	Normative data
CogState battery	2	Construct validity good (3 factor structure), concurrent and convergent validity partially significant, general validity partially significant	Test–retest reliability moderate and moderate to high in another study	–
Continuous Performance Tests (CPT), MOXO-CPT, Conners' Continuous Performance Test (CCPT), computerized Corner's continuous performance test (CPT) – Second Edition	6	Discriminant high apart from impulsivity for MOXO-CPT and discriminant significant for original CPT established. CCPT has partial general validity and specificity is partial. CPT general validity nonsignificant, sensitivity moderate and specificity high	Computerized Corner's continuous performance test (CPT) – Second Edition test–retest reliability moderate	–
QS4-G: Parent Questionnaire for the Developmental Evaluation of 4-Year-Old	1	Sensitivity moderate to high, specificity high, predictive high apart from academic difficulties	–	–
Test of Memory and Learning (TOMAL)	1	Criterion good, discriminant significant, convergent partial significance, factorial analysis produced 5 factors; sensitivity and specificity low to high	–	–
Gordon Diagnostic System (GDS)	1	Construct validity—GDS scores yielded three factors: (a) delay, (b) vigilance correct and distractibility correct, and (c) distractibility errors and vigilance errors; general validity partial	–	–
NIH Toolbox Pattern Comparison Processing Speed Test	1	Convergent and discriminant validity range from low to high depending on test and age group	Test–retest reliability moderate	–
Digit span	2	Discriminant validity partially significant; general validity poor	–	–
Cancellation test	1	Discriminant validity partially significant	–	–
Circle-Tracing Task	1	Discriminant validity partially significant	–	–
Korean Educational Development Institute-Wechsler Intelligence Scales (KEDI-WISC) (subtests include Continuous Performance Test (CPT), Children's Colour Trails Test (CCTT) and Stroop Colour-Word Test (SCWT))	1	General validity partial significance	–	–
Continuous Attention Test for Children (CAT)	1	Discriminant partially significant and convergent weak	–	–
Amsterdam Neuropsychological Tasks (ANT) subtests: baseline speed, focused attention four letters, shifting attentional set–visual (measures vigilance, inhibition, and cognitive flexibility) and sustained attention	1	Discriminant validity partial, sensitivity moderate, specificity moderate	–	–
FAS Verbal Fluency Test	1	General validity partial significance	–	–
Arizona Cognitive Test Battery (ACTB)	1	–	Test–retest partial	–
Cattell-Horn-Carroll (CHC)	1	General validity partial significance	–	–
BaterÃa de EvaluaciÃ³n NeuropsicolÃ³gica Infantil (BENCI)	1	Discriminant validity high	Test–retest reliability moderate to high	–
Cambridge Neuropsychological Test Battery (CANTAB)	3	General validity partially significant; general validity partial significant, construct good	Test–retest reliability low	–
The Cambridge Neuropsychological Test Automated Battery (CANTAB)—subsets include pattern recognition memory (PMR), spatial recognition memory (SRM), spatial span (SSP), Stockings of Cambridge (SOC), intra–extra dimensional set shift (IED),	1	Construct validity good	Internal consistency poor to high	–

Table 4 (continued)

Complex attention measures	Frequency of studies	Validity	Reliability	Normative data
reaction time (RTI), rapid visual information processing (RVP)				
Attentional Network Test (ANT)	5	Criterion validity good; reliability poor; cue validity effect significant; internal validity high	Test–retest reliability low; internal consistency low; test–retest moderate to high	–
Wechsler Intelligence Scale for Children Freedom-from-Distractibility/Working Memory Index (FDI/WMI) and Processing Speed Index (PSI) (both subtests contribute towards FSIQ)	1	Construct validity high and general validity partial	–	–
10 Wechsler Intelligence Scale for Children-Third Edition (WISC-III) subtests and 4 Wechsler Individual Achievement Test (WIAT) subtests	1	External validity partially significant	Reliability good	–
Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV)	1	–	Test–retest reliability low to excellent	–
Wechsler Intelligence Scale for Children Fourth Edition (WISC IV)- General Ability Index (GAI), Full Scale IQ (FSIQ) and Cognitive Proficiency Index (CPI)	2	Sensitivity high; general validity partial—VcSiMrBd subtests highest accuracy estimate for GAI	–	–
CNS Vital Signs (CNSVS)—subtests: verbal and visual memory, finger tapping, symbol digit coding, the Stroop Test, a test of shifting attention and the continuous performance test	1	Concurrent validity moderate and discriminant validity good	Test–retest reliability moderate to high	–
EF scale from the Behaviour Assessment System for Children-Teacher Report	1	Construct validity good, predictive validity weak and partially significant	Reliability high	–
Testbatterie zur Aufmerksamkeitsprufung fur Kinder (KITAP)	1	Discriminant validity partial	–	–
Kilifi Toolkit—subtests include Tower Test, Self-Ordered Pointing Test, Verbal List Learning, Coloured Progressive Matrices, Dots, Contingency Naming Test, Score, People Search	1	Predictive validity partially significant	Internal consistency moderate, test—retest low to moderate	–
Children's Colour Trails Test (CCTT), 1 2 CCTT	2	Construct good—three-factor solution	1 2 CCTT test–retest reliability moderate	Normative data
Brief neurocognitive screener (DIVERGT)—subtests Digit Span Test, The Verbal Fluency Test, The Grooved Pegboard Test and The Trail Making Test	1	Sensitivity and specificity moderate to high, predictive validity significant, discriminant validity significant	Test–retest reliability good	–
Perceived cognitive function (PCF)	1	Discriminant validity significant	–	–
Autism/Tics, AD/HD and other Comorbidities (A&TAC) inventory	1	–	Intrarater and interrater moderate to high	–
Korean Computerized Neurobehavioral Tests (KCNT)—subtests include Simple Reaction Time (response speed), Choice Reaction Time (psychomotor speed), Colour Word Vigilance (attention), Addition (executive functions), Symbol Digit (executive functions) and Finger Tapping Speed (manual dexterity)	1	–	Test–retest moderate	–
Halstead-Reitan Neuropsychological Test Battery for Older Children (HRNB-C)	1	Construct small to large correlations	Reliability high	–
Halstead-Reitan Neuropsychological Test Battery for Children (HRNB-C)	1	Construct validity significant	–	–
Halstead–Reitan Neuropsychological Test Battery—Trail Making Test	1	Sensitivity high, discriminant validity significant	–	–
Test of Variables of Attention (TOVA)	1	–		–

