REVIEW

A Systematic Review of the Validity and Reliability of Assessment Tools for Executive Function and Adaptive Function Following Brain Pathology among Children and Adolescents in Low- and Middle-Income Countries

Kwabena Kusi-Mensah[1,2] · Nana Dansoah Nuamah[3] · Stephen Wemakor[4] · Joel Agorinya[5] ·
Ramata Seidu[5] · Charles Martyn-Dickens[2] · Andrew Bateman[1,6]

Abstract

Minimal but increasing number of assessment instruments for Executive functions (EFs) and adaptive functioning (AF) have either been developed for or adapted and validated for use among children in low and middle income countries (LAMICs). However, the suitability of these tools for this context is unclear. A systematic review of such instruments was thus undertaken. The Systematic review was conducted following the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) checklist (Liberati et al., in *BMJ (Clinical Research Ed.), 339,* 2009). A search was made for primary research papers reporting psychometric properties for development or adaptation of either EF or AF tools among children in LAMICs, with no date or language restrictions. 14 bibliographic databases were searched, including grey literature. Risk of bias assessment was done following the COSMIN (**CO**nsensus-based **S**tandards for the selection of health status **M**easurement **IN**struments) guidelines (Mokkink et al., in *Quality of Life Research, 63*, 32, 2014). For EF, the Behaviour Rating Inventory of Executive Functioning (BRIEF- multiple versions), Wisconsin Card Sorting Test (WCST), Go/No-go and the Rey-Osterrieth complex figure (ROCF) were the most rigorously validated. For AFs, the Vineland Adaptive Behaviour Scales (VABS- multiple versions) and the Child Function Impairment Rating Scale (CFIRS- first edition) were most validated. Most of these tools showed adequate internal consistency and structural validity. However, none of these tools showed acceptable quality of evidence for sufficient psychometric properties across all the measured domains, particularly so for content validity and cross-cultural validity in LAMICs. There is a great need for adequate adaptation of the most popular EF and AF instruments, or alternatively the development of purpose-made instruments for assessing children in LAMICs.

Systematic Review Registration numbers: CRD42020202190 (EF tools systematic review) and CRD42020203968 (AF tools systematic review) registered on PROSPERO website (https://www.crd.york.ac.uk/prospero/).

✉ Kwabena Kusi-Mensah
kpk27@cam.ac.uk

1 Department of Psychiatry, University of Cambridge, Clifford Allbutt Building, Cambridge Biomedical Campus CB2 OAH, Cambridge, UK

2 Komfo Anokye Teaching Hospital, P. O. Box 1934, Kumasi, Ghana

3 Pantang Hospital, Accra, Ghana

4 Department of Psychiatry, University of Michigan Health System, 1500 E Medical Center Dr, Ann Arbor, MI 48109, USA

5 Accra Psychiatric Hospital, Accra, Ghana

6 School of Health and Social Care, University of Essex, Colchester, UK

Introduction

Rationale

Only 10% of research into child and adolescent mental health (CAMH) problems is carried out in low-and-middle-income countries (LAMICs) (Kessler et al., 2007; Merikangas et al., 2009). Meanwhile, over 90% of the world's children live in LAMICs (WHO, 2008) and this proportion is only set to grow with the high birth rate in these countries. Historically the focus of public health in LAMICs has been

on communicable diseases like HIV and Malaria. However, as the standard of living of LAMICs continues to improve, the data shows that in recent years the gradual decrease in infant mortality has resulted in an increasing shift toward non-communicable diseases such as neurodevelopmental disorders (NDDs) (Bakare et al., 2014) and CAMH disorders, which impact cognitive function. Children who would otherwise have died from various infections and birth injuries are now surviving, but surviving with the sequalae of the varied insults suffered at birth and the perinatal period, which are common in LAMICs because of poor obstetric care (Omigbodun & Bella, 2004). The common pathway of many of these conditions are neurobehavioral difficulties often described as Acquired Brain Injury (Bennett et al., 2005; Stuss, 1983) which affects brain function and the mental health and well-being of these children. It is within this context that accurate assessment of executive dysfunction and adaptive functioning, as known sequalae for brain injury, becomes quite important for children in LAMICs.

Executive functions (EF) may be defined as "top-down control processes" of human behaviour (Diamond, 2013) whose primary function is "supervisory control" (Stuss & Alexander, 2000) and includes such abilities as initiation, planning, and decision-making (Diamond, 2013). Better EF is linked to many positive outcomes (Diamond, 2013) such as greater success in school (Duncan et al., 2007; St Clair-Thompson & Gathercole, 2006), while deficits in EF are associated with slow school progress (Morgan et al., 2017) difficulties in peer relationships (Tseng & Gau, 2013) and poor employment prospects (Bailey, 2007). This may be because EF's have also been presented as potential endophenotypes of various childhood mental disorders such as Hyperkinetic Disorder (ICD 10[1] code: F90.1; also commonly called Attention Deficit Hyperactivity Disorder (ADHD)) (Doyle et al., 2005). Behaviourally, EF deficits may manifest as distractibility, fidgetiness, poor concentration, chaotic organization of materials, and trouble completing work (Bathelt et al., 2018). Given the difficulties seen, it is therefore important that mental health and rehabilitation services are able to pinpoint areas of greatest difficulty and target interventions appropriately and cost effectively through accurate assessments (Simblett et al., 2012).

Adaptive functioning on the other hand is an area that is just beginning to be examined (Semrud-Clikeman et al., 2017). It is defined as behaviours necessary for age-appropriate, independent functioning in social, communication, daily living or motor areas (Matson et al., 2009), tapping into the ability to carry out everyday tasks within age and context appropriate constraints (World Health Organization, 2001).

In this present study, the term was restricted to the narrow scope of adaptive function following brain injury/brain pathology.

Adaptive functioning may be viewed as the practical expression of executive functions in an everyday functional context. Executive function abilities are related and have predictive power over adaptive behaviour (Clark et al., 2002) in both typical and atypical populations according several studies (Gardiner & Iarocci, 2018; Gilotty et al., 2010; Gligorović & Đurović, 2014; Low Kapalu et al., 2018; Perna et al., 2012; Pugliese et al., 2015; Sabat et al., 2020; Schonfeld et al., 2007; Ware et al., 2012; Zorza et al., 2016). Specific domains of adaptive behaviours and academic achievement may, in part, depend on executive function capacities (Clark et al., 2002). Specifically, the core domains of EF- working memory, inhibition, and cognitive flexibility- have been shown to relate to the domains of adaptive behaviour as conceptual skills (e.g., language and the understanding of time, money, and number concepts) (Gilotty et al., 2010; Pugliese et al., 2015; Sabat et al., 2020) and practical skills (e.g., personal care, occupational and safety capabilities, use of money and transportation, and following of schedules and routines) (Perna et al., 2012; Sabat et al., 2020) and social skills (Gilotty et al., 2010; Pugliese et al., 2015; Zorza et al., 2016). Therefore, reviewing tools for the two related constructs seemed like an appropriate approach to take. Therefore, assessed together, EF and AF could provide the most utility to LMIC clinicians, depending on whether the goal is to focus interventions from a specific domains' perspective or specific areas of functional deficit in a day-to-day context for interventions.

At this point, one may wonder why of all the cognitive functioning constructs we chose to review tools assessing executive functions and adaptive functioning, and not say IQ. We wanted to review constructs that found the widest applicability trans-diagnostically and the most utility in the clinical setting, and for which interventions could be most directly designed. The information that assessing other psychological constructs related to frontal lobe functioning such as IQ might give, may not be as actionable as what a comprehensive assessment of the domains of EF and AF would provide. For example, an IQ assessment may be useful in diagnosing intellectual disability and in school placement (which are both very important of course), whereas an assessment of EF and AF could lead to identifying domains and areas of functional deficit in the child's life (across several diagnostic labels) that would lend itself most immediately to therapeutic interventions. Furthermore, AF assessment would also allow clinicians to determine the level of support required (Association & Association, 2013). Finally, even in the developed world with the full range of neuropsychological services available, neuropsychological assessments such as EF and AF assessments alone represents

---

[1] ICD 11 is due to be released on Jan 1st, 2022 hence still using ICD 10 nomenclature.

a significant proportion of all assessment services by clinical psychologists (up to 21% according to a large national representative survey among US psychologists (Camara et al., 2000). It therefore seemed more useful for resource constrained LMICs to do a systematic review of such tools for EF and AF than for any other psychological construct at the present time.

Several tools have been developed to assess executive and adaptive function in Western or High-Income-Country (HIC) populations, which perpetuates the trend of skewing research towards wealthier countries as noted above (Kessler et al., 2007; Merikangas et al., 2009). However, not much is known about the nature and quality of tools developed for LAMIC populations. A recent scoping review of EF tools among adolescents was limited in scope, and did not focus particularly on LAMICs (Nyongesa et al., 2019). More importantly, this study did not evaluate risk of bias of the eligible papers (particularly noteworthy was the lack of focus on assessing risk of bias of content validity and cross cultural validity) but only reported on the results declared therein (Nyongesa et al., 2019). Another recent scoping review- this time focused on NDD's among children in LAMICs also reported regarding EF and AF tools used in LAMICs that, only a few tests have been used to assess executive function in children in LAMICs (Semrud-Clikeman et al., 2017). This paper also failed to do a critical appraisal, though that was outside its scope.

There is a high burden of the known causes of brain injury in developing countries (Bitta et al., 2017; Merikangas et al., 2009) and therefore a rigorous critical appraisal of appropriate assessment tools for EF and AF in this specific context will be highly desirable. Towards this goal of elucidating the issue of assessment of executive and adaptive functioning among children in the context of LAMICs, a scoping review of the subject was undertaken by the authors in an earlier paper to broadly map out the kinds of instruments that had either been newly developed or adapted for use in this context, the results of which are reported elsewhere (Kusi-Mensah et al., 2021). However, reviewing the quality of evidence found in the reviewed papers was beyond the scope of that scoping review. This present paper, therefore, is a continuation of that project, seeking to critically appraise the quality of evidence for the results of the scoping review, and to thus make more definitive recommendations on the best instruments with the most high-quality evidence for use among children in LAMICs.

Objectives

The present study seeks to undertake a systematic review of published literature on the reliability and validity of assessment tools for executive functioning and adaptive functioning among children in LAMIC contexts. The purpose of this is to critically appraise and summarise the evidence for the scientific rigour of the methodologies used (risk of bias), and the results presented (psychometric measurement properties established) for the tools which have been developed, adapted or validated among children in developing country contexts, as well as document any knowledge gaps that may exist.

The following research questions were therefore formulated:

1. What is the quality of adaptation (including content validation) of <u>existing</u> HIC-derived assessment tools for executive functioning or adaptive functioning among children in LAMICs?
2. What is the nature and quality of evidence undergirding newly developed and purpose-made tools for assessment of executive functioning or adaptive functioning among children in LAMICs?
3. What is the nature and quality of evidence supporting the psychometric properties (validity and reliability) of <u>all</u> HIC-derived assessment tools for executive functioning or adaptive functioning among children in LAMICs?

Methods

Protocol and Registration

This systematic review was conducted following the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) checklist (Liberati et al., 2009), and our protocol was written according to the PRISMA- Protocol extension (PRISMA-P) guideline (Shamseer et al., 2015). The protocols for the systematic reviews of EF and AF measurement tools were successfully registered separately on the PROSPERO website (see here: https://www.crd.york.ac.uk/prospero/) with registration numbers CRD42020202190 and CRD42020203968 for the EF tools systematic review and AF tools systematic review respectively.

Eligibility Criteria

As alluded to earlier, the papers selected for critical appraisal in this systematic review were selected from the scoping review conducted earlier by the authors. The details of the eligibility criteria therefore can be found in that paper (Kusi-Mensah et al., 2021). However, in summary, we made a search for primary research papers of all study designs that focused on development or adaptation/validation of EF and AF tools used in the context of the target outcomes (executive functioning or adaptive functioning following brain pathology) among children in LAMIC countries, with no date or language restrictions. This meant that

papers published from 1st January 1894 (earliest date of all search engines used) to 15th September 2020 (the last day of update of the search strategy) were included. The paper also had to primarily be concerned with developing, adapting or assessing the validity of the instrument of choice as one of its main stated study aims (if not the main), and not just as an incidental concern, to be eligible. For participants, studies examining children aged 5 years to 18 years (both healthy and clinical populations) living in LAMICs were included. All eligible full articles in any language were included in the search with no a priori language limits set on the search, and an attempt was made to translate non-English articles using Google Translate or volunteer native language speakers (as available). A list of potentially eligible articles that could not be obtained or translated have been provided in Appendix II. Target outcomes were defined as: 1. Papers reporting on the development (specifically concept elicitation and content validation) of a new tool assessing EF or AF; 2. Papers reporting on adaptation (content validation) of an existing HIC-derived EF or AF tool; and 3. Papers reporting psychometric properties of EF or AF tools in a LAMIC. Specifically, psychometric properties that were included as eligible for consideration were:

(a) Internal consistency
(b) Reliability (test–retest, inter-rater and intra-rater reliability)
(c) Validity (structural, cross-cultural, construct and criterion validity)
(d) Measurement Error
(e) Responsiveness

These measurement properties are fully defined in the 'data items' section below. Excluded were animal studies, studies that only used the instrument as an outcome measurement instrument (for instance in randomized controlled trials), studies in which validation was **not** of the EF or AF tool (but rather validation of another instrument for another non-EF/AF construct such as visuospatial ability, intelligence, short-term memory etc.), and all studies that did not meet the inclusion criteria.

Information Sources

The following databases were searched with indicated dates:

1. MEDLINE (OVID interface, 1946 onwards)
2. EMBASE (OVID interface, 1974 onwards)
3. Cochrane library (current issues)
4. PsychINFO (1894 onwards)
5. Global health (1973 onwards)
6. Scopus (1970 onwards)
7. Web of Science (1900 onwards)
8. SciELO (2002 onwards): Latin America focused database providing scholarly literature in sciences, social sciences, and arts and humanities published in leading open access journals from Latin America, Portugal, Spain, and South Africa; this was an important source of non-English language studies from developing countries in Latin America, particularly from Brazil.
9. Education Resources Information Centre (ERIC, 1966 onwards)
10. British Education Index (BEI, 1996 onwards)
11. Child Development & adolescent studies (CDAS, 1927 onwards)
12. Applied Social Sciences Index and Abstracts (ASSIA, 1987 onwards): important source for multidisciplinary papers; includes social work, nursing, mental health and education journals.

GRAY LITERATURE DATA SOURCES

13. Open grey (1992 onwards): includes theses, dissertations, and teaching guides
14. PROSPERO (2011 onwards): repository of pre-registered study protocols for systematic reviews for trial protocols for similar scoping reviews through PROSPERO.
15. Cochrane library (see above)
16. EMBASE (see above)
17. ERIC (see above)
18. CDAS (see above)

In all 14 unique databases were searched initially by 20th March 2020, and finally on **15th September 2020**. We also scanned the reference list of selected papers for other papers of possible interest which might have been missed in the literature search, particularly so for systematic and scoping review papers we found in our search.

Search Strategy

We developed literature search strategies using text words and medical subject headings (MeSH terms) related to the following themes:

– Executive function/Frontal lobe function/Frontal lobe damage/Adaptive Function and their synonyms and variants using truncation
– Assessments/Validation/reliability/norms/reproducibility/standardization of instruments and their synonyms and variants using truncation
– Children/adolescents and their variants using truncation
– Developing countries/lower-middle-income-countries/LMIC and their synonyms and variants using truncation

The search strategy was developed by a member of the study team (KKM) who had undergone extensive training from the Medical Library Services, University of Cambridge in conducting Systematic Reviews and in using search strategies in all the above-named databases. The search strategy was also reviewed by an experienced Medical Librarian who has extensive expertise in systematic review searching. The search terms were entered sequentially first with individual terms/synonyms connected with the Boolean operator "OR" as a theme-group to broaden the inclusivity of potential hits, while theme-groups were then connected with the Boolean operator "AND" entered into the advance search function to enhance the accuracy of potential hits. The full search strategy for MEDLINE is re-produced in Appendix I (see Supplemental Material) with more details of the search strategy. This search strategy was adapted for each of the 14 databases with each of their result documented in Table 7 below (see result section).

Study Selection

This has been extensively described in the scoping review paper. Six reviewers (all authors except AB) reviewed abstracts and critically appraised all papers. A minimum of 2 reviewers independently evaluated and screened each abstract and full paper at the abstract screening and full paper screening phases and compared results at each stage. Discrepancies were resolved by discussion and mutual agreement, or where there was no agreement, by arbitration by a 3rd reviewer. As a final resort, persistent disagreements were arbitrated by AB, the guarantor. Further, the pre-resolution inter-rater agreement ranged between 81.6%–88.9%, which was above the recommended minimum 80% agreement. In accordance with PRISMA recommendations, the selection process was documented in a flow diagram (see Fig. 1 below reproduced from scoping review paper) (Kusi-Mensah et al., 2021).

Data Collection Process

At least two reviewers independently extracted data from the screened articles using purpose-made data extraction charts:
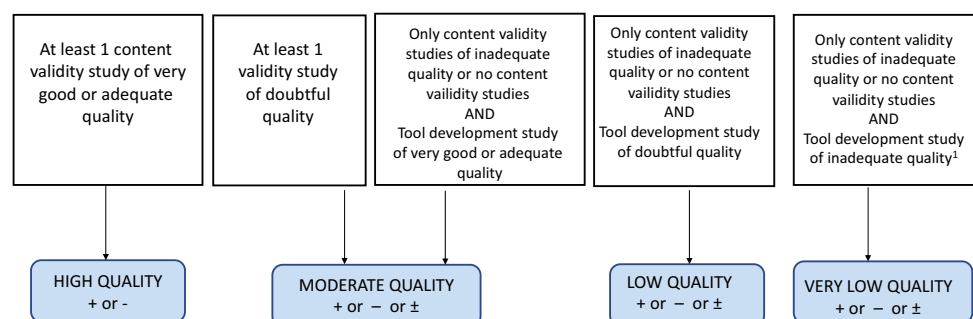
for the preliminary data collection the chart was designed following the PRISMA-ScR and PRISMA-P checklists (see scoping review paper for details); while for the critical appraisal, the chart was designed using items from the risk of bias assessment tool, the **Co**nsensus-based **S**tandards for the selection of health status **M**easurement **In**struments (COSMIN) checklist items guidelines (Mokkink et al., 2018b). First a calibration exercise was done for all 6 reviewers using a sample of 100 abstracts and 10 full paper articles to ensure uniform use of the screening criteria and the charting forms. The search results were uploaded and saved into Mendeley using the '*groups'* function, which allowed online collaboration and discussion among the reviewers. The data were charted using custom-made Excel spreadsheets downloaded from the COSMIN website (see 'help organising your risk of bias ratings' here: https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/).

Data Items

The following data points were collected and critically appraised using the following definitions:

Instrument reference this referred to the name of lead author and publication year of the paper.

Instrument name this referred to the instrument name and version under consideration.

Outcome Variable the outcome variables of interest were executive functioning and adaptive functioning as defined above.

Country Settings the desired setting was low-and-middle-income countries ('LAMIC') setting which was defined according to the World Bank list of lower income country (LIC- GNI per capita less than $1025), and middle-income countries which are split into Lower-middle-Income Country (LMIC- GNI per capita between $1026 to $3995) and upper-middle-income country (UMIC- GNI per capita between $3,996 TO $12,375) list (Cochrane Library, 2012; World Bank Group, 2019).