Table 4 (continued)

Complex attention measures	Frequency of studies	Validity	Reliability	Normative data
			Internal consistency moderate to high	
Kaufman Assessment Battery for Children, second edition (KABC-II)	3	Construct validity high, predictive validity low to high; construct validity good; construct good: yielded five factors (Sequential Processing, Simultaneous Processing, Planning and Learning)	Test–retest reliability moderate to high; reliability good	–
Standardised Assessment of Concussion (SAC)	1	Convergent validity partial	Test–retest poor to good	–
Trail Making Test B (Trails B)	2	Convergent validity non-significant; general validity significant, functional equivalence partial	Test–retest reliability moderate	–
Trail Making Test A (Trails A)	1	General validity significant, functional equivalence partial	–	–
Online version of IMPACT	1	–	Test–retest reliability poor to good	–
Clinical virtual reality VR/Classroom-CPT (VC) (attention)	1	Diagnostic validity significant	–	–
Pediatric ImPACT	1	Convergent significant moderate to high correlations, discriminant significant moderate	–	–
Immediate Post concussion Assessment and Cognitive Testing (ImPACT)	1	–	–	Norms established for children aged 10–-12 years
Parent Report Child Behavioural Checklist (CBCL)	1	Sensitivity high, specificity high, predictive validity significant	–	–
CMS Delayed Verbal Recall>Delayed Recognition memory subtests	1	Specificity high, sensitivity high	–	–
Combination of Kaufman Hand Movements Scale; The Stroop Colour-Word Association Test (Stroop); The Controlled Oral Word Association Test (COWAT); Trail Making Test; Arithmetic and Digit Span subtests of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III; Conners' Continuous Performance Test (CPT)	1	Discriminant validity not significant, sensitivity and specificity low to high	–	–
Neuropsychological Battery: subtests Mental Control; Target Detection Cancellation Test; Visual-Verbal Learning Curve; Rey-Osterrieth Complex Figure Test; Language Comprehension and Working Memory test; Language Fluency test; Wisconsin Card Sorting Test-Abbreviated Version (WCST-A)	1	Construct validity good, discriminant validity poor, sensitivity and specificity poor to high	–	–
Lebby-Asbell Neurocognitive Screening Examination—Children and Adolescent versions (LANSE-C/A)	1	Discriminant validity not significant, convergent validity partial significance	Internal consistency low to high	–
Pediatric Attention Disorders Diagnostic Screener (PADDS)	1	Concurrent validity strong	–	–
Behavioural Assessment and Research System (BARS) (included tests of motor speed and dexterity, attention, memory, and visuospatial coordination)	1	General validity partial	Test–retest low (for tests with alternate forms) to high (for tests without alternate forms)	–
Swanson, Nolan and Pelham Questionnaire (SNAP-IV scale)	1	General validity poor	–	–

Table 4 (continued)

Complex attention measures	Frequency of studies	Validity	Reliability	Normative data
DiViSA—Discriminación Simple de Árboles/Simple Tree Discrimination Test	1	Discriminant good, sensitivity high, specificity high	Reliability high	–
Developmental Neuropsychological Assessment (NEPSY)	2	Discriminant validity significant; construct validity partial correlations, specificity low to high	Reliability moderate to high	–
Go/No-Go paradigm	1	Discriminant validity significant, convergent validity partially significant	–	–
A Maze task	1	Discriminant validity significant, convergent validity partially significant	–	–
	75			

verbal memory; while perceptual motor sub-domains include visuomotor coordination and visuospatial perception. In addition, CogState's reported construct, concurrent and convergent validity means that measurement of neurocognitive deficits is at par with other states especially considering that the CogState battery is computerized; hence, its administration and scoring is easy.

The form of standardization conducted in these tests is diverse, and though not comprehensive in some tests like Tower Test, the tools other psychometric properties have been tested in other settings like in London and Central Finland as is the case with the Tower Test (Ahonniska et al. 2000; Bishop et al. 2001). The validity and reliability findings of the tests in this review were also not widely spread across settings especially in the case of discriminant validity which despite most studies in this narrative review reporting on it, none of the studies conducted in sub-Saharan Africa reported on this form of validity. This is despite some studies having healthy and diseased populations that could be used to calculate discriminant validity of the cognitive tools. This selective testing of validity has been found to be because of authors' preference for what is relevant to them and what is easily obtained (Hubley and Zumbo 1996). Authors tend to choose the type of validity to be tested based on the purposes for which they would like the test to be used. If they want to see whether the tool can measure attention in the same way as another validated attention test, they will choose to do convergent validity testing. When they want to show that a tool can discriminate between children with cognitive insults from the ones that are healthy, they will choose to test for discriminant validity. However distinct the types of validity are, a tool cannot be assumed to work well unless it shows evidence of reliability, correlation with variables that it is expected to correlate with and lack of correlation with variables that it is not expected to correlate as well as evidence that the tool items reflect the cultural construct (Chiang et al. 2015 (October, 13)). In most of the studies reported in this review, reliability and validity were assumed to be different entities; hence, a study could test for validity without testing for reliability. Moreover, most tests had one study reporting on their psychometric properties which should not be the practice with cognitive tools because they are sensitive to cultural experiences in development. Among the tests reviewed, The Developmental Test of Visuo-Motor Integration was the only test that reported on reliability as well as discriminant, convergent and construct validity and in addition had more than one study reporting its psychometric properties. This should be the practice among researchers before assuming that a tool works well. Educators and clinicians should check on these properties before integrating the tools into practice. Interpretation, use and relevance across different cultural settings should be the norm.