Fig. 1 Flowchart for grading quality of evidence based on Risk of Bias (Terwee et al., 2018a, b)

At least 1 content validity study of very good or adequate quality

At least 1 validity study of doubtful quality

Only content validity studies of inadequate quality or no content vaiildity studies AND Tool development study of very good or adequate quality

Only content validity studies of inadequate quality or no content vailidity studies AND Tool development study of doubtful quality

Only content validity studies of inadequate quality or no content vailidity studies AND Tool development study of inadequate quality[1]

HIGH QUALITY + or -

MODERATE QUALITY + or − or ±

LOW QUALITY + or − or ±

VERY LOW QUALITY + or − or ±

[1] Evidence is only based on the reviewer's rating

Type of Study: The critical appraisal was done according to the specific type of study done in the paper, with COSMIN criteria changing for different types of studies. 'Validation of an assessment tool' was defined according to specific items/criteria used for reliability and validity according to the COSMIN manual guidelines (Mokkink et al., 2014). Specific items (including their taxonomy and definitions) that were included as part of validation if they were reported upon were defined according to the COSMIN manual (see page 11, Table 1 of the manual) as follows:

- **RELIABILITY***:* This is a domain which covers the extent to which scores for patients who have not changed on the construct in question, are the same for repeated measurement under several conditions e.g. scores not changing when one uses different sets of items from the same instrument (internal consistency); not changing over time (*test-retest reliability*); not changing even when done by different persons on the same occasion (*inter-rater reliability*); or even when done by the same persons (raters or responders) on different occasions (*intra-rater reliability*). It is a domain that measures degree to which the measurement is free from measurement error. The larger domain "RELIABILITY" comprises of the following measurement properties:

   Internal consistency: The aspect of the 'reliability' domain which looks at the degree of the interrelatedness among the items. In other words, internal consistency is the maintenance of the same score for the same patient when different sets of items (which are related to each other) from the same instrument are used.
   Measurement error: The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured. One source of this error could be a lack of internal consistency, among other sources.
   Reliability: This is a measurement property of the domain "RELIABILITY" that specifically looks at the proportion of the total variance in the measurements which is due to 'true' differences between patients (i.e., variance which excludes all sources of measurement error). This measurement property tests *test–retest, inter-rater* and *intra-rater reliabilities* (see above for definitions).

- **VALIDITY:** This domain refers to the degree to which an instrument measures the construct(s) it purports to measure. This domain consists of the following measurement properties:

   Content validity (including **face validity**): a measurement property of validity which looks at the degree

to which the content or items of an instrument is an adequate reflection of the construct to be measured within the specific cultural context in question. It usually covers the following aspects of its measurement properties: ascertaining whether each item is *relevant* (i.e. refers to concepts that are familiar and common to the experience of the target audience/likely participants), *comprehensible* (i.e. words used can be understood at the language and educational level of the target audience) and *comprehensive* (i.e. items cover all the domains or aspects of the concept as understood by the target audience).

Construct Validity: a measurement property of validity which examines the degree to which the scores of the instrument are consistent with known hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the instrument validly measures the construct to be measured. This will comprise of the following aspects of measurement property subsets: ***Structural Validity*** which is the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured (i.e. if the construct is seen as a single domain, a factor analysis will not produce 2 or more factors for that same construct but just one factor reflecting that single domain); ***Cross-cultural validity***: which is the degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument; and ***Hypothesis-testing (construct validity)***: which is the degree to which scores produced by the instrument are consistent with a known true hypothesis. Hypothesis-testing can be further divided into two sub-categories: *discriminant validation* (where the hypothesis is testing the ability of the instrument to discriminate between 2 group- say a clinical versus healthy group such as an EF instrument's scores discriminating between a population of healthy children and children with ADHD in terms of ability to plan without distraction); *convergent validation* (where the hypothesis is testing the convergence of the instrument of interest- say an EF tool- with another instrument of related but different construct- such as an ADHD screening tool, such that there is good enough correlation between the scores produced by the two different but related tools).

Criterion validity: a measurement property of validity which looks at the degree to which the scores of an instrument are an adequate reflection of a 'gold standard'. For outcome measurement instruments the 'gold standard' is usually taken as the original full version

Table 1 COSMIN criteria for assessing **results** (Summary Measures) of psychometric properties (Mokkink et al., 2018b)

Measurement Property	Rating[a]	Comment
Structural Validity	+	**CTT:** CFA: CFI or TLI or NNFI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08[b] **IRT/Rasch:** No violation of unidimensionality[c]: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08 AND no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 AND no violation of monotonicity: adequate looking graphs OR item scalability > 0.30 AND adequate model fit: IRT: $\chi 2 > 0.01$ Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z- standardized values > -2 and < 2
	?	CTT: Not all information for ' + ' reported IRT/Rasch: Model fit not reported
	–	Criteria for ' + ' not met
Internal Consistency	+	At least low evidence[d] for sufficient structural validity[e] AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale[f]
	?	Criteria for "At least low evidence4 for sufficient structural validity[e]" not met
	–	At least low evidence[d] for sufficient structural validity[e] AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale[f]
Reliability (test–retest, inter rater and intra- rater)	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	–	ICC or weighted Kappa < 0.70
Measurement error	+	SDC or LoA < MIC[e]
	?	MIC not defined
	–	SDC or LoA > MIC[e]
Cross cultural validation/ measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's $R^2 < 0.02$)
	?	No multiple group factor analysis OR DIF analysis performed
	–	Important differences between group factors OR DIF was found
Hypotheses testing: construct val	+	The result is in accordance with the hypothesis[g]
	?	No hypothesis defined (by the review team)
	–	The result is not in accordance with the hypothesis[g]
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for ' + ' reported
	–	Correlation with gold standard < 0.70 OR AUC < 0.70
Responsiveness	+	The result is in accordance with the hypothesis[g] OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	–	The result is not in accordance with the hypothesis[g] OR AUC ≤ 0.70

AUC area under the curve, *CFA* confirmatory factor analysis, *CFI* comparative fit index, *CTT* classical test theory, *DIF* differential item functioning, *ICC* intraclass correlation coefficient, *IRT* item response theory, *LoA* limits of agreement, *MIC* minimal important change, *RMSEA* Root Mean Square Error of Approximation, *SEM* Standard Error of Measurement, *SDC* smallest detectable change, *SRMR* Standardized Root Mean Residuals, *TLI* Tucker-Lewis index

[a]" + " = sufficient,"– " = insufficient, "?" = indeterminate

[b]To rate the quality of the summary score, the factor structures should be equal across studies

[c]unidimensionality refers to a factor analysis per subscale, while structural validity refers to a factor analysis of a (multidimensional) patient-reported outcome measure

[d]As defined by grading the evidence according to the GRADE approach

[e]This evidence may come from different studies

[f]The criteria 'Cronbach alpha < 0.95' was deleted, as this is relevant in the development phase of a PROM and not when evaluating an existing PROM

[g]The results of all studies should be taken together, and it should then be decided if 75% of the results are in accordance with the hypotheses

of an instrument where a shortened version is being evaluated.

- **RESPONSIVENESS:** Another broad domain which looks at the ability of an instrument to detect change over time in the construct to be measured

The above definitions of measurement properties also largely conformed with the definitions given in The Standards for Educational and Psychological Tests manual American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2018). Any paper that reported information on any of the above was considered as eligible for having included an eligible outcome measure. An important judgement decision that was made in respect of classification of what constituted "**a study**" was that any individual validation conducted in any given research project/paper was regarded as a "study" in accordance with COSMIN manual recommendations (Prinsen et al., 2018). So, for example, a given paper might report the conduct of say- construct validation (hypothesis testing), cross cultural validation and structural validation of one instrument all within the same paper. This was thus reported as three studies reported within one paper.

Mode of administration: This was defined as the way in which the instrument was administered, i.e. whether the instrument was a performance-based task, or an informant-based tool (i.e., self-reported or parent/proxy-based questionnaire etc.). For performance-based tasks, the number of trials of a task was considered as the "number of items" when it came to evaluating aspects like adequacy of sample size in Structural validity or Cross-cultural validity studies.

Sub-domains: the number of sub-scales or sub-domains or items of interest of the tool in question reported on.

Sample size: number of participants used.

Demographics: mean age and gender percentages of sample.

Local Settings: whether study was predominantly set in rural or urban settings (or both).

Condition: whether study was conducted among health sample or clinical sample, and if so, what clinical condition.

Language of population: What local language-group was the study conducted among.

Risk of Bias in Individual Studies

Each study was evaluated for both the quality of its methodology (Risk of bias assessment) and the quality of its results following the COSMIN guidelines (Mokkink et al., 2018a). This COSMIN guidelines came in 2 separate manuals: 1. *The COSMIN methodology for assessing the content validity of Patient-Reported Outcome Measures (PROMs)* (C. B.

) which focuses on assessing risk of bias in content validity studies, and 2. *the COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs)* (Prinsen et al., 2018) which focuses on risk of bias for all other types of validity studies. This section describes the risk of bias rating for the **methodology** used while the next section describes that of *results*. The risk of bias evaluation was done at both the study level and the instrument level. A critical appraisal of each study was done independently by at least 2 reviewers and compared, and consensus reached. The methodology was rated using 4 codes: 'very good', 'adequate', 'doubtful' and 'inadequate'. So, for example, in evaluating a Structural Validity study for an instrument developed using classical test theory (CTT), the first item to be evaluated in the methodology would be whether Confirmatory Factor Analysis (CFA) was performed (scored as: 'very good'), as opposed to Exploratory Factor Analysis (EFA score: "adequate"), or no CFA or EFA was performed at all (score: "inadequate"). For giving the overall rating of each study, the 'worst score count' system was used as per the COSMIN guideline, for the simple reason that "poor methodological aspects of a study cannot be compensated by good aspects" (Mokkink et al., 2018a; Caroline B). For a detailed explanation of the individual criteria for each respective rating in all the scoring domains refer to pages 47 – 63 of the 'COSMIN Manual for Systematic Reviews of PROMs' (Mokkink et al., 2018a) which is downloadable for free here: https://www.cosmin.nl/tools/guideline-conducting-systematic-review-outcome-measures/.

Further, any qualitative studies for content validation found for new tool development or adaptation studies were critically appraised using the separate COSMIN guidelines for evaluation of content validation studies (Terwee et al., 2018a, b) (which is different from the COSMIN manual whose methodology was "developed in 2016 in a Delphi study among 158 experts from 21 countries". This is also downloadable for free at the above URL of the COSMIN website. Here also the methodology of the content validation/instrument development study was appraised for risk of bias (also rated as: 'very good', 'adequate', 'doubtful' and 'inadequate'- see COSMIN box 1 and box 2 in excel spreadsheet in supplemental material). The extracted and critically appraised data were then summarised per each identified instrument, and an overall rating given to the quality of evidence according to the COSMIN guideline, using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach (Schünemann et al., 2008), as described below (see Table 2 below).

Summary Measures

This section describes evaluation of **results** reported by the papers. The principal summary measures to be evaluated

Table 2 COSMIN criteria and rating system for evaluating the content validity of Instruments (Terwee et al., 2018a, b)

Name of the Instrument or subscale:	Instrument development study	Content validity study 1	Content validity study 2[b]	Rating of reviewers[*]	OVERALL RATINGS PER PROM[c] (see step 3b in COSMIN manual)	QUALITY OF EVIDENCE (see step 3c in COSMIN manual)
Criteria (see Table… in Appendix for guidelines)	$+/-/\pm/?$[a]	$+/-/\pm/?$[a]	$+/-/\pm/?$[a]	$+/-/\pm/?$[a]	$+/-/\pm/?$[a]	High, moderate, low, very low
Relevance						
1. Are the included items relevant for the construct of interest?[d]						
2. Are the included items relevant for the target population of interest?[d]						
3. Are the included items relevant for the context of use of interest?[d]						
4. Are the response options appropriate?						
5. Is the recall period appropriate?						
RELEVANCE RATING (see COSMIN page 58, Table 3)						
Comprehensiveness						
6. Are all key concepts included?						
COMPREHENSIVENESS RATING (see COSMIN page 58)						
Comprehensibility						
7. Are the Instrument instructions understood by the population of interest as intended?						
8. Are the Instrument items and response options understood by the population of interest as intended?						
9. Are the Instrument items appropriately worded?						
10. Do the response options match the question?						
COMPREHENSIBILITY RATING (see COSMIN page 58)						
CONTENT VALIDITY (see COSMIN page 59, Table 4)						

[*]Rating of Reviewers excluded from this review for reasons stated above

[a]Ratings for the 10 criteria can only be $+/-/?$. The RELEVANCE, COMPREHENSIVENESS, COMPREHESIBILITY, AND CONTENT VALIDITY ratings can be $+/-/\pm/?$

[b]Add more columns if more content validity studies available. 3If ratings are inconsistent between studies, consider using separate tables for subgroups of studies with consistent results

[c]These criteria refer to the construct, population, and context of use of interest in the systematic review

across studies differs according to study type. Table 1 below (reproduced from COSMIN manual for Systematic Reviews) summarises the various summary measures and criteria for all study types except content validity (reported later). After rating the methodology, the results reported were then rated as: '+' for sufficient, '–'for insufficient and '?' for indeterminate. For the construct validation, the acceptable measure of effect of the instrument of interest when compared with another similar instrument was pre-defined by the study team as an expected correlation between the two of at least 0.5 as recommended by the COSMIN guideline (Mokkink et al., 2018a). For comparison of healthy versus clinical populations on EFs or AFs, the pre-defined minimum acceptable difference was a statistically significant difference (Mokkink et al., 2018a) with an odds-ratio of at least 1.5. The results for individual assessment tools from different papers were then qualitatively pooled together and given an overall rating also according to 'sufficient', 'insufficient' and 'indeterminate'.

For the content validation, results were rated as: '+' for sufficient, "– " for insufficient, '?' for indeterminate and "±" for inconsistent. The general rule for giving a sufficient rating per criterion was as follows:

+ ≥ 85% of the items of the instrument (or subscale) fulfill the criterion.
– < 85% of the items of the instrument (or subscale) does fulfill the criteria.
?No(t enough) information available or quality of (part of a) the study inadequate.
± Inconsistent results.

The ten (10) individual criteria for synthesizing results of Content Validity are specified in Table 2 below. The results of each criterion are rated taking into consideration the risk of bias assessment (i.e., quality of methodology- which would have been rated in COSMIN box 1 and box 2) and rated accordingly. The results rating was recorded in Table 2 below.

In using the COSMIN content validation manual (Terwee et al., 2018a, b, we made the following modifications for the following reasons. Firstly, although the COSMIN content validation manual allowed for three types of ratings to be assessed- 1. rating of original development study (if available or relevant) 2. rating of all adaptation or content validation studies, and 3. independent rating of all items by reviewers- following which a summary score would be given for the content validity of the instrument (see page 52 of COSMIN content validity manual), we chose to drop the third type of rating (the reviewers independent rating) and rather just stick with the first two- development study rating and adaptation/content validity studies rating. We followed this course of action for two reasons: firstly obtaining the

actual instrument under review (rather than just the published validation study on that instrument) was not always going to be practical or even possible sometimes since several of them were copyrighted material under commercial licence for which we would have to pay a fee to obtain the instrument; the decision was therefore made that if we were not going to be able to obtain the full instrument for ALL eligible papers, then it would be unfair to evaluate some instruments on those three levels, and compare these with others that were evaluated on just 2 or even 1 level simply because some instruments were freely available and others were not.

Secondly, given our varying levels of expertise/experience and given the wide variety of countries from which we were going to evaluate papers, we also wanted to further minimise the amount of subjectivity that would go into us personally reviewing each item on candidate instruments as to the appropriateness of items for each country across such a diverse array of country-specific sociocultural realities and languages that we were not personally familiar with; and then going on to factor in these subjective impressions into the overall rating of the instrument. We therefore thought it would be better to leave the review of the appropriateness of individual items for individual country contexts (if done at all) to the local experts who might have been involved with the individual projects. We thus thought it best to restrict ourselves to only the review of the published development and adaptation/content validation papers using the specified standard criteria published in the two COSMIN manuals across board. This we felt would be more objective and give all tools an even playing field.

A practical effect of this approach in determining the OVERALL RATING of results of an instrument (summing up ALL available studies) was as follows: the COSMIN manual states that since there is supposed to be a Reviewers' independent rating of each item (apart from the rating of the published development study and rating of the adaptation studies), where the results of the development and adaptation studies are at variance, the Reviewers' own independent rating should be used, hence it should not be possible to give an 'indeterminate' OVERALL RATING for any instrument (see page 60 of COSMIN content validity manual). However, contrary to this, we *did* make it possible to give an 'indeterminate' overall results rating ('?') where the rating for the development study and the adaptation study were at variance since we did not do any Reviewers' rating to fall back on for an overall rating.

Another modification that was made in the application of the COSMIN criteria was in the risk of bias (RoB) evaluation for performance-based tasks. In evaluating the content validity of performance-based tasks, we did not deem it appropriate to assess "comprehensiveness" of the task (how well does the task cover all aspects of the construct at hand)

among the caregivers/patients, because being lay people it seemed a bit unfair and unreasonable to ask them to evaluate how comprehensive such highly specialized tasks were to the construct of interest. Therefore, we only evaluated "comprehensiveness" among subject matter experts. However, for such domains as "relevance", we went ahead and evaluated these performance-based tasks for relevance among caregivers/patients as per COSMIN guidelines because we felt that asking about the relevance of a task to the experience of a patient/caregiver was perfectly legitimate in such circumstances. As noted by Semrud-Clikeman and colleagues, when a task is unfamiliar to a child in a LAMIC but familiar to children in Western cultures, administering such a task to the LAMIC child may measure the ability of LAMIC children to adapt to new situations rather than their ability to complete the actual task, rendering the scores invalid for the domain being measured (Semrud-Clikeman et al., 2017), hence the need to assess even performance-based tasks for relevance.

A final modification made in applying COSMIN was in evaluating the RECALL PERIOD (see item 5 in Table 2 below). Because behaviours being described by the two constructs we evaluated (Executive and Adaptive functioning) were not "symptoms" per se, items concerning RECALL PERIOD were deemed as 'NA- not applicable'. This was because given that these behaviours were not symptoms with a "time of onset", but rather normal behaviours expected at various age-appropriate milestones, it did not seem suitable to evaluate whether authors concerned themselves with recall periods of how far back the item was to be evaluated and would thus have been unfair to have penalized them for not doing so (as per COSMIN guidelines) given the context. Further, for performance-based tasks only, evaluation of appropriateness of RESPONSE OPTIONS (see item 4 in Table 2 below) such as 'often', 'sometimes', 'never' etc. as seen in questionnaires was also deemed 'not applicable'. This question while making perfect sense in the context of an informant-based instrument (which has various response options such as 'sometimes', 'never', 'often' etc.), would not make sense in the context of a performance-based tasks such as Go/No-go or Wisconsin Card Sorting Test where the response is an action.