Cultural Adaptations

Adaptation processes took different dimensions each dependent on the objectives of the studies. Recommendations for cognitive tests adaptation consist of translation, piloting and test modification (Malda et al. 2008). The adaptation processes captured in this review involved changes to the tools in terms of language and items while the objectives of the study at times necessitated just the testing of different psychometric properties of full batteries or their subsets. The reviewed studies partially tapped into the recommended adaptation procedures. It is beyond the objectives of this review to make recommendations on appropriate adaptation of cognitive tests in different cultural contexts. However, some of the adapted tests resulted in cognitive tests with high validity and reliability indicators while others had low indicators. Tests such as the Behavioral Assessment and Research System (BARS) had test–retest validity ranging from low to high depending on the sub-test. The Brazilian Child Hayling Test had high content validity but low specificity; Behavior Assessment System for Children-Teacher Report was found to have high reliability, good construct validity but its predictive validity was found to be weak and partially significant; while the Kilifi toolkit was found to have moderate internal consistency, low

Table 5 Motor and perceptual motor standardized tests among 6–12-year-olds

Motor and perceptual motor tests	Frequency of studies	Validity	Reliability	Normative data
Developmental Test of Visuo-Motor Integration; Beery-Buktenica Developmental Test of Visual Motor Integration test; Beery Developmental Test of Visual-Motor Integration-Third Revision; Beery Visual-Motor Integration (VMI) Test	6	General validity low; predictive validity none significant; discriminant validity high, sensitivity and specificity ranged from low to high depending on cut-off score; 3rd edition has concurrent validity significant and content validity low; discriminant poor, convergent good, construct good, criterion good	Test–retest reliability high in one study and low in another, interrater reliability high	–
CogState battery	2	Construct validity good (3 factor structure), concurrent and convergent validity partially significant; general validity partially significant	Test–retest reliability moderate in one study and moderate to high in another	–
QS4-G: Parent Questionnaire for the Developmental Evaluation of 4-Year-Old	1	Sensitivity moderate to high, specificity high, predictive high apart from academic difficulties	–	–
Five to Fifteen parent questionnaire (FTF)	1	Criterion and discriminant partial significance	Internal consistency high	–
Purdue Pegboard	1	Predictive partially significant	–	–
Pegboard with the dominant (PegsDom) and nondominant (PegsND) hands	1	–	Test–retest reliability moderate to high	–
Matching Figures from the WRAVMA (Wide Range Assessment of Visual Motor Abilities)	1	–	Test–retest reliability moderate to high	–
Visual Learning from the WRAML (Wide Range Assessment of Memory and Learning)	1	–	Test–retest reliability moderate to high	–
Finger Windows from the WRAML (Wide Range Assessment of Memory and Learning)	1	–	Test–retest reliability moderate to high	–
Rey-Osterreith Complex Figure Task (RCFT)	2	Concurrent validity significant and content validity low	–	–
Luria-Nebraska Test for Children (TLN-C, in Portuguese)	1	General validity high	Internal consistency high	–
BaterÃa de EvaluaciÃ³n NeuropsicolÃ³gica Infantil (BENCI)	1	Discriminant validity high	Test–retest reliability moderate to high	–
10 Wechsler Intelligence Scale for Children-Third Edition (WISC-III) subtests and 4 Wechsler Individual Achievement Test (WIAT) subtests	1	External validity partially significant	Reliability good	–
IT—Inspection time (speed of visualization measure)	1	General validity significant	Reliability moderate	–
Pediatric Stroke Outcome Measure (PSOM)	1	Construct validity fair to moderate and partially significant	Interrater reliability high	–
Brief neurocognitive screener (DIVERGT)—subtests Digit Span Test, The Verbal Fluency Test, The Grooved Pegboard Test and The Trail Making Test	1	Sensitivity and specificity moderate to high; predictive validity significant, discriminant validity significant	Test–retest reliability good	–
Autism/Tics, AD/HD, and other Comorbidities (A&TAC) inventory	1	–	Intrarater and interrater moderate to high	–
Reality Monitoring (RM)	1	General validity partially significant	Interrater reliability significant	–
Korean Computerized Neurobehavioral Tests (KCNT)—subtests include Simple Reaction Time (response speed), Choice Reaction Time (psychomotor speed), Colour Word Vigilance (attention), Addition (executive functions), Symbol Digit (executive functions) and Finger Tapping Speed (manual dexterity)	1	–	Test–retest moderate	–
Halstead-Reitan Neuropsychological Test Battery for Older Children (HRNB-C)	1	Construct small to large correlations	Reliability high	–
Halstead-Reitan Neuropsychological Test Battery for Children (HRNB-C)	1	Construct validity significant	–	–
The Bruininks-Oseretsky Test of Motor Proficiency, Second Edition (BOT-2)	1	–	Interrater reliability high, test–retest reliability fair to good	–
Bruininks-Oseretsky Test of Motor Proficiency-SF (BOTMP-SF)	1	Concurrent validity significant, construct validity partially significant	–	–
The Movement Assessment Battery for Children (M-ABC)	2	Concurrent validity partially significant; Construct validity partially significant, concurrent validity significant	–	–
	3			–

Table 5 (continued)