Synthesis of Results

This section describes how methodology and results ratings of individual studies were synthesized and summarized across several papers for each instrument. After rating the methodology and results of individual studies and qualitatively pooling these together to give an overall rating per instrument, the overall **quality of the evidence** for

the reported results of the <u>instrument</u> in question (taking into consideration all published papers for that instrument) was then graded following the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach (Schünemann et al., 2019). In other words, after giving the overall rating of the results of a study type per instrument, this result was then accompanied by a grading of the <u>quality of evidence</u> using the GRADE approach.

The instrument in question is usually assumed to be of high quality of evidence from the start (see Table 3 below reproduced from page 34 of COSMIN manual (Mokkink et al., 2018a)), and progressively downgraded to lower levels, taking into consideration the best rating for Risk of Bias, inconsistencies of results from different studies (i.e. between study variability/heterogeneity), imprecision of results (i.e. down-grading for low pooled sample sizes) and indirectness (i.e. downgrading for studies that were (partly) performed in another population or another context of use than the population or context of use of interest in the systematic review, for example, evidence from a mixed sample with adults rather than just children and adolescents).

The final quality of evidence was then graded as: 'high', 'moderate', 'low' or 'very low' depending on which level the instrument was finally left at following successive downgrades. This grading serves to indicate how confident one can be that the overall rating is trustworthy (Terwee et al., 2018a, b). So, for example, in this scheme, 2 different tools- Instrument A and Instrument B- might both have a rating of "+" (sufficient) in their structural validity, but Instrument A might have a grading of "high" for the quality of evidence supporting its rating of "+", while Instrument B might have

Table 3 Modified GRADE approach for grading the quality of evidence (Mokkink et al., 2018)

QUALITY OF EVIDENCE	LOWER IF
High	Risk of bias:
	-1 Serious
Moderate	–2 Very serious
	–3 Extremely serious
Low	
	Inconsistency:
Very Low	–1 Serious
	–2 Very serious
	Imprecision:
	–1 total n = 50–100
	–2 total n < 50
	Indirectness:
	–1 Serious
	–2 Very serious

a grading of "very low" for the quality of evidence supporting its own "+" rating.

To go into more detail, Risk of Bias (RoB) was given an overall grade of "No risk", "serious", "very serious" or "extremely serious" risk depending on the situation specified against each grade in Table 4 below (reproduced from page 34 of COSMIN manual (Mokkink et al., 2018)). So, for example, an instrument that had multiple studies of 'inadequate' quality, or only one study of 'doubtful' quality in its methodology rating would be rated as having 'very serious' RoB (see Table 3) and would be downgraded by –2 steps from say "High" quality of evidence to "Low" quality of evidence according to GRADE criteria (see Table 4).

A similar principle of progressive downgrading was followed for Inconsistency of Results. The pooled or summarized results were rated based on a majority of individual (study) results as 'sufficient' or 'insufficient' (i.e., if most consistent results were 'sufficient' overall rating would be sufficient, if majority of consistent results were 'insufficient', overall rating for consistency would be 'insufficient') and then downgraded accordingly using the GRADE approach in Table 3. For the purposes of determining the degree to which the downgrade was to be made for inconsistency, COSMIN recommended that the review team come to consensus about that decision. Consequently, the team decided that a downgrade of –1 ('serious') would be made for inconsistency if a super majority of results (above 75%) were consistent, in which case the results would be rated as 'sufficient' but quality of evidence would be downgraded by only –1 (say from 'high' to 'moderate') for inconsistency. However, if less than 75% but more than 50% of studies found agreement (e.g., 60% of studies gave 'sufficient' results and 40% gave 'insufficient' results for any given parameter- say, Test–retest reliability) then the overall rating would be given as "sufficient" but quality of evidence downgraded by –2 levels ('very serious') say from "High" to "Low" quality. If less than 50% of studies gave a consistent result (e.g., 40% of studies gave 'sufficient' and 60% gave 'insufficient' results) then the overall rating would be given as 'insufficient' but quality of evidence downgraded by –2 levels.

Having given this explanation, it must be stated that a slight departure was made in applying this rule when it came to Construct Validation, though this alternative approach was still in accordance with COSMIN recommendations (see page

34 of COSMIN manual). For Construct validation, where the two hypotheses being tested were of a fundamentally different nature (e.g., discriminant validation, as opposed to say a convergent validation), the strategy adopted was to summarize the results by sub-groups (i.e. summarize according to each individual hypothesis) rather than pooling all the results together. So, in this case, the results of all convergent validity studies will be pooled separately from the results of all discriminant validation studies treating both as "sub-groups" of construct validation, rather than pooling together all "construct validity" studies and using the 75% majority rule.

For Imprecision, the total sample size of all the included studies (on the validation study in question) were simply pooled together and the quality of evidence simply downgraded with one level when the total sample size of the pooled or summarized studies was below 100, and with two levels when the total sample size was below 50, as per the COSMIN manual (see page 35) (Mokkink et al., 2018). For Indirectness, this was typically not a major concern in this study since part of the screening criteria was that the study be performed primarily among children (aged 5 – 18 years). However, in the few instances where there were mixed study populations involving say young adults (19 – 25 years), only a single level downgrade ('serious') of quality of evidence was made, as recommended by the COSMIN manual.

This downgrading was done successively from "high" downwards following evaluation of 'risk of bias', 'inconsistency of results', 'imprecision' and 'indirectness' as explained above. The results of the quality of evidence for all instruments were then summarized and presented below in the results section. The interpretation of what each of the ratings (individual studies) and the grading (overall quality of evidence for individual instruments) is shown in Table 5 below.

For the synthesis of results of the content validation, following the separate COSMIN guidelines for evaluation of content validation studies (Terwee et al., 2018a, b) a similar but slightly different approach was used for results synthesis. First, as stated above with the other types of validations, each criterion for results of content validity was given a rating (±/?) that took into consideration the methodology of content validation rating (see COSMIN box 1 and box 2 in Excel spreadsheet in supplemental material), which rating was recorded in the corresponding item in Table 2. After rating each criterion, a summary rating is then given to the

Table 4 Instructions on Downgrading for Risk of Bias (Mokkink et al., 2018a)

Risk of bias	Downgrading for Risk of Bias
No	There are multiple studies of at least adequate quality, or there is one study of very good quality available
Serious	There are multiple studies of doubtful quality available, or there is only one study of adequate quality
Very serious	There are multiple studies of inadequate quality, or there is only one study of doubtful quality available
Extremely serious	There is only one study of inadequate quality available

Table 5 COSMIN Ratings/grade and their Interpretations (Mokkink et al., 2018b)

RATINGS	DEFINITION/INTERPRETATION
Tool Development/Adaptation[a]	
+ or 'sufficient'	≥ 85% of the items of the instrument (or subscale) fulfil the criterion
– or 'insufficient'	<85% of the items of the instrument (or subscale) does fulfil the criteria
? or 'indeterminate'	Not enough information available or quality of (part of) the study inadequate
± or 'inconsistent'	Not used in rating individual items/criteria, but only used in rating the RELEVANCE, COMPREHENSIVENESS, COMPREHESIBILITY, and overall CONTENT VALIDITY of a study where results for each section are mixed
GRADE quality of evidence	
High	We are very confident that the true measurement property lies close to that of the estimate* of the measurement property
Moderate	We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different
Low	Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property
Very Low	We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property

These definitions were adapted from the GRADE approach (Schünemann et al., 2019)

*Estimate of the measurement property refers to the pooled or summarized result of the measurement property of an instrument

[a]refers to assessment of 'relevance', 'comprehensibility' and 'comprehensiveness'

RELEVANCE box (criteria 1 – 5), COMPREHENSIVENESS box (criteria 6), and COMPREHENSIBILITY box (criteria 7 – 8) as either ±/?/± (see COSMIN content validity manual, page 58, Table 3 for rules for summarising rating per box).

Then, a CONTENT VALIDITY rating is given **for the study** according to COSMIN guidelines (see COSMIN content validity manual, page 59, Table 4 for rules for summarising content validity rating per study) also as ±/? /±. Finally, the content validity ratings from all available studies are summarised **for the instrument** and an overall quality of evidence is given according to GRADE approach as high', 'moderate', 'low' and 'very low', all following the specified COSMIN guidelines (see steps 3b and 3c of COSMIN Content Validity Manual on page 60–62).

This overall rating is done factoring in both the score of the content validity of the eligible adaptation study/studies under consideration, as well as the score of the content validity of the original instrument development study of that instrument following a critical appraisal as per usual COSMIN guidelines whether or not this original development study was captured by the search strategy. This would mean attempting to obtain the original instrument development study and appraising it, even where it would <u>not</u> otherwise fall into the eligibility criteria (for example, when appraising instruments developed in HICs that were being adapted for use in a LAMIC- the original development study (in the HIC) will have to be critically appraised, and the score factored into the **overall content validity rating <u>of the instrument</u>**). Where the original development study could not be

obtained (for whatever reasons) to be critically appraised, an estimation of their probable score based on publicly available information and an assumption of the best case scenario of their likelihood of fulfilling (or failing to fulfil) the COSMIN criteria was made, and this was then used in estimating their overall content validation score while factoring in the score of the adaptation studies that were eligible as just described, with cogent reasons provided for their estimated score. Wherever this type of estimation was done, this has been clearly indicated in the results and in the supplemental material for the sake of transparency.

The criteria for rating quality of evidence are reproduced in Table 6 and Fig. 1 below. Because of the decision to

Table 6 Modified GRADE approach for grading the quality of evidence for Content Validity (Terwee et al., 2018a, b)

STUDY DESIGN	QUALITY OF EVIDENCE	LOWER IF
At least 1 content validity study	High	Risk of bias: –1 Serious
No content validity studies	Moderate	–2 Very serious –3 Extremely serious
	Low	Inconsistency:
	Very Low	–1 Serious –2 Very serious Indirectness: –1 Serious –2 Very serious

exclude 'Reviewer rating' level, the COSMIN guideline for summarising and giving an overall content validity rating per instrument was modified to allow for an 'indeterminate' rating (?). As recommended by COSMIN, where multiple content validity studies were reported for any single instrument, 'indeterminate' studies were ignored, and only 'sufficient' or 'insufficient' results were considered and summarised. However where only a single study of 'indeterminate' result was found for content validity, rather than ignore the "?" rating, we chose to report this (?) as the final summary rating. The quality of evidence was then determined as 'high', 'moderate', 'low' and 'very low' according to the modified GRADE approach as outlined in COSMIN guidelines.

Downgrading the GRADE quality of evidence rating for risk of bias was done as follows. Downgrade for 'serious RoB' (–1 level, e.g., from high to moderate) was made if the adaptation/content validity study was of 'doubtful' methodological quality. If there were no content validity studies (or only of inadequate quality) and the instrument development study was of doubtful quality, downgrade for 'very serious RoB' (–2 levels) was made. If there was no adaptation/content validity study (or if only one of 'inadequate' quality) and the instrument development study was also of 'inadequate' quality, RoB downgrade was –3 levels (extremely serious) to 'very low' quality of evidence. Figure 1 (reproduced from

COSMIN content validity manual) summarizes the flowchart for downgrading for RoB. For inconsistency, downgrade to 'serious' (–1 level) was made if the results rating for the instrument development study and the adaptation/content validation study were inconsistent.

Risk of Bias Across Studies

Meta-biases like publication bias were reduced by searching the 'grey literature' as indicated above. However, it was difficult to guard against selective reporting within studies as this was a largely qualitative systematic review which therefore precluded requesting for original datasets from authors to conduct a "meta-analysis" of results.

Results

Study Selection

Table 7 (reproduced from the scoping review) summarises the results of the initial search in each individual data source, along with the dates of coverage of the search in each database.

Following de-duplication, manual screening of full papers for eligibility, 51 full papers were found to be eligible for

Table 7 Database Results and Dates

Database	Date of search coverage	Initial search by 20/03/20	Initial de-duplication	Updated search by 15/09/20	Updated de-duplication
MEDLINE via OVID	1946 – 05/09/20	803	662	939	760
EMBASE[a]	1974 – 05/09/20	27	26	29	27
Cochrane Library[a]	*11/09/20	67	49	68	48
PsychINFO	1894 – 11/09/20	1314	985	1490	812
Global Health	1973 – 11/09/20	38	14	38	13
Scopus	1970 – 11/09/20	1093	1080	1134	910
Web of Science	1900 – 11/09/20	1730	1254	1820	1184
SciELO	2002 – 15/09/20	29	28	29	26
ERIC[a]	1966 – 15/09/20	19	2	19	2
BEI	1996 – 15/09/20	2	2	3	1
CDAS[a]	1927 – 15/09/20	13	3	13	2
ASSIA	1987 – 15/09/20	84	81	87	48
Open Gray[a]	1992 – 15/09/20	4	4	4	4
PROSPERO[a]	2011 – 15/09/20	1	0	2	0
Other (reference lists etc.)	–	–	–	0	0
TOTAL ABSTRACTS				**5675**	**3837**

ERIC Education Resources Information Centre, *ASSIA* Applied Social Sciences Index and Abstracts, *CDAS* Child Development & Adolescent studies, *BEI* British Education Index

*Although first established in 1996, composite nature of Cochrane library means it does not have a "start date" as other biomedical databases do

[a]Includes theses, dissertations, teaching guides and other such non- peer-reviewed "grey" literature

full data extraction and critical appraisal. Web of Science showed a surprisingly higher number of results than PsychINFO possibly because of the search strategy that was used: because the search terms, ab initio, did not preclude studies that only used these assessment tools in psychometrically unrelated papers without necessarily reporting psychometric properties (these were later excluded on manual inspection of abstracts and full review papers). Therefore, Web of Science, which is a database dedicated to multidisciplinary research (and hence also likely to have a lot more abstracts than the more specialised PsychINFO), was likely to include a significant proportion of abstracts of such "use studies" which were later excluded as ineligible upon manual screening by the reviewers. The screening process and reasons for exclusion are summarised in the PRISMA flow diagram below in Fig. 2.
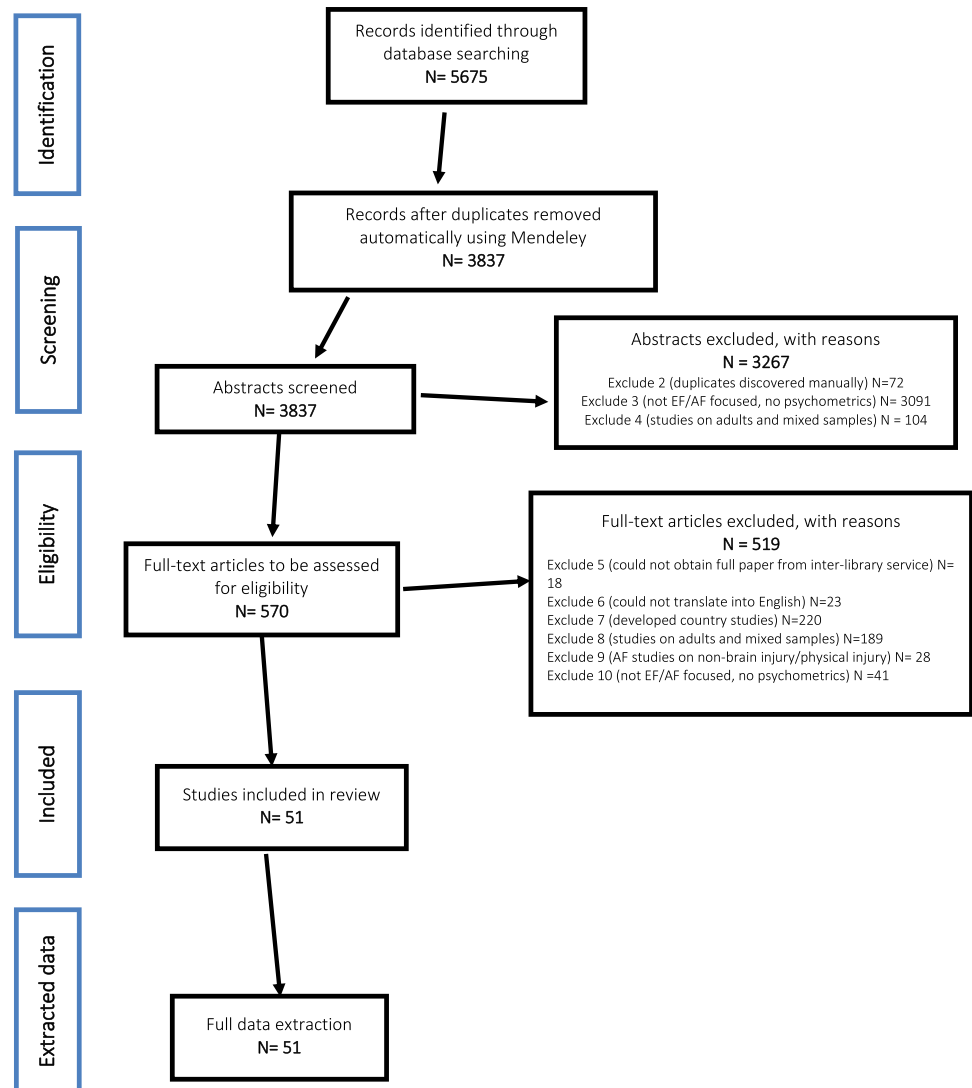
PRISMA Flow Diagram for Study Selection

Study Characteristics

Details of the study characteristics of each paper are reported elsewhere in the scoping review paper by the authors (Kusi-Mensah et al., 2021), and provided in the Appendix of this paper. However, some key highlights are pertinent here. A total of 163 studies (as defined under *Data Items* above) were reported in 51 papers. The most frequently conducted types of study were Structural Validity and Construct Validity/Hypothesis testing studies at 38 studies each (23.3% each of total individual studies), followed by Internal consistency studies at 27 (16.6%). In terms of geographical regions, the Americas (Central and South America) (30.4% of papers), Sub-Saharan Africa

Fig. 2 PRISMA flow diagram of study selection

PRISMA Flow Diagram for Study Selection

Identification

Records identified through database searching
N= 5675

Screening

Records after duplicates removed automatically using Mendeley
N= 3837

Abstracts screened
N= 3837

Abstracts excluded, with reasons
N = 3267
Exclude 2 (duplicates discovered manually) N=72
Exclude 3 (not EF/AF focused, no psychometrics) N= 3091
Exclude 4 (studies on adults and mixed samples) N = 104

Eligibility

Full-text articles to be assessed for eligibility
N= 570

Full-text articles excluded, with reasons
N = 519
Exclude 5 (could not obtain full paper from inter-library service) N= 18
Exclude 6 (could not translate into English) N=23
Exclude 7 (developed country studies) N=220
Exclude 8 (studies on adults and mixed samples) N=189
Exclude 9 (AF studies on non-brain injury/physical injury) N= 28
Exclude 10 (not EF/AF focused, no psychometrics) N =41

Included

Studies included in review
N= 51

Extracted data

Full data extraction
N= 51

(21.4% of papers) and the Middle East (17.9% of papers) were the top 3 performing regions in terms of number of papers published from LAMICs. Most studies conducted involved urban populations at (76.7%), and exclusively healthy populations (60.8%) as opposed to clinical populations (with health controls).