Motor and perceptual motor tests	Frequency of studies	Validity	Reliability	Normative data
Kaufman Assessment Battery for Children, second edition (KABC-II)		Construct validity high, predictive validity low to high; construct validity good; construct good: yielded five factors (sequential processing, simultaneous processing, planning and learning)	Test–retest reliability moderate to high and in another study reliability was good	
Trail Making Test A (Trails A)	1	General validity significant, functional equivalence partial	–	–
Rorschach Performance Assessment System	1	General validity significantly partial	Interrater fair to high	–
Ten Questions' Questionnaire (TQQ)	1	Sensitivity high, specificity high	Test–retest fair to excellent, interrater good to excellent	–
Combination of Kaufman Hand Movements Scale; The Stroop Colour-Word Association Test (Stroop); The Controlled Oral Word Association Test (COWAT); Trail Making Test; Arithmetic and Digit Span subtests of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III); Conners' Continuous Performance Test (CPT)	1	Discriminant validity not significant, sensitivity and specificity low to high	–	–
Touwen examination	1	–	Test–retest poor to high, inter-assessor moderate to high, intra-assessor moderate to high	–
Neuropsychological Battery: subtests Mental Control; Target Detection Cancellation Test; Visual-Verbal Learning Curve; Rey-Osterrieth Complex Figure Test; Language Comprehension and Working Memory test; Language Fluency test; Wisconsin Card Sorting Test-Abbreviated Version (WCST-A)	1	Construct validity good, discriminant validity poor, sensitivity and specificity poor to high	–	–
Lebby-Asbell Neurocognitive Screening Examination—Children and Adolescent versions (LANSE-C/A)	1	Discriminant validity not significant, convergent validity partial significance	Internal consistency low to high	–
Conjunction Visual Search—CVS	1	External validity significant, internal validity significant	Reliability high	–
Behavioural Assessment and Research System (BARS) (included tests of motor speed and dexterity, attention, memory, and visuospatial coordination)	1	General validity partial	Test–retest low (for tests with alternate forms) to high (for tests without alternate forms)	–
Developmental Neuropsychological Assessment (NEPSY)	2	Discriminant validity significant; construct validity partial correlations, specificity low to high	Reliability moderate to high	–
Assessment of Motor and Process Skills (AMPS)	1	General validity high	–	–
Dean-Woodcock Sensory-Motor Battery (DWSMB)	1	Discriminant validity good	–	–
Test of Visual Perceptual Skills – Third Edition (TVPS) (Visual Discrimination, Visual Memory, Visual Spatial Relationships).	1	–	Test–retest reliability low	–
	49			

to moderate test retest reliability and partially significant predictive validity (Kitsao-Wekulo et al. 2013b; Rohitrattana et al. 2014; Sadeh et al. 2012; Siqueira et al. 2016). The variability in psychometric indicators could be as a result of many factors including differences in test population, differences in individual task scores that may affect reliability or also the adapted test items do not reflect the cultural construct (Cooper et al. 2017).

Implications for Domains Well Covered

A total of seventy-seven and seventy-five of the studies tested the psychometric properties of tools that measure executive function and complex attention respectively.

Executive function domain has been extensively covered among preschoolers and children in early school years despite development of this domain starting at around 3–5 years and its maturity being in adolescence (Best and Miller 2010). This

Table 6 Learning standardized tests among 6–12-year-olds

Learning tests	Frequency of studies	Validity	Reliability	Normative data
CogState battery	2	Construct validity good (3 factor structure), concurrent and convergent validity partially significant; general validity partially significant	Test–retest reliability (moderate); test–retest reliability (moderate to high)	–
Test of Memory and Learning (TOMAL)	1	Criterion good, discriminant significant, convergent partial significance, factorial analysis produced 5 factors; sensitivity and specificity low to high	–	–
Five to Fifteen parent questionnaire (FTF)	1	Criterion and discriminant partial significance	Internal consistency high	–
Memory Screening Index (MSI) from the WRAML (Wide Range Assessment of Memory and Learning)	1	Factor structure good, criterion significant	–	–
Rey's Auditory-Verbal Learning Test (AVLT)	1	–	Test–retest reliability low to high	–
Korean Educational Development Institute-Wechsler Intelligence Scales (KEDI-WISC) (subtests include Continuous Performance Test (CPT), Children's Colour Trails Test (CCTT) and Stroop Colour-Word Test (SCWT))	1	General validity partial significance	–	–
Wechsler Intelligence Scale for Children Fourth Edition (WISC IV)- General Ability Index (GAI), Full Scale IQ (FSIQ) and Cognitive Proficiency Index (CPI)	2	Sensitivity high; general validity partial-VcSiMrBd subtests highest accuracy estimates for GAI	–	–
Differential Ability Scales (DAS), Differential Ability Scales - Second Edition (DAS II)	2	Discriminant validity good; predictive validity for DASII high	–	–
A brief computerized test, incorporated into the Discrete Trial Trainer (c)	1	Concurrent validity partial, sensitivity high	Test–retest reliability high	–
Internet based measures:- Peabody Individual Achievement Test (PIAT); GOAL Formative Assessment in Literacy for Key Stage 3; Woodcock-Johnson III Reading Fluency Test; Language tests Listening Gramma, Figurative Language and Making Inferences; items from National Foundation for Educational Research 5–14 Mathematics Series; General cognitive ability was measured using WISCIII-PI Multiple Choice Information (General Knowledge) and Vocabulary Multiple Choice subtests for verbal measures and for nonverbal measures WISC-III-UK Picture Completion and Raven's Standard Progressive Matrices. The Spatial Reasoning series	1	Concurrent validity good	Internal consistency reliability high	–
Children's Category Test – Level 2 (CCT-2)	1	Criterion partial, discriminant partially significant, sensitivity partial	–	–
Autism/Tics, AD/HD, and other Comorbidities (A&TAC) inventory	1	–	Intrarater and interrater moderate to high	–
Immediate Post concussion Assessment and Cognitive Testing (ImPACT)	1	–	–	Norms established for children aged 10–12 years
Lebby-Asbell Neurocognitive Screening Examination—Children and Adolescent versions (LANSE-C/A)	1	Discriminant validity not significant, convergent validity partial significance	Internal consistency low to high	–
Developmental Neuropsychological Assessment (NEPSY)	2	Discriminant validity significant; construct validity partial correlations, specificity low to high	Reliability moderate to high	–
Go/No-Go paradigm	1	Discriminant validity significant, convergent validity partially significant	–	–
A Maze task	1	Discriminant validity significant, Convergent validity partially significant	–	–
	21			