Results: Risk of Bias Within and Across Studies and Results of Individual Studies

Table 8 shows a summary of the results (synthesis of results for all studies for an instrument pertaining to each measurement property) of the critical appraisal for all the types of validity except content validity (see Table 9 for content validity results). The quality of evidence was summarised using the GRADE approach (Schünemann et al., 2019) shown by the colour coding in the table. For the details of the rating of individual studies, see the following boxes in the Supplemental Material: for Risk of Bias of various measurement properties of validations in individual studies- COSMIN boxes 3–10, for results of individual studies- COSMIN Supplemental Table 2; for details of the reasons behind each rating, see comments on 'Summary Psychometric Ratings' in Supplemental Material.

In this systematic review, 40 unique tools coming in 49 version/variants were identified as having been either developed or adapted/validated for use among children in LAMIC countries. Figures 3 and 4 below provide a snapshot summary of the overall results of this critical appraisal in pie chart format. Figure 3 shows the overall summary of the results ratings of the various measurement properties, while Fig. 4 shows the summary of the overall GRADE ratings of the quality of evidence for the various measurement properties. None of these tools showed full validation (i.e., validation in all domains assessed) in LAMICs. Only 11% of adaptation/development studies showed "sufficient" content validation, while 33% showed sufficient results for reliability and 36% did same for cross cultural validity. However, for internal consistency, structural, convergent and discriminant validities results were good at 91.3%, 53%, 75% and 84% "sufficient" results respectively. When it came to the quality of evidence as well (see Fig. 4), the pattern of results was similar with 100% of content validation studies, 86% of cross-cultural validations and 90.5% of reliability studies being of either low or very low quality. Internal consistency was generally of high quality (87%), with 50.1% of structural validity, 75% of convergent/concurrent studies and 71.8% of discriminant validity studies showing either high or moderate quality of evidence supporting their results.

Going into the specific instruments (see Table 8), the BRIEF (in its various versions) (Gioia et al., 2000), Wisconsin Card Sorting Test (WCST) (Berg, 1948), Go/No-go (GNG) (Luria, 1973) and the Rey-Osterrieth complex figure

(ROCF) (Rey, 1941) had the most validations undertaken, both in terms of the variety of validations and the number of studies. None of them though had either 'high' or even 'moderate' quality evidence for 'sufficient' psychometric properties across all the measured psychometric domains. The BRIEF (in all its versions) showed high quality of evidence for 'sufficient' internal consistency, high quality evidence for 'sufficient' construct validity (for BRIEF-pre-school version) and moderate to low quality evidence for sufficient construct validity (mostly convergent validity) for BRIEF-parents and BRIEF-teachers respectively. The WCST showed moderate quality evidence for 'sufficient' internal consistency and low-quality evidence for 'sufficient' construct validity across several studies, while the ROCF showed high quality evidence for 'sufficient' construct validity (discriminant validity). The GNG showed high quality evidence for both 'sufficient' structural validity and 'sufficient' internal consistency. For cross-cultural validity, only two instruments had moderate quality evidence for a 'sufficient' rating: GNG and the Behavioural Assessment System for Children (BASC) (Reynolds & Kamphaus, 1992). They both also showed high quality evidence for sufficient structural validity and internal consistency. However, BASC showed low quality evidence for sufficient construct validity in the countries of interest.

For adaptive functioning, the most assessed tools were the Vineland Adaptive Behaviour Scales (VABS) (Sparrow et al., 1984, 2005) and the Child Function Impairment Rating Scale (CFIRS) (Tol et al., 2011), a newly developed tool from Indonesia. Of these two, while both had high quality evidence for 'sufficient' internal consistency, VABS showed low-quality evidence for 'insufficient' structural validity, and indeterminate cross-cultural validity. The CFIRS showed low quality evidence for sufficient construct validity but reported no formal assessment of comprehensibility and comprehensiveness among subjects during tool development (hence was rated 'indeterminate' in both) with methodological shortcomings in relevance assessment (hence rated 'insufficient').

Table 9 shows a summary of the results of the critical appraisal for content validity. First few columns show results of **individual studies** content validation studies, with the "overall rating" column showing the **synthesis of results** for all content validation studies pertaining to **an instrument**, and the colour coding showing the GRADE rating for quality of evidence. For details of the reasons behind each results rating, see comments on 'Summary Content validity' in supplemental material; and for Risk of Bias of content validation for individual instrument development studies, see COSMIN box 1, while that for content validation of adaptation studies is in COSMIN box 2 in the supplemental material.

Here, 16 unique tools coming in 18 version/variants had either instrument development studies (i.e., brand new

Table 8 Summary of Results (Synthesis of Results) for Critical Appraisal of Psychometrics of Executive Functions and Adaptive Function tools Validated for use Among Children in LAMICs

Instrument	Structural validity				Internal consistency				Cross cultural validity				Reliability				Construct validity				Responsiveness/Measurement Error (ME)			
	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall	Ref.	N	meth. qual	Overall	Ref.
BRIEF (parent) (Gioia et al., 2000)	277	RoB: -3 1 study of 'inadequate' qual	? Did not do EFA or CFA only PCA	Carim D. 2012	462	RoB: 0 2 studs of 'very good' qual.	+ Cronb. Alpha = 0.88 to 0.93	Carim 2012; Burkey 2015	81	RoB -3, Imp: -1 1study of 'inadequate' qual	? No MG CFA done, rather did PCA	Baker E. 2011	603	RoB: -3 1 study of 'inadequate' qual.	- ICC ranged 0.5 to 0.57	Chernoff M. 2018	663	RoB 0, -1 Imp: -1, 0 2 studies for discriminant and convergent validity	- (convergent); + (discriminant)	Chernoff M. 2018; Zarrabi M. 2015	ME: 603	ME: RoB -3 1 std of 'inadequate' qual.	ME: ? MIC' not defined	Chernoff M. 2018
BRIEF (teacher) (Gioia et al., 2000)	642	RoB: -3 2 stds of 'inadequate' qual	+ CFI bet. 0.99 to 1.0, RMSEA bet. 0.0 to 0.33	Carim D. 2012; Amani M. 2018	642	RoB 0 Inc: 0 Imp: 0 Ind: 0 2 stds of 'very good' qual.	+ Cronb. Alph = 0.79 - 0.99	Carim D. 2012; Amani M. 2018									360	RoB -1 Ind: -1 1 std of 'adequate' qual.	+ convergent validity for WISC	Amani M. 2018				
BRIEF (self) (Gioia et al., 2000)	112	RoB: -3 1 std of 'inadequate' qual	? Did not do EFA or CFA only PCA	Carim D. 2012	112	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 std	+ Cronb. Alpha = 0.95	Carim D. 2012																
BRIEF (pre-school) (Isquith et al., 2004)	537	RoB -2 2 stds: 1 'inadequate', 1 'doubtful' qual.	? Did not do EFA or CFA	Rincon-Diaz M. 2017; Selvam S. 2018	537	RoB 0 Inc: 0 Imp: 0 Ind: 0 2 stds of 'very good' qual.	+ Cronb. Alph. Bet. 0.60 and 0.93	Rincon-Diaz M. 2017; Selvam S. 2018	412	RoB -3 1 std of 'inadequate' qual.	? no MG CFA done, only did student t test	Selvam S. 2018	25	RoB -1 Imp: -2 2 stds of 'doubtful' qual	+ ICC for inter-rater bet. 0.90 to 0.94	Rincon-Diaz M. 2017; Selvam S. 2018	537	RoB 0 Inc: 0 Imp: 0 Ind: 0 2 stds for discriminant and convergent validity	+ + results in line with both hypothesis (discriminant; convergent)	Rincon-Diaz M. 2017; Selvam S. 2018				
ROCF (Rey, 1941)	621	RoB -2 1 std of 'doubtful' qual.	? Did not do CFA or EFA but PCA	Pineda D. 2007					786	RoB -3 1 study of 'inadequate' qual.	? MG CFA performed but not well	Holding P. 2018	786	RoB -3 1 study of 'inadequate' qual.	? ICC not done for inter rater; Even then, Kappa <0.7	Holding P. 2018	621	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 std of 'very good' qual.	+ result in line with hypothesis (discriminative)	Pineda D. 2007	R: 786	R: RoB -2 1 std of 'doubtful' qual.	R: - results not in line with hypothesis	Holding P. 2018
EFS/EFE (Korzeniowski & Ison, 2019)	307	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 std	+ CFI = 0.96, TLI 0.95	Korzeniowski C. 2019	307	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 std	+ Cronb. Alpha = 0.95	Korzeniowski C. 2019					307	RoB -2 1 study of 'doubtful' qual.	? ICC not done, only Pearsons	Korzeniowski C. 2019								
CCTT of DKEFS (Delis et al., 2001)	281	RoB -1 2 stds; 1 'adequate', 1 'inadequate'	+ EFA done and confirm factor model	Ruffieux N. 2010; Roste H. 2018									125	RoB -3 1 study of 'inadequate' qual.	? no ICC or weighted kappa	Ruffieux N. 2010								
Design fluency of DKEFS (Delis et al., 2001)	156	RoB -3 1 study of 'inadequate' qual.	? Did not do EFA or CFA only PCA	Roste H. 2018																				
Verbal Fluency of DKEFS (Delis et al., 2001)	902	RoB -1 3 stds: 1 'adequate', 1 'inadequate'; 1 'doubtful' qual.	+ EFA done and confirm factor mode	Ruffieux 2010; Roste H. 2018; Pineda D. 2007									125	RoB -3 1 study of 'inadequate' qual.	? no ICC or weighted kappa	Ruffieux N. 2010	621	RoB 0 Inc: 0 Imp: 0 1 std of 'very good' qual.	+ result in line with hypothesis (discriminant)	Pineda D. 2007				
KITAP (Zimmermann et al., 2002)									143	RoB -3 1 std of 'inadequate' qual.	? No MG CFA or DIF performed	Sobeh J. 2012					143	RoB -2 1 std of 'doubtful' qual.	++ in line with 2 discriminant hypothesis	Sobeh J. 2012				
JBT (Shallice et al., 2002)	121	RoB -3 1 std of 'inadequate' qual.	? Did not do EFA or CFA	Senturk 2014													121	RoB -1 1 std of 'adequate' qual.	- not in line with convergent hypothesis	Senturk 2014				
RACER (Ford et al., 2019)									2725	RoB -2 1 std of 'doubtful' qual.	? No MG CFA performed	Ford C. 2019												
CPAS (Amini et al., 2016)	700	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study	+ GFI, NNFI, CFI > 0.9, RMSEA bet. 0.03 - 0.08	Amini M. 2017	700	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study	- Cronb. Alpha bet. 0.44 - 0.91	Amini M. 2017					31	RoB -1 Imp: -2 1 stud of 'adequate' qual.	+ ICC between 0.74 - 0.96	Amini M. 2017	122	RoB -1 1 study of 'adequate' qual.	+ in line with hypothesis (convergent)	Amini M. 2017				

Table 8 (continued)

Instrument	Structural validity				Internal consistency				Cross cultural validity				Reliability				Construct validity				Responsiveness/Measurement Error (ME)			
	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall	Ref.	N	meth. qual	Overall	Ref.
WCST (E. A. Berg, 1948)	621	RoB -2 1 std of 'doubtful' qual.	? Did not report results of CFA or EFA	Pineda D. 2007	65	RoB 0 Imp: -1 1 std of 'very good' qual.	+ Cronb. Alph a bet. 0.88 to 0.94	Nampijja M. 2010	64	RoB -3 Imp: -1 1 study of 'inadequate' qual.	? No MG CFA performed	Nampijja M. 2010	64	RoB -2 Imp -1 1 study of 'doubtful' qual.	? ICC or weigthed kappa not done, did Pearsons with no proof of no systematic change	Nampijja M. 2010	1786	RoB 0 Inc: -1 Ind -1 3 stds; 1 'very good'; 2 'doubtful' with one of mixed adult popl.; 75% confirm hypo. so Inc: -1	+++ - 3 of 4 hypothesis of same kind (discriminative) confirmed	Tascu A. 2012; Nampijja M. 2010; Pineda D. 2007				
Design Copying of NEPSY (Korkman, 1998)	185	RoB -3 1 std of 'inadequate' qual.	? Did not do EFA or CFA only PCA	Kashala E. 2005													185	RoB 0 1 std of 'adequate' qual.	- not in line with discriminant hypothesis	Kashala E. 2005				
Tower Test# of NEPSY /DKEFS/computer administered ToH	449	RoB -2 2 stds of 'inadequate' qual.	? Did not do EFA or CFA only PCA or ICC	Kashala E. 2005; Pluck G. 2019	264	RoB -3 1 std of 'inadequate' qual.	- Cronb. Alpha = 0.576	Pluck G. 2019	1311	RoB -2 1 std of 'doubtful' qual	+ partial scalar invariance, RMSEA 0.03, CFI 0.98, TLI 0.98	Xu C. 2020	22	RoB -2 Imp: -2 1 study of 'doubtful' qual.	? ICC or weighted Kappa not done, and even Pearson's bet. -0.01 to 0.7	Pluck G. 2019	185	RoB -1 1 study of 'adequate' qual	- not in line with discrimimant hypothesis	Kashala E. 2005				
Digit Span of NEPSY (Korkman, 1998)	185	RoB -3 1 std of 'inadequate' qual.	? Did not do EFA or CFA only PCA	Kashala E. 2005													185	RoB -1 1 study of 'adequate' qual	- not in line with discrimimant hypothesis	Kashala E. 2005				
Go/No-go (GNG) (Luria, 1973)	306	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study	+ CFI = 1.0, RMSEA = 0.0	Sartori R. 2020	306	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study	+ Cronb. Alpha = 0.80	Sartori R. 2020	306	RoB -1 1 std of 'adequate' qual.	+ invariance for age, gender RMSEA .00; GFI .99; CFI 1.00	Sartori R. 2020	786	RoB -3 1 study of 'inadequate' qual.	? ICC not reported for inter rater. Even then, Kappa <0.7	Holding P. 2018	306	RoB -1 1 study of 'adequate' qual.	++ in line with 2 hypothesis (convergent, discrimimant)	Sartori R. 2020	R: 786	R: RoB -2 1 std of 'doubtful' qual.	R: - not in line with hypothesis (age effect reversed)	Holding P. 2018
Letter numbering sequence (Ruffieux et al., 2010)	125	RoB -1 1 std of 'adequate' qual.	+ EFA performed and confirm m factor model	Ruffieux N. 2010									125	RoB -3 1 std of 'inadequate' qual.	? ICC or weighted kappa not done	Ruffieux N. 2010								
Shift (P Holding et al., 2018)									786	RoB -3 1 study of 'inadequate' qual.	? MG CFA performed but not well	Holding P. 2018	786	RoB -3 1 study of 'inadequate' qual.	? ICC not done for inter rater; Even then, Kappa <0.7	Holding P. 2018					R: 786	R: RoB -2 1 std of 'doubtful' qual.	R: - not in line with hypothesis	Holding P. 2018
EF Touch (Willoughby et al., 2019)																	193	RoB -2 1 study of 'doubtful' qual.	+ in line with discrimimant hypothesis	Willoughby M. 2019				
BASC (Reynolds & Kamphaus, 1992)	1003	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study	+ CFI 0.942, TLI 0.931, RMSEA 0.072	Garcia-Barrera M. 2015	1003	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study	+ Cronb. Alpha bet. 0.77 - 0.87	Garcia-Barrera M. 2015	1003	RoB -1 1 study of 'adequate' qual.	+ invariance achieved across age, gender and ADHD	Garcia-Barrera M. 2015					1003	RoB -2 1 study of 'doubtful' qual.	+ in line with hypothesis (discrimimant for ADHD)	Garcia-Barrera M. 2015				
BSFT (Rosetti et al., 2016)																	106	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual. Looking at 2 hypothesis	+; - in line with discrimimant but not convergence	Rosetti M. 2018				

Table 8 (continued)

Instrument	Structural validity				Internal consistency				Cross cultural validity				Reliability				Construct validity				Responsiveness/Measurement Error (ME)			
	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall	Ref.	N	meth. qual	Overall	Ref.
VABS I & II (Sparrow et al., 1984, 2005)	120	RoB -3 1 study of 'inadequate' qual.	- CFI 0.93-0.97, but RMSEA was 0.113-0.095	Goldberg M. 2009	206	RoB 0 Inc: 0 Imp: 0 Ind: 0 2 stds of 'very good' qual.	+ Cronb. Alpha 0.76 to 0.99	Goldberg 2009; Tombokan- Runtukahu 1992	498	RoB -2 2 stds of 'inadequate' qual.	? No MG CFA performed	Selvam 2016; Tombokan- Runtukahu 1992	498	RoB -2 2 stds: 1 'inadequate' qual.; 1 'doubtful' qual.	+ ICC for test-retest 0.83 to 0.90, for inter-rater 0.93 to 0.97	Selvam 2016; Tombokan- Runtukahu 1992	237	RoB 0 Inc: 0 Imp: 0 Ind: 0 2 stds of 'adequate' qual.	+++ sufficient; all 3 discriminant hypothesis confirmed	Selvam 2016; Tombokan- Runtukahu 1992				
CPQ (Rosenberg et al., 2010)	120	RoB -3 1 std of 'inadequate' qual.	+ CFI was 0.99; RMSEA was 0.04	Amini M. 2017	120	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual.	+ Cronb. Alpha bet. 0.64 - 0.99; most > 0.7	Amini M. 2017					120	RoB -1 1 study of 'adequate' qual.	+ ICC > 0.9	Amini M. 2017	120	RoB -1 1 study of 'adequate' qual.	- results not in line with convergent hypothesis	Amini M. 2017				
Stroop Colour Word Test (Stroop, 1935)													50	RoB -2 Imp: -2 1 study of 'doubtful' qual.	? Pearson for test-retest between 0.64 - 0.83	Malek A. 2013	180	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual	+ in line with hypothesis (discriminative)	Malek A. 2013				
SOPT (Petrides & Milner, 1982)																	248	RoB 0 Ind: -1 2 std: 1 'very good'; 1 'adequate' Ind.-1 because compares EF tool with psychopathology tool	++ results in line with 2 hypothesis (discriminant- and concurrent)	Sallum I. 2017				
PADL (Penny Holding & Kitsao-Wekulo, 2009)	116	RoB -3 1 std of 'inadequate' qual.	? Did not do EFA or CFA only PCA.	Holding P. 2009	116	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual.	+ Cronb. Alpha 0.85 and 0.91	Holding P. 2009									116	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual.	++ in line with 2 hypothesis (2 discriminative)	Holding P. 2009				
Child Function Impairment Rating Scale CFIRS (Tol et al., 2011)	403	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual.	+ NNFI = 0.99, RMSEA = 0.00	Tol W. 2011	403	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual.	+ Cronb. Alpha child = 0.77, parent = 0.74	Tol W. 2011					53	RoB -2 Imp: -1 1 std of 'doubtful' qual.	+ CC "approaching 1.0"	Tol W. 2011	403	RoB -3 Ind: -1 1 std 'inadequate'. Also used comparator tool that doesn't measure same construct for "convergent val."	-;? confirms 1 hypothesis (concurrent); ? used wrong method for discriminant	Tol W. 2011				
BDEFS-CA (Barkley, 2012)	2295	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual	+ CFI=0.98, TLI = 0.96, RMSEA =0.063	Mashhadi A. 2020	2295	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual	+ Cronb. Alpha all > 0.98	Mashhadi A. 2020					90	RoB -2 Imp: -1 1 std of 'doubtful' qual.	? ICC or weighted Kappa not reported	Mashhadi A. 2020	2295	2 hypotheses tested: RoB 0 for convergent o/a 1 std 'very good'; but RoB -2 for discriminant o/a 1 std of 'doubtful'	++ results in line with both hypothesis (convergent; discriminant)	Mashhadi A. 2020				
TEXI (Thorell et al., 2020)	306	RoB -1 1 std of 'adequate' qual.	+ did EFA and showed 2 factor model	Thorell L. 2020	306	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 std 'very good'	+ Cronb. Alpha 0.85 - 0.90	Thorell L. 2020					306	RoB -2 1 std of 'doubtful' qual.	+ ICC = 0.82 for both subscales	Thorell L. 2020	306	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 study of 'very good' qual.	++ in line with 2 hypothesis (discriminant for age, gender)	Thorell L. 2020				
CHEXI (Thorell & Nyberg, 2008)	408	RoB -1 1 std of 'adequate' qual.	- EFA failed to confirm 4 factor model	Trevisan B 2017	408	RoB 0 Inc: 0 Imp: 0 Ind: 0 1 std 'very good'	+ Cronb. Alpha 0.94 to 0.98	Trevisan B 2017									408	RoB -1 Ind: -1 1 std of 'adequate' qual.	+ set out to do 'convergr', but showed 'concurren'	Trevisan B 2017				
ENFEN (Portellano et al., 2009)	367	RoB -3 1 std of 'inadequate' qual.	? Did not do EFA or CFA only PCA.	Rodriguez Barreto L. 2018													367	RoB -1 1 std of 'adequate' qual.	++ in line with 2 hypothesis (convergent, discriminant)	Rodriguez Barreto L. 2018				

Table 8 (continued)

Instrument	Structural validity				Internal consistency				Cross cultural validity				Reliability				Construct validity				Responsiveness/Measurement Error (ME)			
	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall rating	Ref.	N	meth. qual	Overall	Ref.	N	meth. qual	Overall	Ref.
DEX (Wilson et al., 1996)	1586	RoB 0, 1 std	+ CFI 0.95, RMSEA 0.10	Yang Z. 2017																				
IBAS (Munir et al., 1999)					1404	RoB 0, 1 std of 'very good' qual.	+ Cronb. Alpha 0.92 - 0.97	Munir S. 1999					1404	RoB -3, 1 std of 'inadequate' qual.	? no ICC or weighted Kappa, Pearsons when possibility of systematic error	Munir S 1999	1404	RoB -1, 1 std of 'adequate' qual.	++ in line with discriminant- age, clinical disorders	Munir S. 1999				
CVS of TAC (Richard's et al., 2017a)																	41	RoB 0 Imp -2, 1 std of 'very good' qual. But low sample	+ in line with convergence for WISC)	Richard's M. 2017				
TAC	103	RoB -1, 1 std of 'adequate' qual.	+ EFA confirms factor model	Richard's M. 2017																				
AWMA (Alloway, 2007)					210	RoB 0, 1 std of 'very good' qual	+ Cronb. Alpha 0.81 - 0.95	Injoque-Ricle I. 2011									210	RoB -1, 1 std of 'adequate' qual.	+ in line with convergence	Injoque-Ricle I. 2011				
CANTAB (Huppert et al., 1995)									826	RoB -3, 1 std of 'inadequate' qual.	? No MG CFA performed	Green R. 2019					826	RoB 0, 1 std of 'very good' qual	+ in line with discriminative validity for age)	Green R. 2019				
SFIRS (Du et al., 2018)	361	RoB 0, 1 std of 'very good' qual.	+ CFI > 0.999, RMSEA < 0.001	Du Y. 2018	361	RoB 0, 1 std of 'very good' qual.	+ Cronb. Alpha 0.82 to 0.97	Du Y. 2018					91	RoB -3 Imp -1	? ICC or weighted Kappa not reported	Du Y. 2018	421	RoB -1, 1 std of 'adequate' qual.	+ in line with convergence with BRIEF	Du Y. 2018				
EFICA-Parents (Arruda et al., 2020)	1649	RoB 0, 1 std of 'very good' qual.	+ CFI 0.965, TLI 0.963, RMSEA 0.053	Arruda M. 2020	1649	RoB 0, 1 std of 'very good' qual.	+ Cronb. Alpha >0.9	Arruda M. 2020									165	RoB 0, 1 std of 'very good' qual.	+ in line with discriminant for ADHD	Arruda M. 2020				
EFICA-Teachers (Arruda et al., 2020)	1273	RoB 0, 1 std of 'very good' qual.	+ CFI 0.991, TLI 0.991, RMSEA 0.077	Arruda M. 2020	1273	RoB 0, 1 std of 'very good' qual.	+ Cronb. Alpha >0.9	Arruda M. 2020																
BES (Wong et al., 2012)	149	RoB -3, 1 std of 'inadequate' qual.	? Did not do CFA or even EFA	Wong A. 2012	149	RoB -3, 1 std of 'inadequate' qual.	+ Cronb. Alpha 0.8	Wong A. 2012																
PSRA (Smith-Donald et al., 2007)									471	RoB -3, 1 std of 'inadequate' qual.	+ invariance achieved CFI .997, TLI .991, RMSEA .035	Gonen M. 2019	471	RoB -1, 1 std of 'adequate qual.	+ ICC between 0.75 and 0.92	Gonen M. 2019								
Computer version-Spatial span, Figure Matching and Stop Signal Task (form of GNG)									1311	RoB -2, 1 std of 'doubtful' qual.	+ partial scalar invariance achieved, RMSEA 0.03, CFI 0.98, TLI 0.98	Xu C. 2020												

*MIC- Minimal Important Change; **RoB**- Risk of Bias; **Inc**- Inconsistency; **Imp**- Imprecision; **Ind**- Indirectness

#Different Tower Tests appear in various batteries, but essentially the same legacy test. Hence aggregated results of various versions in different batteries and critically appraised as a single test.

KEY FOR QUALITY OF EVIDENCE: 'High' quality of evidence = green

'Moderate' quality of evidence = yellow

'Low' quality of evidence = orange

'Very Low' quality of evidence = red

Lighter shades of the respective colours indicate where the results rating 'sufficient' (+) even though the quality of evidence was low (orange) or 'very low' (red).

BRIEF- Behavior Rating Inventory of Executive Function; **ROCF**- Rey-Osterrieth complex figure; **EFE/EFS**- Executive Function Scale for Children; **CCTT of DKEFS**- Children's Color Trails Test of Delis Kaplan Executive Function System; **KITAP**- Testbatterie zur Aufmerksamkeitsprüfung für Kinder; **JBT**- Junior Brixton Test; **RACER**- Rapid Assessment of Cognitive and Emotional Regulation; **CPAS**- Children's Participation Assessment Scale; **WCST**- Wisconsin Card Sorting Test; **NEPSY**- Developmental NEuroPSYchological Assessment; **BASC**- Behavioural Assessment System for Children; **BSFT**- Ball Search Field Task; **VABS**- Vineland Adaptive Behaviour Scales; **CPQ**- Child Participation Questionnaire; **SOPT**- Self-Ordered Pointing Task; **PADL**- Participation in Activities of Daily Living; **CFIRS**- Child Function Impairment Rating Scale; **BDEFS-CA**- Barkley Deficits in Executive Functioning Scale–Children and Adolescents; **TEXI**- Teenage Executive Functioning Inventory; **CHEXI**- Children's Executive Function Inventory; **ENFEN**- (Spanish acronym) Neuropsychological Assessment of Executive Functions; **IBAS**- Independent Behaviour Assessment Scale; **DEX**- Dysexecutive Questionnaire; **CVS of TAC**- Conjunction Visual Search task of tareas de autorregulaCión Cognitiva battery; **AWMA**- The Automated Working Memory Assessment; **CANTAB**- The Cambridge Neuropsychological Test Automated. Battery; **SFIRS**- Symptoms and Functional Impairment Rating Scale; **EFICA**- Executive Function Inventory for Children and Adolescents; **BES**- Ballet Executive Scale; **PSRA**- Preschool Self-Regulation Assessment

Table 9 Summary of Results of Critical Appraisal of Content Validation/Adaptation of EF and AF tools Validated for use Among Children in LAMICs

Instrument	Instrument Development Study			Content Validity/ adaptation study 1			Content Validity/ adaptation study 2			Content Validity/ adaptation study 3			OVERALL RATING FOR INSTR.	COMMENT*
	Meth. Qual. (RoB)	Results Rating	Ref	Meth. Qual	Rating	Ref.	Meth. Qual. (RoB)	Rating	Ref.	Meth. Qual. RoB	Rating	Ref.		
CPAS (Amini et al., 2016)	Inadequate.	? o/a relevance, comprehensiveness and comprehensibility all rated ?	Amini. M. 2016										? RoB -3 downgrade o/a only 1 dev. Study of 'inadequate' qual. Hence 'VERY LOW' qual. Of evidence	'inadequate' RoB rating o/a no formal qualitative analysis done; final set tested in only 2 subjects and even at that a quantitative survey done rather than qualitative
PACS (C. Berg & LaVesser, 2006)	Doubtful	± o/a Relevance rated +, comprehensiveness +; but comprehensibility ?	Berg 2006	Inadequate	? relevance and comprehensibility rated as ?; comprehensiveness +	Malkawi S. 2015							± RoB -2 downgrade to LOW o/a 1 content validity std. of 'inadequate' quality and 1 Instrument dev. study of doubtful quality	Adaptation study had professional relevance box rated inadequate o/a convened expert panel did not comment on relevance, also items tested on low number of experts; also final version unlikely to have been tested for comprehensibility or comprehensiveness
EFS/EFE (Korzeniowski & Ison, 2019)	inadequate	? o/a 2 parts- comprehensibility and comprehensive ?; but relevance ±	Korzeniowski C. 2019										? RoB -3 downgrade o/a only 1 dev. Study of 'inadequate' qual. Hence 'VERY LOW' qual. Of evidence	'inadequate' RoB rating o/a doesn't appear concept elicitation was done at all. Only "prepared items adhoc based on EF theory" and tested with expert panel.
ROCF (Rey, 1941)	Could not access development study o/a a legacy instrument (test developed in 1940's) and also article in French.		(Rey, 1941)	Inadequate	± o/a relevance rated ±; comprehensiveness - ; and comprehensibility ?.	(Penny Holding et al., 2018)							± Since no instrument dev. study available, used only the adaptation study. Per this, RoB -2 downgrade o/a only 1 adaptation study of 'inadequate' quality. Since estimated best case scenario rating of the dev. Study (per COSMIN) was estimated as 'doubtful' (see comment), we graded as 'LOW' at best	'inadequate' RoB of adaptation study o/a patients/caregivers not involved in determining relevance of task to their experience, nor directly asked about comprehensibility of instructions; also piloting study not described at all. Also, dev. Study not evaluated, but given the legacy nature of the ROCF (test developed in 1940's), highly unlikely to have met COSMIN criteria with a rating of 'very good' or 'adequate' (see supplemental material for details). Thus probably best case rating would have been 'doubtful'.
Go/No-Go (GNG) (Luria, 1973)	Could not access development study o/a a legacy instrument (test developed in 1970's). Appeared to have developed this version themselves, but didn't say how		(Luria, 1973)	Inadequate	? o/a relevance and comprehensibility rate ?, while comprehensiveness rated -	(Penny Holding et al., 2018)	Inadequate	± o/a relevance ?, comprehensiveness -, comprehensibility +	Nampijja 2010	Inadequate	± o/a relevance ?, comprehensiveness -, comprehensibility +	Sartori R. 2020	± Since no instrument dev. study available, used only the 3 adaptation studies. Per this, RoB -2 downgrade o/a 3 adaptation studies of 'inadequate' quality. Since estimated best case scenario rating of the dev. Study (per COSMIN) was estimated as 'doubtful' (see comment).), we graded as 'LOW' at best	'inadequate' RoB of adaptation study o/a patients/caregivers not involved in determining relevance of task to their experience, nor directly asked about comprehensibility of instructions; also piloting study not described at all. Also, dev. Study not evaluated, but given the legacy nature of the Go/No-go (test developed in 1970's), highly unlikely to have met COSMIN criteria with a rating of 'very good' or 'adequate' (see supplemental material for details). Thus probably best case rating would have been 'doubtful'.
Shift (P Holding et al., 2018)	No Instrument dev. Study bcos authors appeared to have developed this themselves, but didn't say how.			Inadequate	? o/a relevance and comprehensibility rate ?, while comprehensiveness rated -	(Penny Holding et al., 2018)							? Since no instrument dev. study available, used only the adaptation study. Per this, RoB -3 downgrade o/a only 1 adaptation study of 'inadequate' quality. Since estimated best case scenario rating of the dev. Study (per COSMIN) was estimated as 'inadequate' (given the 'inadequate' rating of their methodology for their adaptation study), we graded as 'VERY LOW' at best	'inadequate' RoB of adaptation study o/a patients/caregivers not involved in determining relevance of task to their experience, nor directly asked about comprehensibility of instructions; also piloting study not described at all. Because this appears to be a newly developed test by authors, and given the inadequate rating of the adaptation study, its highly unlikely that the original dev. study of this test would have been anything other than 'inadequate'. Hence estimating instrument dev. rating as 'inadequate'.
VABS (Sparrow et al., 1984, 2005)	'inadequate'	? o/a comprehensiveness and comprehensibility ?; and relevance ±	Sparrow 1984, 2016	Inadequate	+ o/a relevance and comprehensibility +, but comprehensiveness ?	(Tombokan-Runtukahu & Nitko, 1992)	inadequate	? o/a all 3 rated as ?	Goldberg M. 2009				+ based on best result rating rule. Thus, RoB -3 o/a 2 adaptation studies both of 'inadequate' quality, and 1 instrument development study of 'inadequate' quality.	'inadequate' RoB of development study o/a parents not asked about relevance or comprehensiveness. Same for 1st adaptation study where neither parents nor professionals were asked about comprehensiveness. In Goldberg study, no cognitive interviewing was done at all. Also in both studies, not much detail given about pilot study
PADL (Penny Holding & Kitsao-Wekulo, 2009)	Inadequate	? o/a all 3 rated ?	(Penny Holding & Kitsao-Wekulo, 2009)										? RoB -3 downgrade o/a only 1 dev. Study of 'inadequate' qual. Hence 'VERY LOW' qual. Of evidence	inadequate' RoB rating o/a no formal coding of data done; comprehensibility of items or instructions, and comprehensiveness not assessed
CFIRS (Tol et al., 2011)	inadequate	? o/a all 3 rated ?	(Tol et al., 2011)										? RoB -3 downgrade o/a only 1 dev. Study of 'inadequate' qual. Hence 'VERY LOW' qual. Of evidence	inadequate' RoB rating o/a no formal coding of data done; no clear definition of construct to be measured, no details of piloting to assess methodological approach
BRIEF (parent) (Gioia et al., 2000)	Inadequate	? o/a comprehensive and comrehensible ?, relevance +	Gioia G. 1996, 2015	Inadequate	? o/a all 3 rated ?	(Chernoff et al., 2018)							? RoB -3 o/a 1 adaptation study of 'inadequate' quality, and 1 instrument dev. study of 'inadequate' quality. Hence VERY LOW'	inadequate' RoB of adaptation study o/a - no cognitive interviewing was done at all. For the development study o/a patients/caregivers not asked about comprehensibility (although teachers were), or comprehensiveness. Also not enough detail on pilot study to assess methodology properly.

Table 9 (continued)

BRIEF (pre-school) (Isquith et al., 2004)	Inadequate	? o/a comprehensive and comrehensible?, relevance +	Gioia G. 1996, 2015	Inadeq uate	+ o/a comprehensiven ess and comprehensibility +; relevance ±	(Diaz & Anac ona, 2017)							*- based on best result rating rule. Thus RoB -3 o/a 1 adaptation study of 'inadequate' quality, and 1 instrument development study of 'inadequate' quality. Hence 'VERY LOW' Qual. Of evidence	inadequate' RoB of adaptation study o/a – relevance not asked about among parents, only comprehensibility. For the dev. study 'inadequate' rating o/a patients/caregivers not asked about comprehensibility, or comprehensiveness. Also, not enough detail on pilot study (for both studies) to assess methodology properly.
WCST (E. A. Berg, 1948)	Could not access development study o/a legacy instrument (test developed in 1940's).		(E. A. Berg, 1948)	Inadeq uate	± o/a relevance ?, comprehensiven ess -, comprehensibility +	(Na mpijj a et al., 2010)							± Since no instrument dev. study available, used only the adaptation study. Per this, RoB -2 downgrade o/a only 1 adaptation study of 'inadequate' quality. Since estimated rating of the dev. Study (per COSMIN) based on best case scenario was 'doubtful' (see comment), we graded as 'LOW' at best	'inadequate' RoB of adaptation study o/a patients/ caregivers not involved in determining relevance of task to their experience; comprehensiveness not really assessed at all Also, dev. Study not evaluated, but given the legacy nature of the WCST (test developed in 1940's), highly unlikely to have met COSMIN criteria with a rating of 'very good' or 'adequate' (see supplemental material for details) . Thus probable best case rating would have been 'doubtful'.
TEXI (Thorell et al., 2020)	Inadequate	? o/a all 3 rated ?	(Tho rell & Nybe rg, 2008 b)	Inadeq uate	? o/a relevance and comprehensiven ess ?, comprehensibility +	(Tho rell et al., 2020)							? RoB by -3 o/a 1 adaptation study of 'inadequate' quality and 1 instrument dev. study of 'inadequate' quality	'inadequate' RoB of adaptation study o/a target population (adolescents) not asked about relevance or comprehensiveness; also not enough detail on pilot study (for both studies) to assess methodology properly. For the dev. study 'inadequate' rating o/a used inappropriate method of data analysis at concept elicitation phase and failed to mention any formal coding of data. Also no cognitive interviewing done at piloting stage.
CHEXI (Thorell & Nyberg, 2008)	Inadequate	? o/a all 3 rated ?	(Tho rell & Nybe rg, 2008 b)	Inadeq uate	? o/a all 3 rated?	(Trev isan et al., 2017)							? RoB by -3 o/a 1 adaptation study of 'inadequate' quality and 1 instrument dev. study of 'inadequate' quality	'inadequate' RoB of adaptation study o/a no cognitive interviewing was done at all. For the dev. Study, see above for reasons for 'inadequate' rating .
Child Hayling Test (P. W. Burgess & Shallice, 1997)	Could not access Instrument dev. study because the development study is not freely available online. And Manual with dev. study had to be purchased.		(P. Burg ess & Shalli ce, 1997)	Inadeq uate	± o/a relevance ±, comprehensiven ess -, comprehensibility +	(Siqu eira et al., 2016)							± Since no instrument dev. study available, used only the adaptation study. Per this, RoB -2 downgrade o/a only 1 adaptation study of 'inadequate' quality. Since estimated rating of the dev. Study (per COSMIN) based on best case scenario was 'doubtful' (see comment), we graded as 'LOW' at best	'inadequate' RoB of adaptation study o/a patients/ caregivers not involved in determining relevance of task to their experience; comprehensiveness not really assessed at all Also, dev. Study not evaluated, but given the fact that on the comprehensiveness criteria alone, CHT only assesses 2 domains of EF (Inhibition and Cognitive flexibility), CHT is therefore unlikely to score anything higher than doubtful in the comprehensiveness box per COSMIN criteria. Hence grading as 'low'.
AWMA (Alloway, 2007)	Could not access Instrument dev. study because the development study is not freely available online. And Manual with dev. study had to be purchased.		(Allo way, 2007)	Inadeq uate	± o/a relevance ?, comprehensiven ess -, comprehensibility +	(Injo que- Ricle et al., 2011)							± Since no instrument dev. study available, used only the adaptation study. Per this, RoB -2 downgrade o/a only 1 adaptation study of 'inadequate' quality. Since estimated rating of the dev. Study (per COSMIN) based on best case scenario was 'doubtful' (see comment), we graded as 'LOW' at best	'inadequate' RoB of adaptation study o/a patients/ caregivers not involved in determining relevance of task to their experience; comprehensiveness not really assessed at all; also used survey method in piloting Also, dev. Study not evaluated, but given the fact that on the comprehensiveness criteria alone, AWMA only assesses 1 domains of EF (Working Memory), AWMA is therefore unlikely to score anything higher than doubtful in the comprehensiveness box per COSMIN criteria. Hence grading as 'low'.
SFIRS (Du et al., 2018)	Inadequate	? o/a all 3 rated ?	Du Y. 2018										* RoB -3 downgrade o/a only 1 dev. Study of 'inadequate' qual. Hence 'VERY LOW' qual. Of evidence	'inadequate' RoB rating o/a patients/ caregivers not involved in determining comprehensiveness of the items for their child's functional experience living with ADHD; also methodological shortfalls such as only written records with no recordings and verbatim transcription, no mention of data coding method, Didn't say HOW they arrived at the item pool following the interviews etc.
EFICA (Arruda et al., 2020)	Inadequate	? o/a comprehensiveness and comprehensibility ?, relevance ±	(Arr uda et al., 202 0)										? RoB -3 downgrade o/a only 1 dev. Study of 'inadequate' qual. Hence 'VERY LOW' qual. Of evidence	'inadequate' RoB rating o/a target group not consulted in concept elicitation and item generation, final version of (corrected) tool likely not re-tested for comprehensibility before validation, doubtful that COMPREHENSIVENESS was assessed in FGD of target group

*only highlights of reasons for RoB rating discussed here. For detailed explanation refer to supplemental material 'dev. Study' = Development study
"ref" = refereemce; 'meth. Qual.' = methodology quality
KEY FOR QUALITY OF EVIDENCE: 'High' quality of evidence = green
'Moderate' quality of evidence = yellow
'Low' quality of evidence = orange
'Very Low' quality of evidence = red
Lighter shades of the respective colours indicate where the results rating was 'sufficient' (+) even though the quality of evidence was low (orange) or 'very low' (red).