Table 7 Language standardized tests among 6–12-year-olds

Language tests	Frequency of studies	Psychometric output	Reliability	Normative data
QS4-G: Parent Questionnaire for the Developmental Evaluation of 4-Year-Old	1	Sensitivity moderate to high, specificity high, predictive high apart from academic difficulties	–	–
Five to Fifteen parent questionnaire (FTF)	1	Criterion and discriminant partial significance	Internal consistency high	–
Expressive One-Word Picture Vocabulary Test - Revised	1	–	Test–retest reliability moderate	–
Luria-Nebraska Test for Children (TLN-C, in Portuguese)	1	General validity high	Internal consistency high	–
FAS Verbal Fluency Test	1	General validity partial significance	–	–
BaterÃa de EvaluaciÃ³n NeuropsicolÃ³gica Infantil (BENCI)	1	Discriminant validity high	Test–retest reliability moderate to high	–
10 Wechsler Intelligence Scale for Children-Third Edition (WISC-III) subtests and 4 Wechsler Individual Achievement Test (WIAT) subtests	1	External validity partially significant	Reliability good	–
Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV) Vocabulary subtest	1	General validity low	–	–
Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV)	1	–	Test–retest reliability low to excellent	–
Internet based measures: Peabody Individual Achievement Test (PIAT); GOAL Formative Assessment in Literacy for Key Stage 3; Woodcock-Johnson III Reading Fluency Test; Language tests Listening Gramma, Figurative Language and Making Inferences; Items from National Foundation for Educational Research 5–14 Mathematics Series; General cognitive ability was measured using WISCIII-PI Multiple Choice Information (General Knowledge) and Vocabulary Multiple Choice subtests for verbal measures and for nonverbal measures WISC-III-UK Picture Completion and Raven's Standard Progressive Matrices. The Spatial Reasoning series	1	Concurrent validity good	Internal consistency reliability high	–
Pediatric Stroke Outcome Measure (PSOM)	1	Construct validity fair to moderate and partially significant	Interrater reliability high	–
Autism/Tics, AD/HD, and other Comorbidities (A&TAC) inventory	1	–	Intrarater and interrater moderate to high	–
Seashore Rhythm Test (SRT)	1	–	Reliability moderate	–
Two forms of the Speech Sounds Perception Test (SSPT)	1	–	Reliability moderate	–
Aphasia Screening Test (AST)	1	–	Reliability moderate	
Halstead-Reitan Neuropsychological Test Battery for Older Children (HRNB-C)	1	Construct small to large correlations	Reliability high	–
Halstead-Reitan Neuropsychological Test Battery for Children (HRNB-C)	1	Construct validity significant	–	–
Evaluación Neuropsicológica Infantil (ENI)	1	General validity significant	–	–
Buschke Selective Reminding Test (SRT)	1	Convergent validity partial	Test–retest poor to good	–

Table 7 (continued)

Language tests	Frequency of studies	Psychometric output	Reliability	Normative data
Woodcock Reading Mastery Test	1	Convergent validity significant, concurrent validity high similarity	–	–
Omnibus test of cognitive functioning; Trail Making A (attention), Continuous Performance Task (CPT) (attention)]; Trail Making B (Executive Function); Cog Set Shifting (Executive Function), Controlled Oral Word Association Test (COWAT) (Executive Function); Digit Span (Working Memory), Spatial Span (Working Memory), and California Verbal Learning Test (CVLT)(Verbal Memory)	1	General validity significantly strong	Interrater reliability high	–
Timo's Adventure	1	Discriminant validity high, sensitivity high, specificity high	–	–
Combination of Kaufman Hand Movements Scale; The Stroop Colour-Word Association Test (Stroop); The Controlled Oral Word Association Test (COWAT); Trail Making Test; Arithmetic and Digit Span subtests of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III; Conners' Continuous Performance Test, (CPT)	1	Discriminant validity not significant, sensitivity and specificity low to high	–	–
Neuropsychological Battery: subtests Mental Control; Target Detection Cancellation Test; Visual-Verbal Learning Curve; Rey-Osterrieth Complex Figure Test; Language Comprehension and Working Memory test; Language Fluency test; Wisconsin Card Sorting Test-Abbreviated Version (WCST-A)	1	Construct validity good, discriminant validity poor, sensitivity and specificity poor to high	–	–
Revised Token Test (RTT)	1	Discriminant validity high	–	–
Lebby-Asbell Neurocognitive Screening Examination—Children and Adolescent versions (LANSE-C/A)	1	Discriminant validity not significant, convergent validity partial significance	Internal consistency low to high	–
Developmental Neuropsychological Assessment (NEPSY)	2	Discriminant validity significant; reliability moderate to high, construct validity partial correlations, specificity low to high	–	–
	28			

trend presupposes that these studies are inclined to find out the developmental trend rather than whether or not the function has reached maturity. In addition to this, the tests such as BaterÃa de EvaluaciÃn NeuropsicolÃgica Infantil (BENCI), Developmental Neuropsychological Assessment (NEPSY) and subtests of Behavioral Assessment of Dysexecutive Syndrome for Children (BADS-C) monitor executive dysfunction progression and recovery. This interest in executive function development means many tools are likely to be developed and standardized for measurement of these domains compared to other domains. In addition, the interest leads to development of different versions of the same tools in different settings.

Key Gaps and Areas for Intervention

Most of the tools have been standardized to be used in the USA, yet each setting has different cultural practices that give a different orientation to cognitive functioning. One's environment determines cognitive development trajectory. In the USA, processing speed of information is valued in education, which may underpin quality of information which is inadvertently valued among Hispanics (Casaletto and Heaton 2017).

Only seven studies reported on the development of normative data with other studies reporting on the decision to changing the tools to make them valid. Though there is still a debate on which option to pick before integration of a tool in a certain setting, it is interesting to note that the researchers are hesitant to develop normative data. It is important for test results to be interpreted with regard to the general population as clinical data may not cover the full range of possible scores. Normative data is able to tell whether a child's functioning score is well within that of the general population in reference to age or not (Ellingsen 2016). Normative data studies are difficult to conduct as several methods of data collection need to be integrated to obtain an ethnically diverse sample that is truly representative of the general population (Nolte et al. 2015).

Table 8 Other cognitive domains standardized tests among 6–12-year-olds

Other tests	Frequency of studies	Validity	Reliability	Normative data	Cognitive domain
Zareki-R. Arithmetic subtest of WISC-III	1	Construct validity partial	–	Normative data	Arithmetic
KeyMath-Revised Inventory (KM-R)	1	Construct validity good, discriminant validity inadequate	–	–	Arithmetic
Wechsler Individual Achievement Test-Second Edition (WIAT-II) reading subtest (measured Cognitive reserve)	1	General validity low	–	–	Cognitive reserve
Reynolds Intellectual Assessment Scale (RIAS)—subtests include: Composite Intelligence Index (CIX), Nonverbal Intelligence Index (NIX) and Verbal Intelligence Index (VIX)	1	Construct validity partial	–	–	Intelligence
Five to Fifteen parent questionnaire (FTF)	1	Criterion and discriminant partial significance	Internal consistency high	–	Social skills
EF scale from the Behaviour Assessment System for Children-Teacher Report	1	Construct validity good, predictive validity weak and partially significant	Reliability high	–	Social cognition
Autism/Tics, AD/HD, and other Comorbidities (A&TAC) inventory	1	–	Intrarater and interrater moderate to high	–	Social cognition
Human figure drawings (Matching Familiar Figure Test)—two drawings were used: person and house, tree and person	1	Discriminant validity significantly partial	Interrater reliability high	–	Social cognition
Cartoons, pictures and video vignettes	1	Discriminant partial significant	–	–	Social cognition
Cambridge Neuropsychological Test Battery (CANTAB)	3	General validity partially significant; general validity partial significant, construct good	Test–retest reliability low	–	Representational competence
WISC-RN (the Dutch version of the WISC-R)	1	Construct validity poor, diagnostic validity no significant difference	Reliability high	–	Intellectual ability
Woodcock Johnson III Tests of Achievement	1	–	Test–retest reliability low to high	–	Academic cchievement
	14				

Conclusion

The narrative review indicates that more needs to be done in cultural adaptation and standardization of neuropsychological tools. There is a need to extensively standardize other DSM-5 cognitive domains; adapt and standardize tools in diverse settings; and integrate diverse validity and reliability measures, as well as courageously do normative data studies.

Strengths and Limitations

A narrative review was conceptualized to be the best form of interrogating the research questions due to its nature of critically looking at and discussing the knowledge of interest. That said, there are other forms of studies that would have complemented the findings of this review such as systematic reviews. The review concentrated on studies done between 1987 and 2017 hence studies falling off this timeline were not integrated. Further, the search sites were limited to PubMed, Web of Science and Psych Infor yet there are other databases that would have generated more information. However, even though we concentrated on these three search sites, we had duplication of data a situation that would have inadvertently been described as having reached saturation level. The terms used for the search were limited to "neuropsychological" or "neurocognitive" and "assessment" or "test". Using other terms with these ones could have increased the comprehensiveness and impact of the work. However, during screening other search terms were tried out but they resulted in the same and, in some cases, fewer results meaning this study search terms resulted in optimal and unique studies. Moreover, this search criteria resulted in many studies being screened in comparison to other cognitive review studies that

have found fewer search results. In addition, narrative reviews are said to be subjective and their search criteria may not have explicit specifications (Ferrari 2015). The search did not include studies with non-English-reported findings due to lack of resources for hiring translators. However, reviews that have not included non-English publications have been found to not have systematic bias (Morrison et al. 2012; Nussbaumer-Streit et al. 2020). The search did not include data published in test manuals that was not published in research journals.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40817-021-00096-9.

Data Availability The data from which the results of this review were extracted can be accessed by contacting the corresponding author.

Compliance with Ethical Standards

Ethics Approval and Consent to Participate This study was part of a larger study whose ethical approval was sort from Tilburg University's Research Ethics Committee (REC#2017/25) and University of Nairobi, Kenyatta National Hospital Ethics and Research Committee (P556/07/2016).

Consent for Publication We give consent for publication of this paper.

Financial Disclosure This study was funded through a seed grant for early career researchers organized by Partnerships for Mental Health Development in Sub-Saharan Africa (PaM-D) (NIMH award number U19MH98718) and the Kenyatta National Hospital's Research & Programs Department.

References

Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A., & Lyytinen, H. (2000). Repeated assessment of the Tower of Hanoi test: reliability and age effects. *Assessment, 7*(3), 297–310.

Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A., & Lyytinen, H. (2001). Practice effects on visuomotor and problem-solving tests by children. *Perceptual and Motor Skills, 92*(2), 479–494.

Bangirana, P., Seggane-Musisi., Allebeck, P., Giordani, B., John, C., Opoka, O.,. .. MJ., B. (2009). A preliminary examination of the construct validity of the KABC-II in Ugandan children with a history of cerebral malaria. *African Health Sciences, 9*(3).

Bangirana, P., Sikorskii, A., Giordani, B., Nakasujja, N., & Boivin, M. J. (2015). Validation of the CogState battery for rapid neurocognitive assessment in Ugandan school age children. *Child and Adolescent Psychiatry and Mental Health, 9*.

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development, 81*(6). https://doi.org/10.1111/j.1467-8624.2010.01499.x.

Bishop, D. V. M., Aamodt-Leeper, G., Creswell, C., McGurk, R., & Skuse, D. H. (2001). Individual differences in cognitive planning on the Tower of Hanoi task: Neuropsychological maturity or measurement error? *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 42*(4), 551–556.

Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: a critical review of tests. *Journal of Psychoeducational Assessment, 19*(1), 19–44. https://doi.org/10.1177/073428290101900102.

Budtz-Jorgensen, E., Keiding, N., Grandjean, P., & Weihe, P. (2002). Estimation of health effects of prenatal methylmercury exposure using structural equation models. *Environmental Health, 1*(1), 2.

Carone, D. A. (2014). Young child with severe brain volume loss easily passes the word memory test and medical symptom validity test: Implications for mild TBI. *The Clinical Neuropsychologist, 28*(1), 146–162. https://doi.org/10.1080/13854046.2013.861019.