PACS- Preschool Activity Card Sort; **CPAS**- Children's Participation Assessment Scale; **EFE/EFS**- Executive Function Scale for Children; **ROCF**- Rey-Osterrieth complex figure; **VABS**- Vineland Adaptive Behaviour Scales; **PADL**- Participation in Activities of Daily Living; **CFIRS**- Child Function Impairment Rating Scales; **BRIEF**- Behavior Rating Inventory of Executive Function; **WCST**- Wisconsin Card Sorting Test; **TEXI**- Teenage Executive Functioning Inventory; **CHEXI**- Children's Executive Function Inventory; **AWMA**- The Automated Working Memory Assessment; **SFIRS**- Symptoms and Functional Impairment Rating Scales; **EFICA**- Execution Function Inventory for Children and Adolescent

Fig. 3 Summary of Results Ratings for Measurement Properties
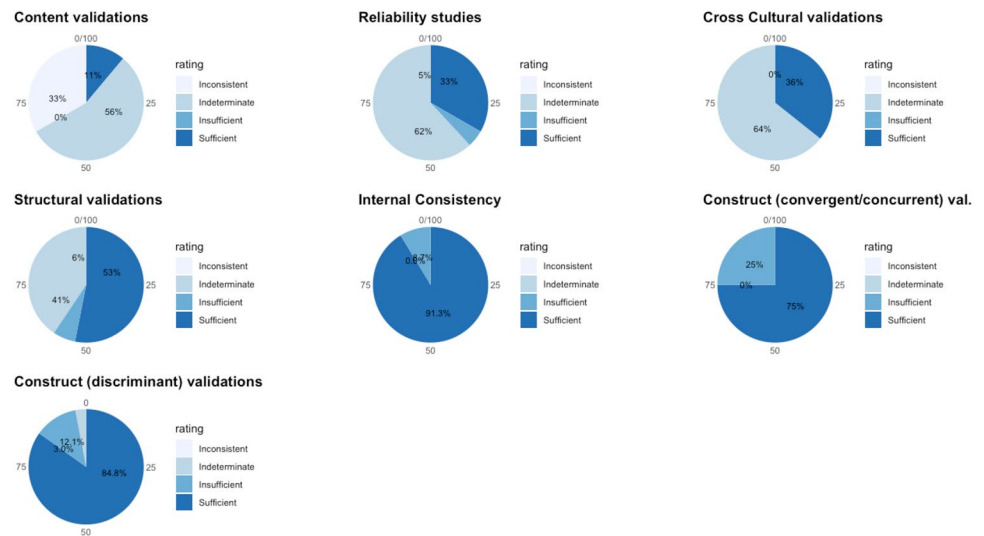
tools developed for the countries of interest) or adaptation studies reporting on their content validity in those countries, which were critically appraised. None of these tools showed full content validity (i.e., 'sufficient' rating of high or moderate quality evidence in all three sub-domains of RELEVANCE, COMPREHENSIBILITY and COMPREHENSIVENESS) in the LAMICs considered. Of these, only the BRIEF (pre-school version) (Isquith et al., 2004) and the VABS (Sparrow et al., 1984, 2005) had studies showing 'sufficient' content validity but of very low quality of evidence, with all the others showing either 'indeterminate' (?) or 'inconsistent' (±) content validity results of either low or very low quality of evidence. For several of

the legacy tests such as the Wisconsin Card Sorting Test (WCST) (Berg, 1948), Go/No-go (GNG) (Luria, 1973) and the Rey-Osterrieth complex figure (ROCF) (Rey, 1941), the original development studies could not be obtained to be critically appraised. Therefore we estimated their probable score based on the best case scenario of their likelihood of fulfilling (or failing to fulfil) the COSMIN criteria for development studies, which was used in estimating their overall content validation score (while factoring in the score of the adaptation studies that were eligible as described in *Synthesis of Results* section of Methodology, and in COSMIN guideline pages 62–63), with reasons provided for their estimated score.

Fig. 4 Summary of Quality of Evidence (GRADE) Ratings for Measurement Properties
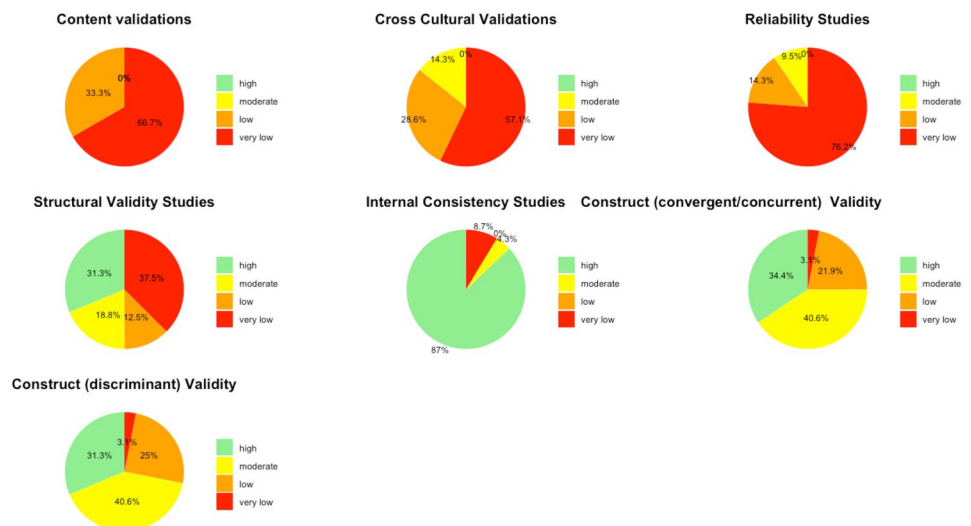
Discussion

This systematic review was carried out to critically appraise the validation and adaptation studies that had been conducted for use of executive and adaptive functioning tools among children and adolescents in LAMICs, to make recommendations for clinical use and for further research in such cross-cultural contexts.

Summary of the evidence: The General State of the Evidence Available.

The first noteworthy observation is that there is an obvious lack of validation studies for executive functioning (EF) and adaptive functioning (AF) instruments in LAMICs, period. The many 'potholes' in the summary tables of Tables 8 and 9 is ample evidence of this lack of validity studies. Further, even where the evidence exists, the quality of evidence leaves much to be desired. Particularly concerning were the quality of evidence for content validation, cross-cultural validation, reliability and (to a lesser extent) structural validation studies, where most of these types of studies were of either 'low' or 'very low' quality as shown in Figs. 3 and 4.

For cross-cultural validation studies, only 5 out of 14 (35%) showed 'sufficient' cross cultural validity (see Fig. 3), and even then, the quality of evidence for 12 out of 14 (86%) was 'low' or 'very low' (see Fig. 4). One of the major methodological problems that led to this poor result was that a vast majority of papers purporting to do a cross-cultural validation failed to use an appropriate approach to compare the two groups (comparing original culture to target cultural context) such as Multi-group Confirmatory Factor Analysis (MGCFA) with most opting for a Principal Component Analysis (PCA), or even ANOVAs and *student t test* approaches. In order to adequately ascertain measurement invariance of a tool in one cultural context as compared to its use in its original context, one must assess how each item in the tool functions in the target population- whether the items achieved scalar invariance, metric invariance and configural invariance- to be able to confidently conclude that the tool functions in the same way in the target population as it does in the original population in which it was developed (Fischer & Karl, 2019). Only an appropriate method such as MGCFA can provide this information. A PCA for example, based on its underlying assumption of only common variance and no unique variance among items, will only be useful as a data reduction tool that can essentially "boil down" a complex set of variables to an essential set of composite variables (i.e. reduce the number of necessary items in a tool), and not to uncover the latent constructs underlying the variables or to ascertain how items behave in another population. Therefore, use of PCA for cross-cultural validation is considered an inappropriate methodology, and was the major reason for

most of the cross-cultural validation studies being rated so poorly, as per COSMIN criteria.

Reliability studies also had a poor showing. Only 6 out of 21 (29%) showed 'sufficient' reliability, and even then, the quality of evidence for 19 out of 21 (90%) was 'low' or 'very low'. The reason for this poor showing for many of the studies was either time interval between assessments not stated, or too long an interval (whether for test–retest or inter-rater reliability), where many studies went well beyond the recommended (by COSMIN) 2 week-period (some studies as long as 48 – 96 weeks between assessments), taking too long a time interval to be sure that the underlying construct (e.g. EF) had not actually changed in reality, especially considering that EF does change with age. A second reason was an inappropriate method of data analysis, where some papers used a Pearson's correlation to calculate reliability without any evidence provided that no systematic change had occurred (since Pearson's does not take systematic error into account), rather than using the more appropriate Intra Class Coefficient (ICC) (which does take systematic errors into account) and describing the model of ICC used.

For content validity, only 2 out of 16 studies (13%) had 'sufficient' content validity with all of them showing either 'low' or 'very low' quality of evidence. The commonest reasons for this were that many adaptation or instrument development studies did not involve the target audience in either concept elicitation (for instrument development studies) or the determination of relevance, comprehensiveness or comprehensibility (for adaptation studies), with most relying heavily on "expert panel" input for these things. Given the modern sweep towards "co-production" methods, this "expert dominated" synthesis cannot continue to be encouraged. Also, most did not report on enough detail of the pilot studies to allow for adequate assessment of methodology, hence the many 'indeterminate' ratings of content validity (according to COSMIN, where not enough information is provided to make a judgement one way or another whether evidence is 'sufficient' or not, it is recommended to make a rating of 'indeterminate'). Further, in a few cases where either legacy instruments (such as WCST and ROCF) or some proprietary instruments were involved, it was not always possible to obtain the original development studies if they were not freely available online (some proprietary instruments only published their development study in their instrument manuals which need to be purchases along with the tools), and therefore an estimation rating had to be made for the development tools based on information that was publicly available, assuming a best-case scenario each time.

Having noted the above, the best performing tools for use among children in LAMICs (judging by quality of evidence for psychometric properties) were BRIEF for executive function (EF) and VABS for adaptive function (AF). For EFs, the

main weakness of the close contenders to the BRIEF- the legacy instruments/tasks of WCST, GNG and ROCF- was their lack of comprehensiveness in assessing ALL domains of EF as currently widely accepted. The BRIEF (in all its various versions) showed a respectable coverage of almost all the domains of validity that were assessed with positive results in a good number of them: there was high quality evidence for 'sufficient' internal consistency (seen in 7 studies across all versions of BRIEFs); 'sufficient' construct validity (mostly convergent and some discriminant validity) whose quality of evidence ranged from low to high (seen in 5 studies across 3 out of 4 versions of BRIEFs); and moderate to low quality evidence for 'sufficient' construct validity (mostly convergent validity) for BRIEF-parents and BRIEF-teachers versions respectively. However evidence for good content validity, structural validity and cross-cultural validity of the BRIEFs among children in LAMIC was lacking: there was very low quality evidence for one 'indeterminate' rating of content validity (parent version) and one 'sufficient' content validity (pre-school version) seen in 2 studies; low to very low quality evidence for mostly 'indeterminate' structural validity seen in 6 studies across all versions of BRIEF (except for one study showing low quality evidence of 'sufficient' structural validity of BRIEF-teacher version); and low quality evidence for 'indeterminate' cross-cultural validity (seen in 2 studies but only for 2 versions- parent and pre-school).

Having noted the above, an honourable mention is the TEXI/CHEXI tool, which although a relatively new tool with fewer studies conducted in LAMICs showed remarkably good results with decent quality of evidence: high quality evidence for 'sufficient' internal consistency (seen in 2 studies across all versions), low to high quality evidence for 'sufficient' construct validity (seen in 2 studies across all versions), moderate quality evidence for 'sufficient' structural validity (2 studies across all versions) and low quality evidence for 'sufficient' reliability (seen in 1 study for TEXI). The EFICA tool- developed in Brazil- was also noteworthy because although there was only 1 study evaluating it, that single study produced high quality evidence of sufficient structural validity, internal consistency and (for the parent version only) construct validity (discriminant validity) across its 2 versions- parent and teachers. Both of these other tools though did not have good evidence for content validity in their reported studies. Glaringly missing in all EF instruments though was any tool that explicitly assesses the domain of meta-cognition, but this is probably due to the continuing evolution and refinement of the concept even in Western thought.

For adaptive functioning (AF), VABS was the most validated as earlier mentioned. The main weakness of the closest contender, the CFIRS, was the fact that outside of its home country of Indonesia it had not been validated in a cross-cultural setting, which is understandable given its relatively newer status. As has been noted elsewhere in the literature, adaptive behaviour and expectations of daily living skills and socialization are culturally dependent, hence local input into the appropriate types of questions to be asked in adaptive behavior assessment is crucial (Semrud-Clikeman et al., 2017). The VABS (in all its editions) also showed a decent coverage of almost all domains of validity evaluated with positive results: high quality evidence of sufficient internal consistency (seen in 2 studies) and construct validity (backed by 2 studies); low quality evidence for 'sufficient' reliability (backed by 2 studies); and very low-quality evidence for sufficient content validity (backed by an 'inadequate' adaptation study and an 'inadequate' development study). However, results for cross-cultural validity and structural validity were not the best: there was low quality evidence for an 'indeterminate' rating of cross culturally validity (2 studies); and very low-quality evidence of insufficient structural validity (backed by 1 study).

Evidence from High-Income Countries

Evidence from HICs indicate that assessments of EF and adaptive behavior form a very significant proportion of all assessments performed by clinical psychologists and neuropsychologists with approximately 50% of clinical psychologists practicing in this area (Camara et al., 2000) which indicates its importance. In this nationally representative survey of 1002 neuropsychologists and 1500 clinical psychologists, the most frequently used EF instruments were such legacy tests as Trail making test (TMT) and WCST while that for adaptive behaviour was the VABS (Camara et al., 2000), which was confirmed by a broader-based and more recent follow up survey among 2004 north American neuropsychologists (Rabin et al., 2005), with the ROCF and NEPSY also featuring more prominently in this survey. To provide some context though, it is worth noting that these large surveys are rather dated as the data were collected over 20 years ago, hence would not have included many of the instruments included in this review which were developed in the last 20 years.

Interestingly, in a similar recently published systematic review of EF performance-based instruments used within the context of Occupational therapy (which incidentally also used the COSMIN manual for assessing risk of bias, and which was not limited to LMICs), the following instruments, none of which have been adapted for use in LAMICs hence failed to show up in our review, were all found to have a "low" GRADE rating of quality of evidence: the Behavioural Assessment of the Dysexecutive Syndrome for Children, Children's Cooking Task, Children's Kitchen Task Assessment, Do-Eat, and Preschool Executive Task Assessment (Gomez et al., 2021). Another such systematic review

performed by researchers from Italy (also using COSMIN, and also not limited to LAMICs, but considering a much limited array of databases compared to this present study) also identified 19 EF tools which scored poorly on their methodological rigour, but also found the BRIEF (2 of its variants) to have the best risk of bias rating, noting that it particularly did well in internal consistency, but reported inadequate indices for reliability and other measurement properties (Berardi et al., 2021). This study also noted the relatively good performance of another test not included in our review (for lack of adaptation in LMIC)- the Lion Game (Van de Weijer-Bergsma et al., 2014) developed for study of visuo-spatial working memory in Dutch. Finally another paper reviewing EF tools used in Brazil was also found (Guerra et al., 2020) however no critical appraisal was performed in this paper. These findings from the wider literature therefore largely confirm the findings in this systematic review.

Cross Cultural Considerations

Since one of the desired outcomes of this paper is to be able to make recommendations of assessment tools for use in LAMICs, an important consideration for the authors was which of these instruments had adequate content validity and cross-cultural validity in LAMICs. The importance of adequate cultural adaptation or development of instruments in accurately measuring such psychological constructs cannot be overstated. This very point was made by Geisinger in his seminal paper on cross-cultural normative assessment (Geisinger, 1994) when he highlighted the importance of going beyond simple language translation and considering cultural backgrounds during adaptation of instruments among sub-populations in the US. Concept formation is goal-directed. Which means, if the socio-cultural milieu within which a child is placed does not place certain types of goals before him/her, he/she will not form concepts to solve the problems standing as obstacles to that goal, a point echoed by scholars across the decades from Vygotsky first in the 1930's (Vygotsky, 1980) through to Ardila (1995), and Nell (1999) more recently while discussing the work of Luria (1973). For a word to have any meaningful meaning, it must represent a concept that is widely shared with the speakers to whom the communication is directed. Or put in the words of Vygotsky: "This is why certain thoughts cannot be communicated to children even if they are familiar with the necessary words" (Vygotsky, 1980). Any task or set of questions placed before this child (or caregiver) that has not been rigorously determined to be within the conceptual understanding of said child, therefore, is bound to fail in assessing what it purports to assess without a proper rigorous cultural (and conceptual) adaptation. This type of "cultural specificity" of conceptualization has been well documented in the literature in studies from across the world, for example by

Venter in the context of the acquisition of Piagetian tasks in cognitive development (Venter, 2000).

Unsurprisingly, the EF instruments that performed best in cross-cultural settings were legacy performance-based tools such as the GNG (moderate quality evidence of sufficient cross-cultural validity) and Tower Test (low quality evidence of sufficient cross-cultural validity), with the one interview-based instrument that did reasonably well being the BASC (moderate quality evidence of sufficient cross-cultural validity). Of these, the only instrument for which a formal adaptation was reported was the GNG, which showed low quality evidence for 'inconsistent' rating of content validity. Further, it is reported elsewhere in the literature from LAMICs that a version of the GNG (the knock tap test) should be used with extreme caution because of poor reliability (Semrud-Clikeman et al., 2017).

An important implication of these finding is that potentially performance-based tasks/tools are perhaps most easily translatable across different cross-cultural contexts to produce consistent results across different population groups. An obvious shortcoming of this though is the issue of most of these existing (legacy) tasks lacking comprehensiveness in terms of our current understanding of the various domains of EF. A potential solution to this dilemma might be the development of more comprehensive, culturally appropriate performance-based measures of EF which tap into almost all domains of EF. These could either be universally adaptable age-appropriate tasks- such as cooking or running errands- or more regionally peculiar tasks- such as fetching water, playing of particular local games or occupational activities like farming- if they could potentially be calibrated to assess all domains of EF and AF as we currently understand them.

Tool Recommendations

One of the important goals of this paper was to be able to make evidence-backed recommendations for appropriate tools for EF and AF among children and adolescents in LAMICs. To fulfil this goal, one potential approach could have been to make recommendations for the best tools for each domain of EF and AF and discuss reasons why accordingly. However, apart from how bulky this approach would have made this paper, again we were guided by the insightful observation of Vygotsky on the matter of assessing only the constituent parts of a complex whole thus (Vygotsky, 1980): rather than assessing EF and AF only in terms of their constituent domains (as some EF tools for example appear to do), we chose to treat EF and AF as complex wholes because of their intricate inter-linkages and their conceptual unity in our current modern understanding of them.

According to the COSMIN guidelines (as well as the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2018)), content validation is the most

important validation as every other psychometric property of the instrument in question hinges on the items/content making sense in the population to be measured (with respect to the construct of interest- as understood by the population) (Terwee et al., 2018a, b). That notwithstanding, a word of caution from the manual about content validation is in order: "The content validity of an [instrument] may be different when using the [instrument] for measuring different constructs, in different populations, or in different contexts of use. Researchers do not validate an instrument but rather the **application** of it" (Terwee et al., 2018a, b) (emphasis ours). Considering all this and considering their overall performance in the other domains and measurement properties of validity, the BRIEF and VABS would again merit some measured recommendation for use among children in LAMICs as they both have studies showing 'sufficient' content validity, even though of very low-quality evidence. All other instruments evaluated showed either 'inconsistent' or 'indeterminate' content validity of either 'low' or 'very low' quality. Further, all of those others showed slightly better quality of evidence ('low' as opposed to 'very low') only because their instrument development studies could not be obtained and hence a best-case-scenario estimation of their development study quality of evidence (i.e., assuming the best, given available evidence) was a 'doubtful' overall rating, giving them 'low' (as opposed to 'very low') quality of evidence.

Having said that, one of the issues with the BRIEF in Western countries is the lack of positive correlation with direct measures of executive functioning (Semrud-Clikeman et al., 2017). The implications of this is that the BRIEF may reflect a child's inability to perform these skills in day-to-day life even though individual measures of executive functioning may show that the child can perform the task in a structured, one-to-one setting (Semrud-Clikeman et al., 2017). But in its defense, this problem of a disconnect between legacy "experimental" tests of EF and the "real life situations" at home is what prompted the developers of the BRIEF to come out with their "ecologically valid" assessment instrument in the first place (Gioia et al., 2000).

Other reports in the literature largely support our findings in this review. In the scoping review paper for NDD assessment tools for children in LAMICs, among the EF tools they reported as having been used in LAMICs were the NEPSY (in South Africa, Zambia, Romania and Iran) where they found that strong language skills are needed for success, even for English-speaking subjects; the Behaviour Rating Inventory of Executive Function (BRIEF) in Uganda, and the Wisconsin Card Sorting Test (WCST) in Uganda where they reported poor reliability due to the nature of the test (Semrud-Clikeman et al., 2017). For adaptive functioning, this paper reported that the Vineland Adaptive Behavior Scales (VABS) (Sparrow et al., 2005) as the commonest reported in LAMICs, which has been used successfully in

South Africa with families with HIV (Allen et al., 2014). Semrud-Clikeman and colleagues did go ahead to finally recommend selected subtests of the NEPSY-II for testing executive function, and the VABS and the IBAS for testing adaptive behavior (with limitations) which although not backed by a rigorous critical appraisal of the evidence, nonetheless largely supports our recommendations.

Future Directions

One may wonder though, after such an apparently less-than-ideal review, if there is hope for a more universally applicable assessment of executive and adaptive functions. To this, we offer the well-known aphorism: the absence of evidence is not the evidence of absence; that is to say, more evidence needs to be collected before utterly rejecting many of these tools. To be fair to most of these tests, the COSMIN criteria being a recent development is rather biased in favor of tools developed along the contemporary principles of co-production and patient/stakeholder involvement in research design and implementation, which is a relatively recent phenomenon. Therefore, for several of these legacy instruments (being developed 20 or more years ago when such principles were not yet contemporary), it is not too surprising that the COSMIN criteria penalized them for this. However, if LAMICs are to keep abreast with current best practices, it is imperative that researchers and clinicians are made aware of how available existing tools hold up against these best practices, and the standards that the scientific community will expect in adapting or developing tools for local use.

Content validation, being the most important measurement property of an instrument upon which all other properties are hinged (Terwee et al., 2018a, b), bears singling out for comment. BRIEF and VABS showed the most promise for use in LAMICs, but can have better content validation studies and cross-cultural studies undertaken. For example, in the adaptation study of the BRIEF (parents) by Chernoff et al. (Chernoff et al., 2018), no cognitive interviewing (Jobe & Mingay, 1989) was done at all which is a fatal flaw in ensuring content validity, while in that of the BRIEF (preschool) by Diaz et al. (Diaz & Anacona, 2017) a form of cognitive interviewing was done but parents were apparently not specifically asked about the relevance of items in their sociocultural experience. This aspect of adaptation of western tools for cross-cultural use needs to be improved by researchers, with more involvement of parents, adolescents, and teachers in a cognitive interviewing approach to assess relevance and comprehensibility of items and instructions (and due modification where appropriate) going forward.

Finally, researchers in LMICs might want to consider more performance-based task assessment tools that are ecologically based (such as the Children's Cooking Task (Chevignard et al., 2010)) and are culturally appropriate

for future adaptation. As was noted above under the "cross cultural considerations" sub-section, the best performing assessment tools in cross cultural contexts were performance-based tools such as the GNG and Tower Test. Some of these tasks (with some minor cultural fine-tuning of course) might find greater universal applicability (simple cooking tasks for example are likely to find applicability in almost all cultures around the world) and might represent a potentially more universal basis of comparison across various cultures than the various questionnaire-based assessment tools. They are also likely to better solve the problems of subjectivity in interpretation of items by respondents and the need for literacy (a particularly acute problem in LAMICs) in the use of questionnaire-based assessment instruments.

Strengths and Limitations of this Study

This systematic review was conducted according to the rigorous guidelines of PRISMA checklist for reporting Systematic Reviews (Liberati et al., 2009), with the risk of bias assessment conducted following the high standards of both the COSMIN guidelines for evaluation of content validation studies (Terwee et al., 2018a, b) (for content validation of adaptation and development studies) and the 'COSMIN Manual for Systematic Reviews of PROMs' (Mokkink et al., 2018a) (for all other validations). The complex and varied sources of evidence were qualitatively summarized in an easy-to-digest format, while giving the benefit of doubt where it was reasonable to do so, to make some evidence-based recommendations where they could be made (and to stay silent where evidence was simply not enough to conclude one way or the other). We also searched an exceptionally large number of databases (14 in all, including grey literature), and included as many non-English papers as we could translate in this study.

Unfortunately, there were some noteworthy limitations to our study. Firstly, not all eligible abstracts could be obtained or translated. These have been listed in Appendix II. Secondly, The COSMIN guidelines that formed the backbone of the risk of bias assessment of methodology was primarily designed for use for "patient reported outcome measures" or PROMs- i.e., for interview-based questionnaires. The criteria therefore had to be adapted by authors for the evaluation of performance-based tasks/instruments, in places where the questionnaire-targeted criteria would not be appropriate for such performance-based instruments. However, we have endeavoured to be as transparent as possible where these adaptations were made and why and are of the opinion that these adaptations did not affect the overall impression of the quality of evidence for each performance-based task as most COSMIN criteria could be reasonably applied to such tasks (as well as to questionnaires). Thirdly, we did not assess practical considerations like copyright issues and cost surrounding use of these tools, which are all important and

relevant since these can all affect the availability and use of such tools particularly for resource-poor LAMICs. Finally, in a few cases where either legacy instruments (such as WCST and ROCF) or some proprietary instruments were involved, it was not possible to obtain the original development studies if they were not freely available online (some proprietary instruments only published their development in their instrument manuals which need to be purchases along with the tools), and therefore an estimation rating had to be made for the development tools based on information that was publicly available, assuming a best-case scenario each time.

Conclusion

To the best of our knowledge, this is the first systematic review of EF and AF tools focused on children in LAMICs of the scope and depth that we have done. In it we sought to critically appraise the kind and quality of evidence available rigorously for the psychometric soundness of instruments developed or adapted for use in LAMIC and made recommendations based on the evidence we found. From this careful evaluation, the BRIEF and the VABS appeared to have the best evidence for use in this context. However, we must stress that the conclusions or findings from this review are not absolute. They are limited to findings of evidence (or lack thereof) as of the time of research, and (more importantly) for the country-settings of interest. Therefore, our recommendations notwithstanding, it is not to categorically say that that the other instruments are inadequate for use, or that good evidence for each of these (other) instruments cannot be generated in the future, or that even now there is not good evidence for their use in HICs or other settings (other than LAMICs). Therefore, interpretation of results must be limited only to current evidence available as of time of publication and to the specified country-contexts evaluated.

Appendices

Appendix I: Search Strategy for MEDLINE

1. "executive function*".ti,ab.
2. "executive dysfunction*".ti,ab.
3. "dysexecutive syndrome".ti,ab.
4. "frontal lobe* function*".ti,ab.
5. "frontal lobe* dysfunction*".ti,ab.
6. "function* of the frontal lobe*".ti,ab.
7. "frontal lobe* syndrome".ti,ab.
8. "frontal lobe* damage".ti,ab.
9. "prefrontal cortical damage".ti,ab.
10. "prefrontal damage".ti,ab.
11. "prefrontal cortical dysfunction*".ti,ab.

12. "prefrontal dysfunction*".ti,ab.
13. "prefrontal cortical function*".ti,ab.
14. "prefrontal function*".ti,ab.
15.   "adaptive function*".ti,ab.
16. "adaptive function* impair*".ti,ab.
17. "activities of daily living"/
18. "daily living".ti,ab.
19. "ADL".ti,ab.
20. or/1-14
21. or/15-19
22. "children".ti,ab.
23. 21 and 22
24. "valid*".ti,ab.
25.   "reliab*".ti,ab.
26. "unreliab*".ti,ab.
27. standard*".ti,ab.
28. "norm*"*".ti,ab.
29.   "reproducibl*".ti,ab.
30. "replica*".ti,ab.
31. "coefficient of variation".ti,ab.
32. "internal consistency".ti,ab.
33. "responsive*".ti,ab.
34. "assessment tool*".ti,ab.
35. "assessment instrument*".ti,ab.
36. "outcome assessment".ti,ab.
37. "outcome measure*".ti,ab.
38. "clinimetr*".ti,ab.
39. "clinometr*".ti,ab.
40. "cross-cultural".ti,ab.
41. "reproducibility of results"/
42. "sensitivity and specificity"/
43. behavior rating scale/
44. neuropsychological tests/
45. psychometrics/
46. discriminant analysis/
47. Or/24-46
48. Developing Countr*.**sh,kf.**
49.   (low adj3 middle adj3 countr*).**ti,ab.**
50. (lmic or lmics or third world or lami countr*).**ti,ab.**
51. transitional countr*.**ti,ab.**
52. ((developing or less* developed or under developed or underdeveloped or middle income or low* income or underserved or under served or deprived or poor*) adj (countr* or nation? or population? or world)).**ti,ab.**
53. ((developing or less* developed or under developed or underdeveloped or middle income or low* income) adj (economy or economies)).**ti,ab.**
54. (low* adj (gdp or gnp or gross domestic or gross national)).**ti,ab.**
55.   (Africa or Asia or Caribbean or West Indies or South America or Latin America or Central America).**hw,kf,ti,ab,cp.**
56. (Afghanistan or Albania or Algeria or Angola or Antigua or Barbuda or Argentina or Armenia or Armenian or Aruba or Azerbaijan or Bahrain or Bangladesh or Barbados or Benin or Byelarus or Byelorussian or Belarus or Belorussian or Belorussia or Belize or Bhutan or Bolivia or Bosnia or Herzegovina or Hercegovina or Botswana or Brasil or Brazil or Bulgaria or Burkina Faso or Burkina Fasso or Upper Volta or Burundi or Urundi or Cambodia or Khmer Republic or Kampuchea or Cameroon or Cameroons or Cameron or Camerons or Cape Verde or Central African Republic or Chad or Chile or China or Colombia or Comoros or Comoro Islands or Comores or Mayotte or Congo or Zaire or Costa Rica or Cote d'Ivoire or Ivory Coast or Croatia or Cuba or Cyprus or Czechoslovakia or Czech Republic or Slovakia or Slovak Republic or Djibouti or French Somaliland or Dominica or Dominican Republic or East Timor or East Timur or Timor Leste or Ecuador or Egypt or United Arab Republic or El Salvador or Eritrea or Estonia or Ethiopia or Fiji or Gabon or Gabonese Republic or Gambia or Gaza or Georgia Republic or Georgian Republic or Ghana or Gold Coast or Greece or Grenada or Guatemala or Guinea or Guam or Guiana or Guyana or Haiti or Honduras or Hungary or India or Maldives or Indonesia or Iran or Iraq or Isle of Man or Jamaica or Jordan or Kazakhstan or Kazakh or Kenya or Kiribati or Korea or Kosovo or Kyrgyzstan or Kirghizia or Kyrgyz Republic or Kirghiz or Kirgizstan or Lao PDR or Laos or Latvia or Lebanon or Lesotho or Basutoland or Liberia or Libya or Lithuania or Macedonia or Madagascar or Malagasy Republic or Malaysia or Malaya or Malay or Sabah or Sarawak or Malawi or Nyasaland or Mali or Malta or Marshall Islands or Mauritania or Mauritius or Agalega Islands or Mexico or Micronesia or Middle East or Moldova or Moldovia or Moldovian or Mongolia or Montenegro or Morocco or Ifni or Mozambique or Myanmar or Myanma or Burma or Namibia or Nepal or Netherlands Antilles or New Caledonia or Nicaragua or Niger or Nigeria or Northern Mariana Islands or Oman or Muscat or Pakistan or Palau or Palestine or Panama or Paraguay or Peru or Philippines or Philipines or Phillipines or Phillippines or Poland or Portugal or Puerto Rico or Romania or Rumania or Roumania or Russia or Russian or Rwanda or Ruanda or Saint Kitts or St Kitts or Nevis or Saint Lucia or St Lucia or Saint Vincent or St Vincent or Grenadines or Samoa or Samoan Islands or Navigator Island or Navigator Islands or Sao Tome or Saudi Arabia or Senegal or Serbia or Montenegro or Seychelles or Sierra Leone or Slovenia or Sri Lanka or Ceylon or Solomon Islands or Somalia or South Africa or Sudan or Suri-

name or Surinam or Swaziland or Syria or Tajikistan or Tadzhikistan or Tadjikistan or Tadzhik or Tanzania or Thailand or Togo or Togolese Republic or Tonga or Trinidad or Tobago or Tunisia or Turkey or Turkmenistan or Turkmen or Uganda or Ukraine or Uruguay or USSR or Soviet Union or Union of Soviet Socialist Republics or Uzbekistan or Uzbek or Vanuatu or New Hebrides or Venezuela or Vietnam or Viet Nam or West Bank or Yemen or Yugoslavia or Zambia or Zimbabwe or Rhodesia).**hw,kf,ti,ab,cp.**

57. Or/48-57
58. 20 and 23 and 47 and 57

NB: for search items 48 – 56, a special search filter for 'developing countries' authored by the Cochrane Library (Cochrane Library, 2012) was used verbatim.