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological assessment: past and future. *Journal of the International Neuropsychological Society, 23*(9–10), 778–790. https://doi.org/10.1017/S1355617717001060.

Chen, X., Chen, H., Li, D., & Wang, L. (2009). Early childhood behavioral inhibition and social and school adjustment in Chinese children: a 5-year longitudinal study. *Child Development, 80*(6), 1692–1704. https://doi.org/10.1111/j.1467-8624.2009.01362.x.

Chiang, I. A., Jhangiani, R. S., & Price, P. C. (2015 (October, 13)). Research methods in psychology: reliability and validity of measurement (2 ed.). Canada: BC Campus.

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: a case study of the AX-CPT. *Frontiers in Psychology, 8*, 1482–1482. https://doi.org/10.3389/fpsyg.2017.01482.

Ellingsen, K. M. (2016). Standardized assessment of cognitive development: Instruments and issues. In *Early Childhood Assessment in School and Clinical Child Psychology* (pp. 25-49).

Ezeamama, A. E., Bustinduy, A. L., Nkwata, A. K., Martinez, L., Pabalan, N., Boivin, M. J., & King, C. H. (2018). Cognitive deficits and educational loss in children with schistosome infection—a systematic review and meta-analysis. *PLoS Neglected Tropical Diseases, 12*(1), e0005524. https://doi.org/10.1371/journal.pntd.0005524.

Fasfous, Peralta-Ramirez, M. I., Perez-Marfil, M. N., Cruz-Quintana, F., Catena-Martinez, A., & Perez-Garcia, M. (2015a). Reliability and validity of the Arabic version of the computerized Battery for Neuropsychological Evaluation of Children (BENCI). *Child Neuropsychology, 21*(2), 210–224. https://doi.org/10.1080/09297049.2014.896330.

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing, 24*(4), 230–235. https://doi.org/10.1179/2047480615z.000000000329.

Holding, P. A., Taylor, H. G., Kazungu, S. D., Mkala, T., Gona, J., Mwamuye, B., et al. (2004). Assessing cognitive outcomes in a rural African population: development of a neuropsychological battery in Kilifi District, Kenya. *Journal of the International Neuropsychological Society : JINS, 10*(2), 246–260. https://doi.org/10.1017/S1355617704102166.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: where we have been and where we are going. *The Journal of General Psychology, 123*(3), 207–215. https://doi.org/10.1080/00221309.1996.9921273.

Hwang, Y., Hosokawa, T., Swanson, H. L., Ishizaka, I., Kifune, N., Ohira, D., & Ota, T. (2006). A Japanese short form of the

Swanson cognitive processing test to measure working memory: reliability, validity, and differences in scores between primary school children of the United States and Japan. *Psychological Reports, 99*(1), 27–38. https://doi.org/10.2466/pr0.99.1.27-38.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Oxford: World Book Co.

Kitsao-Wekulo, P. K., Holding, P. A., Taylor, H. G., Abubakar, A., & Connolly, K. (2013a). Neuropsychological testing in a rural African school-age population: evaluating contributions to variability in test performance. *Assessment, 20*(6), 776–784. https://doi.org/10.1177/1073191112457408.

Kitsao-Wekulo, P. K., Holding, P. A., Taylor, H. G., Kvalsvig, J. D., & Connolly, K. J. (2013b). Determinants of variability in motor performance in middle childhood: a cross-sectional study of balance and motor co-ordination skills. *BMC Psychology, 1*(1), 29–29. https://doi.org/10.1186/2050-7283-1-29.

Konstantopoulos, K., Vogazianos, P., Thodi, C., & Nikopoulou-Smyrni, P. (2015). A normative study of the Children's Color Trails Test (CCTT) in the Cypriot population. *Child Neuropsychology, 21*(6), 751–758.

Llorente, A. M., Voigt, R. G., Williams, J., Frailey, J. K., Satz, P., & D'Elia, L. F. (2009). Children's color trails test 1 2: -retest reliability and factorial validity. *The Clinical Neuropsychologist, 23*(4), 645–660.

Malda, M., Vijver, F. J. R. V. D., Transler, C., Sukumar, P., Srinivasan, K., & Rao, K. (2008). Adapting a cognitive test for a different culture: an illustration of qualitative procedures. *Psychology Science Quarterly, 50*(4), 451–468.

Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., et al. (2012). The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *International Journal of Technology Assessment in Health Care, 28*, 138–144. https://doi.org/10.1017/S0266462312000086.

Mung'ala-Odera, V., Meehan, R., Njuguna, P., Mturi, N., Alcock, K., Carter, J. A., & Newton, C. R. (2004). Validity and reliability of the 'Ten Questions' questionnaire for detecting moderate to severe neurological impairment in children aged 6-9 years in rural Kenya. *Neuroepidemiology, 23*(1–2), 67–72. https://doi.org/10.1159/000073977.

Nolte, M. T., Shauver, M. J., & Chung, K. C. (2015). Analysis of four recruitment methods for obtaining normative data through a Web-based questionnaire: a pilot study. *Hand (N Y), 10*(3), 529–534. https://doi.org/10.1007/s11552-014-9730-y.

Nussbaumer-Streit, B., Klerings, I., Dobrescu, A. I., Persad, E., Stevens, A., Garritty, C., et al. (2020). Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. *Journal of Clinical Epidemiology, 118*, 42–54. https://doi.org/10.1016/j.jclinepi.2019.10.011.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45–50.

Reitan, R. M., & Wolfson, D. (2004). The Trail Making Test as an initial screening procedure for neuropsychological impairment in older children. *Archives of Clinical Neuropsychology, 19*(2), 281–288.

Reynolds, E., Fazio, V. C., Sandel, N., Schatz, P., & Henry, L. C. (2016). Cognitive development and the immediate postconcussion assessment and cognitive testing: a case for separate norms in preadolescents. *Applied Neuropsychology: Child, 5*(4), 283–293. https://doi.org/10.1080/21622965.2015.1057637.

Rohitrattana, J., Siriwong, W., Suittiwan, P., Robson, M. G., Strickland, P. O., Rohlman, D. S., & Fiedler, N. (2014). Adaptation of a neurobehavioral test battery for Thai children. *Roczniki Państwowego Zakładu Higieny, 65*(3), 205–212.

Rose, S. A., Feldman, J. F., Jankowski, J. J., & Van Rossem, R. (2011). Basic information processing abilities at 11 years account for deficits in IQ associated with preterm birth. *Intelligence, 39*(4), 198–209.

Roy, A., Allain, P., Roulin, J.-L., Fournet, N., & Le Gall, D. (2015). Ecological approach of executive functions using the Behavioural Assessment of the Dysexecutive Syndrome for Children (BADS-C): developmental and validity study. *Journal of Clinical and Experimental Neuropsychology, 37*(9), 956–971. https://doi.org/10.1080/13803395.2015.1072138.

Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., & Petersen, R. C. (2014). Classifying neurocognitive disorders: the DSM-5 approach. *Nature Reviews. Neurology, 10*(11), 634–642. https://doi.org/10.1038/nrneurol.2014.181.

Sadeh, S. S., Burns, M. K., & Sullivan, A. L. (2012). Examining an executive function rating scale as a predictor of achievement in children at risk for behavior problems. *School Psychology Quarterly, 27*(4), 236–246. https://doi.org/10.1037/spq0000012.

Siqueira, L. S., Goncalves, H. A., Hubner, L. C., & Fonseca, R. P. (2016). Development of the Brazilian version of the Child Hayling Test. *Trends Psychiatry Psychother, 38*(3), 164–174. https://doi.org/10.1590/2237-6089-2016-0019.

Spironello, C., Hay, J., Missiuna, C., Faught, B. E., & Cairney, J. (2010). Concurrent and construct validation of the short form of the Bruininks-Oseretsky test of motor proficiency and the movement-ABC when administered under field conditions: Implications for screening. *Child: Care, Health and Development, 36*(4), 499–507. https://doi.org/10.1111/j.1365-2214.2009.01066.x.

Stad, F. E., Wiedl, K. H., Vogelaar, B., Bakker, M., & Resing, W. C. M. (2019). The role of cognitive flexibility in young children's potential for learning under dynamic testing conditions. *European Journal of Psychology of Education, 34*(1), 123–146. https://doi.org/10.1007/s10212-018-0379-8.

Stadskleiv, K. (2020). Cognitive functioning in children with cerebral palsy. *Developmental Medicine and Child Neurology, 62*(3), 283–289. https://doi.org/10.1111/dmcn.14463.

Stinnett, T. A., Oehler-Stinnett, J., Fuqua, D. R., & Palmer, L. S. (2002). Examination of the underlying structure of the NEPSY: a developmental neuropsychological assessment. *Journal of Psychoeducational Assessment, 20*(1), 66–82.

Syvaoja, H. J., Tammelin, T. H., Ahonen, T., Rasanen, P., Tolvanen, A., Kankaanpaa, A., & Kantomaa, M. T. (2015). Internal consistency and stability of the CANTAB neuropsychological test battery in children. *Psychol Assess, 27*(2), 698–709. https://doi.org/10.1037/a0038485.

Teglasi, H., Nebbergall, A. J., & Newman, D. (2012). Construct validity and case validity in assessment. *Psychological Assessment, 24*(2), 464–475. https://doi.org/10.1037/a0026012.

Termine, C., Luoni, C., Fontolan, S., Selvini, C., Perego, L., Pavone, F., et al. (2016). Impact of co-morbid attention-deficit and hyperactivity disorder on cognitive function in male children with Tourette syndrome: a controlled study. *Psychiatry Research, 243*, 263–267.

Thaler, N. S., Allen, D. N., McMurray, J. C., & Mayfield, J. (2010). Sensitivity of the test of memory and learning to attention and memory deficits in children with ADHD. *Clinical Neuropsychology, 24*(2), 246–264. https://doi.org/10.1080/13854040903277305.

Thomas, E., Maruff, P., Paul, J., & Reeve, R. (2016). Spatial sequence memory and spatial error monitoring in the Groton Maze Learning Task (GMLT): a validation study of GMLT sub-measures in healthy children. *Child Neuropsychology, 22*(7), 837–852. https://doi.org/10.1080/09297049.2015.1038989.

van Nieuwenhuijzen, M., Vriens, A., Scheepmaker, M., Smit, M., & Porton, E. (2011). The development of a diagnostic instrument to measure social information processing in children with mild to borderline intellectual disabilities. *Research in Developmental Disabilities, 32*(1), 358–370. https://doi.org/10.1016/j.ridd.2010.10.012.

Vriezen, E. R., & Pigott, S. E. (2002). The relationship between parental report on the BRIEF and performance-based measures of executive function in children with moderate to severe traumatic brain injury. *Child Neuropsychology, 8*(4), 296–303. https://doi.org/10.1076/chin.8.4.296.13505.

Williams, M. E., Sando, L., & Soles, T. G. (2014). Cognitive tests in early childhood: psychometric and cultural considerations. *Journal of Psychoeducational Assessment, 32*(5), 455–476. https://doi.org/10.1177/0734282913517526.

Woodward, H., & Donders, J. (1998). The performance of children with traumatic head injury on the wide range assessment of memory and learning-screening. *Applied Neuropsychology, 5*(3), 113–119. https://doi.org/10.1207/s15324826an0503_1.

Affiliations

Maina Rachel[1,2] · Van De Vijver J. R. Fons[2,3] · Abubakar Amina[4,5] · Miguel Perez-Garcia[6] · Kumar Manasi[7]

1 Department of Clinical Medicine and Therapuetics, University of Nairobi, Nairobi 10834-00400, Kenya

2 Department of Culture Studies, Tilburg University, Tilburg, Netherlands

3 Department of Psychology, Higher School of Economics, Moscow, Russia

4 Neurosciences Unit, KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya

5 Institute for Human Development, Aga Khan University, Nairobi, Kenya

6 Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain

7 Department of Psychiatry, University of Nairobi, Nairobi, Kenya