Appendix II: Untranslated and/or Unavailable but Potentially Eligible Non-English Articles

1. Zhang Q. 2003 Development of Adaptive Skill Rating Scale for school age children: *only Chinese PDF available*
2. Yao S. 1999 Development of the Adaptive Skill Rating Scale for Young Children (ASRSYC): *online version not available*
3. Rosselli-Cock M. 2004 Neuropsychological assessment of children: A test battery for children between 5 and 16 years of age. A Colombian normative study: *online version not available*
4. Pereira A. 2018 Executive functions in childhood: Assessment and preliminary normative data for Portuguese preschoolers: *study conducted in Brazil but online pdf version not available*
5. Du Q. 2010 Reliability and validity of the Behavior Rating Inventory of Executive Function-Adult Version Self-Report Form in China: *pdf version not available online*
6. Singh S. 2019 A comparative study of vineland adaptive behavior scale ii and vineland social maturity scale on children and adolescents with intellectual disability: *online version not readily available*
7. Pawlowski J. 2014 Evidences of construct validity of the NEUPSILIN using confirmatory factorial analysis: *only Portuguese pdf available, was unable to get translator*
8. Mashhadi A. 2014 Psychometric properties of the Behavior Rating Inventory of Executive Functioning-Preschool Version (Teacher Form): *only Persian version of pdf available, hence was difficult to translate using google translate.*
9. Pawlowski J. 2013 Reliability of the Brief Neuropsychological Assessment Instrument Neupsilin: *only Portuguese pdf version available, hence difficult to translate*

10. Ramos-Galarza C. 2019 EFECO Scale for Assessing Executive Functions in Self-Report Format: *for this study done in Ecuador, only Spanish version in pdf was available, hence difficult to translate*
11. Butman J. 2000 Spanish verbal fluency. Normative data in Argentina: *could not obtain full article online.*
12. Barreyro J. 2015 Validity and Reliability of the Running Memory Span: *only Spanish pdf version available online.*
13. Roselli Cock M. 2005 Neuropsychological Assessment of Children: a test battery for children between 5 and 16 years of age. A Colombian normative study: *only Spanish pdf version available online.*
14. Qian Y. 2011 Reliability and validity of the Chinese version of Weiss Functional Impairment Scale-Parent form for school age children: *no pdf version available at all online*
15. Dadsetan P. 2010 Kindergarten Inventory of Social/Emotional Tendencies: A cross-validation study: *no pdf available at all online*
16. Shi-jie Z. 2005 Development of the working memory battery and its validity in primary school students: *no pdf available online*
17. Lu T. 2017 Validity and reliability of the Behavior Rating Scale of Executive Function-Preschool Version parent form in China: *no pdf available online*
18. Ebrahimi A. 2016 Psychometric properties of the Behavior Rating Inventory for Executive Function-Preschool (BRIEF-P) among preschool children: *only Persian version of pdf available*
19. Andrea K. 2010 The executive functions from a neuropsychometric perspective: study from Hungary, but *no pdf available online*
20. de Oliveira A. 2014 Construction of a Scale to Assess Cognitive Planning: *study conducted in Brazil but only Portuguese pdf version available*
21. y Vila Molina G. 2010 Taylor's Figure Standardization on Mexican population: *study conducted in Mexico but only Spanish pdf version available*
22. Musso M. 2009 Assessment of executive functions in children: Analysis and adaptation of tasks in a school context: *study conducted in Argentina but only Spanish pdf version available*
23. Qian Y. 2009 Reliability and validity of the Behavior Rating Inventory of Executive Function Teacher Form for school age children in China: *no pdf available*
24. Butman J. 2000 Spanish verbal fluency test. Normative data in Argentina: *only Spanish pdf version available*
25. Abedi A. 2012 Standardization of the neuropsychological test of NEPSY on 3-4 years old children: *only Persian version of pdf available*
26. Ramos-Galarza C. 2019 EFECO scale for assessing executive functions in self-report format: *only Spanish pdf version available*

Appendix III:

Table 10 Study Characteristics of All Critically Appraised Papers (Kusi-Mensah et al., 2021)

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(Taşcu et al., 2012)	Wisconsin Card Sorting Test WCST	EF	Romania	construct validation (discriminant)	Mostly children and few adults	interview-based task	single task	1101	26.7	urban	clinical and healthy controls	Romanian
(Amini et al., 2016)	Iranian Children's Participation Assessment Scale (CPAS)	Adaptive functioning	Iran	Scale development	Children	parent report/ self-report	8 subscales: ADL, Instrumental ADL, Play, Leisure, Social Participation, education, Work, Rest/Sleep	40 kids, 21 parents	11.5 (2.8); range 6–18 years	urban	healthy	Persian
(Ford et al., 2019)	Rapid Assessment of Cognitive and Emotional Regulation (RACER)	EF	Lebanon, Niger	cross cultural validation	children	interview-based, computer administered task	2: inhibition; working memory	2725; 866–Niger, 1859 Lebanon	Lebanon: 9.2 (2.3); Niger: 9.2 (1.4)	urban	healthy	French; Arabic
(Malkawi et al., 2015)	Arabic Preschool Activity Card Sort (A-PACS)	adaptive functioning, ADL	Jordan	development	school-age children	parent report	7 domains: self-care, community mobility, high physical demand leisure and low physical demand leisure, social inter-action, domestic and education	115 caregivers	4.8	rural and urban	healthy	Arabic
(Senturk et al., 2014)	Junior Brixton Test (JBT)	EF	Turkey	Internal structure (structural validity), construct validation (convergent)	School-age children	interview-based task	single task	121	7.2	urban	healthy	Turkish
(Sobeh & Spijkers, 2012)	KITAP	Attention, inhibition, cognitive flexibility	Syria	Cross cultural validation, Construct validation (discriminant)	school-age children	performance-based task	Flexibility (The Dragon's House), Go/No-Go subscales	143	8	urban	healthy	Syrian Arabic
(Rosete, 2018)	Delis Kaplan Executive Function System (DKEFS); Children's Color Trails Test (CCTT); Color Trails Test (CTT)	EF	Thailand	Structural validation	adolescents	interview-based tasks	Design Fluency and Verbal fluency subscales of DKEFS	156	16.44 (2.04)	urban	healthy	Thai

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(de Bustamante Carim et al., 2012)	Behavior Rating Inventory of Executive Function (BRIEF)	EF	Brazil	structural validity and internal consistency	children and adolescents	self-report, parent and proxy (teacher) report	8 sub-scales: inhibition, flexibility and emotional control (behavioural regulation index), initiative, working memory, planning / organization, material organization and monitoring (metacognition index)	277 parents; 282 teachers; 112 adolescents	9.8 (3.4)	urban	healthy	Portuguese
(Korzeniowski & Ison, 2019)	Executive Function Scale for Children EFS	EF	Argentina	content validation, structural validity, internal consistency, reliability (test–retest)	school-age children	parent report	6 sub-scales: attention control, inhibitory control, metacognition, organization, planning and cognitive flexibility	307	7.7 (1.07)	urban	healthy	Spanish
(Burkey et al., 2015)	BRIEF	EF	Uganda	Internal consistency	school-age children	parent report	8 sub-scales: as above for BRIEF	185 children: 28 ADHD, 157 healthy	8.5 years	urban	clinical (ADHD) and healthy	Luganda
(Kashala et al., 2005)	Design Copying, Tower Test (from NEPSY), Digit Span (backwards)	EF: planning, working memory	DR Congo	Structural validation, construct validation (discriminant)	children	interview-based task	Design Copying, Tower Test (from NEPSY), Digit Span (backwards)	185 (28 cases, 157 controls)	7–9.9 years, mean 8.5 years	urban	mixed: 15% clinical (ADHD), 85% healthy	French; assorted African dialects
(Holding et al., 2018)	Rey-Osterrieth complex figure (ROCF), Go/ No-Go, Shift	EF: Working Memory, Inhibition, cognitive flexibility	Bangladesh, Ghana, Tanzania	adaptation, internal consistency, cross cultural validation, reliability (test–retest, inter rater), responsiveness	children	interview-based task	3 scales: each 1 item/task given multiple times	786 total: 166 Ghana, 323 Tanzania, 297 Bangladesh	range: 7–18; mean: 13	rural	healthy	Kasem and Nankam (Ghana), Swahili (Tanzania), Chatagonian (Bangladesh)
(Willoughby et al., 2019)	EF Touch (computer-based battery)	EF: inhibition, cognitive flexibility and working memory	Kenya	construct validation (discriminant)	pre-school children	computer-based, interview administered task	5 tasks/subscales: Go/no-go, variant of Stroop, Spatial conflict arrows, Pick the picture, Something's the same	193	range: 3–6yrs	urban	healthy	Kiswahili, English
(Bakar et al., 2011)	BRIEF	EF	Turkey	construct validation (convergent and discriminative)	school-age children	parent report	8 sub-scales: as above for BRIEF	80: 61 ADHD, 19 controls	6–11 years;	urban	ADHD and healthy	Turkish

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(Garcia-Barrera et al., 2015)	Behavioural Assessment System for Children (BASC)	EF	Colombia	internal consistency, cross cultural validation, Construct validation (discriminative)	school-age children	parent report	Not stated	848 healthy; 155 clinical	6–11 years	urban	ADHD and healthy	Spanish
(Rosetti et al., 2018)	Ball Search Field Task (BSFT), BRIEF	EF	Mexico	construct validation (convergent and discriminative (age), internal consistency (for BRIEF)	children and adolescents	interview-based task; parent report (BRIEF)	single task	106	range: 6–16; mean: 10.2 (2.7)	urban	ADHD only	Spanish
(Tombokan-Runtukahu & Nitko, 1992)	Indonesian-Vineland Adaptive Behaviour Scales (VABS)	adaptive functioning	Indonesia	adaptation (development), internal consistency, reliability (intra-rater and inter rater), cross cultural validation, Construct validation (discriminant)	children	parent report	4 domains: Communication, Daily Living Skills, Socialization, and Maladaptive Behavior	86: 43 ID, 43 healthy	range: 6- 18; mean: 11.8 (2.9)	semi-urban	Intellectual disability, healthy	Indonesian
(Amini et al., 2016)	Iranian- Child Participation Questionaire (I-CPQ)	adaptive functioning	Iran	Structural validity, internal consistency, reliability, construct validity (convergent)	pre-school children	parent report	6 subscales: self-care, home participation, play, leisure, social participation, and educational environment	120	range: 4–6; mean: 5.2	urban	cerebral palsy	Persian
(Malek et al., 2013)	Stroop Colour Word Test (Victoria Edition)	EF: inhibition	Iran	construct validation (discriminative), reliability (test–retest)	adolescents	interview-based task	single task	180: 150 healthy, 30 ADHD	range: 12–17 years	urban	ADHD and healthy	Persian and Turkish
(Sallum et al., 2017)	Self-Ordered Pointing Task (SOPT)	EF: working Memory	Brazil	construct validation (discriminative and convergent), ecological validity	pre-school children	interview-based task	single task	248	range: 3–5	urban	healthy	Portuguese
(Zarrabi et al., 2015)	BRIEF	EF	Iran	construct validation (discriminative)	school-age children	parent report	8 sub-scales: as above for BRIEF	60: 30 cases, 30 controls	range: 7–12 years; mean: 9yrs (1.41)	urban	ADHD and healthy	Persian

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(Holding & Kitsao-Wekulo, 2009)	Participation in Activities of Daily Living (PADL)	adaptive functioning	Kenya	development, internal consistency, construct validity (discriminative)	children	interview-based questionnaire	2 subscales: participation and limitations to participation	study development: 92, validation: 116	range: 6–18; mean: 11.7 (2.9)	rural	assorted chronic conditions (Sickle Cell, DM, etc.); acute encephalopathy; healthy controls	Kiswahili
(Mashhadi et al., 2020)	Barkley Deficits in Executive Functioning Scale–Children and Adolescents (BDEFS-CA)	EF	Iran	structural validation, internal consistency, reliability (test–retest), convergent validity, discriminative validity	children and adolescents	parent report/ self report	5 subscales: Self- Management to Time, Self-Organization/ Problem Solving, Self-Restraint, Self-Motivation and Self-Regulation of Emotion	2295	range: 6–18; mean: 11.6 (3.34)	rural and urban	mostly healthy, 4.7% clinical	Persian
(Tol et al., 2011)	Child Function Impairment Rating Scale (CFIRS)	adaptive functioning	Indonesia	development, Structural Validity, internal consistency, reliability (test-retest, inter rater), construct validity (convergent and discriminant)	children	FGD, diaries, interview-administered questionnaire	1 unidimensional scale with 4 factors; 11 items,	development: 53, validation: 403 children, 385 parents	mean: 9.9 (1.21)	rural	healthy but exposed to political violence	Indonesian
(Sartori et al., 2020)	Go/No-go app version	EF: inhibition	Brazil	content, structural validity, internal consistency, cross-cultural, and construct (convergent and discriminant)	school-age children	interview-based task	4 tasks of inhibition	306: 253 healthy, 53 DCD	range: 8–10 years	urban	Development Coordination Disorder, healthy	Portuguese
(Chernoff et al., 2018)	BRIEF	EF	South Africa, Malawi, Zimbabwe, Uganda	adaptation, reliability (test–retest), measurement error, convergent validity	school-age children	parent report	8 sub-scales: as above for BRIEF	603: 244 HIV, 179 HIV-exposed,180 no HIV)	range: 5–10	rural and urban	HIV, HIV exposed but healthy, no HIV exposure	Luganda, Afrikaans, Xhosa, Shona, Zulu, Chichewa, Sesotho, and Setswana
(Nampijja et al., 2010)	WCST, Knock-Tap Game (Go/No-Go)	EF: cognitive flexibility, planning	Uganda	adaptation, reliability (test–retest), discriminative validity	pre-school children	interview-based task	single task (each)	64	range: 4–6; mean: 5.2	rural and urban	healthy	Luganda

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(Thorell et al., 2020)	Teenage Executive Functioning Inventory (TEXI)	EF: Working memory, inhibition	Serbia	adaptation, structural validation, internal consistency, reliability (inter-rater)	adolescents	self-report, parent and proxy (teacher) report	20 items, 2 subscales: inhibition and working memory	302 adolescents	range: 13–19; mean 16.4 (1.8)	rural and urban	healthy	Serbian
(Barreto et al., 2018)	Neuropsychological Assessment of Executive Functions (ENFEN)	EF	Colombia	structural validation, construct validation (convergent and discriminant)	school-age children	interview-based task	Not stated	367 children	range: 6–12; mean 8.9 (1.8)	rural and urban	healthy	Spanish
(Yang et al., 2018)	Dysexecutive Questionnaire (DEX)	EF	China	structural validation	adolescents	self-report	20 items	1586	mean: 18.9 (5.4)	urban	healthy	Mandarin
(Munir et al., 1999)	Independent Behaviour Assessment Scale (IBAS)	adaptive functioning	Bangladesh	reliability, internal consistency, construct validity (discriminant)	pre-school and school-age children	parent report, interview-based tasks	4 subscales: Motor skills, Socialisation skills, Communication skills, Daily living skills	1404 healthy, 222 clinical	range: 2–9,	rural and urban	healthy and clinical	Bengali
(Selvam et al., 2016)	VABS II- Indian Version	adaptive functioning	India	cross cultural validation	pre-school children	parent report	4 domains: as above for VABS	412	2–6 years,	urban	healthy	Kannada
(de Siqueira et al., 2016)	Child Hayling Test- Brazilian version	EF: inhibition, cognitive flexibility	Brazil	adaptation (development)	school-age children	interview-based task	24 items	139	range: 6–12	urban	healthy	Portugese
(Richard's et al., 2017b)	Conjunction Visual Search (CVS) task of tareas de autorregulaCión Cognitiva (TAC) battery	EF: inhibition	Argentina	development (but of computer-based task), convergent validity	school-age children	interview-based, computer administered task	1 task: 120 trials	41	range: 6–11, M 8.49; SD 1.47	urban	healthy	Spanish
(Injoque-Ricle et al., 2011)	The Automated Working Memory Assessment (AWMA)	EF: working Memory	Argentina	adaptation (development), internal consistency, construct validity (convergent)	school-age children	interview-based task	12 tasks: Digit Recall, Word Recall, Non-word Recall, Dot Matrix, Block Recall, Mazes Memory, Listening Recall, Counting Recall, Backward Digit Recall, Odd One Out, Mr. X Spatial Span	26 adaptation; 210 validation	range 6–11	urban	healthy	Spanish
(Selvam et al., 2018)	BRIEF- pre-school version	EF	India	structural validity, internal consistency, reliability (inter rater), cross cultural validation, discriminant validation	pre-school children	parent report	5 sub-scales: inhibition, change, emotional control, working memory and planning / organization	412	2–5 years	urban	healthy	Kannada

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(Amani et al., 2018)	BRIEF- teacher	EF	Iran	structural validation, internal consistency, construct validity (convergence)	school-age children	proxy (teacher) report	8 sub-scales: as above for BRIEF	360 children, 12 teachers	range: NR	urban	healthy	Persian
(Amini et al., 2017)	CPAS- parent version	adaptive functioning	Iran	structural validation, reliability (test–retest), construct validation (convergent)	school-age children	parent- report	8 subscales: ADL, Instrumental ADL, Play, Leisure, Social Participation, education, Work, Rest/Sleep	700 parents	range: 6–12; mean 9.45 (1.76)	urban	healthy	Persian
(Green et al., 2019)	CANTAB	EF: working memory, inhibition, and attention shifting	Mexico	construct validation (discriminative)	children and adolescents	interview-based task, computer based	6 sub-tests (out of 23): stockings of Cambridge, delayed matching sample, extra dimensional shift, rapid visual information processing, stop signal task and match to sample visual search task	826 children	range 5–15; mean 9.27 (2.6)	urban	healthy	Spanish
(Richard's et al., 2018)	Tareas de Autorregulación Cognitiva Battery (TAC)	EF: inhibition, working memory and cognitive flexibility	Argentina	structural validity	children	interview-based task	3 sub-scales used: perceptual inhibition task, 4 WM tasks; 1 cognitive flexibility task	103 children	range: 9–12 years; M= 10.84 SD= 0.88	urban	healthy	Spanish
(Du et al., 2018)	Symptoms and Functional Impairment Rating Scale (SFIRS)	EF in ADHD: working memory, planning, time management, self monitoring, and emotional control	China	development, structural validity, internal consistency, reliability (test-retest), construct validity	school-age children	parent report	6 sub-scales: hyperactivity-impulsivity, self-control, inattention, self-management, school performance, social interaction	20 development; 412 validation	range: 6–12, 8.7 (1.42)	urban	ADHD and healthy	Mandarin
(Pineda et al., 2007)	bespoke battery including: WCST-abbreviated, verbal fluency test, token test, ROCF	EF in ADHD	Colombia	structural validation, discriminant validation	school-age children	interview-based task	3 tasks	621 total: 249 cases, 372 controls	range: 6–11,	urban	ADHD and healthy	Spanish

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(Trevisan et al., 2017)	Children's Executive Function Inventory (CHEXI)	EF in ADHD	Brazil	adaptation (development), structural validation, internal consistency, construct validity (convergence or concurrent)	pre-school children	parent report/ proxy report	4 subscales: working memory, inhibition, self-regulation and planning	408 parents and teachers	range 4–7, M=5.51, SD=0.59	urban	ADHD and healthy controls	Portuguese
(Arruda et al., 2020)	EFICA (Executive Function Inventory for Children and Adolescents)- parent and teacher versions	EF	Brazil	structural validation, internal consistency, reliability, construct validity (discriminant)	children and adolescents	parent report/ proxy (teacher) report	Not stated	3284 healthy, 165 ADHD	range 5–18, mean 8.2 (2.0)	rural and urban	ADHD and healthy	Portuguese
(Pluck et al., 2019)	Tower Test of DKEFS	EF: planning, working memory	Ecuador	Structural validity, internal consistency, reliability (test-retest)	children and adolescents	interview-based task	1 task	264	range 10–20, mean 15.05	urban	healthy, but some street children and foster children with controls	Spanish
(Wong et al., 2012)	Ballet Executive Scale (BES)	EF in dance/ ballet	Cuba	structural validation, internal consistency, concurrent validity	adolescents	self-report	5 subscales: Strategic Planning, Organization of Dance Behavior, Motivation to Dance, Impulse Control in Dance Behavior, Empathy toward Other Dancers	149	range: 11–18	urban	healthy	Spanish
(Diaz & Ancona, 2017)	BRIEF- pre-school version	EF	Colombia	adaptation, structural validation, internal consistency, test-retest reliability, construct validation (convergent)	pre-school children	parent report	5 sub-scales: inhibition, change, emotional control, working memory and planning / organization	125	range: 2–5 years; mean 4.3	urban and rural	healthy	Spanish
(Goldberg et al., 2009)	VABS- Vietnam version	adaptive functioning	Vietnam	adaptation, structural validity, internal consistency, construct validation (discriminant)	pre-school children	parent report	4 domains: as above for VABS	120 healthy, 31 ID children	range: 3–6yrs; mean 4.9 (1.1)	urban	intellectual disability and healthy	Vietnamese

Instrument Reference (authors, year)	Name and version of instrument	Outcome measure assessed	Country/ Study sites	Type of Study*: development, validation	Target Population	Mode of Admin	Sub-scales/ number of items of interest	Sample size	Age range and mean	Local Setting	Condition (clinical/ healthy)	Language(s) of population
(Ruffieux et al., 2010)	Assessment Battery containing: Color Trail Test, Hand movements, Verbal Fluency, letter numbering sequence	EF in Sickle Cell Disease	Cameroon	Structural validity, inter rater reliability	children and adolescents	interview-based task	14 tasks: but only 5 relevant for EFs	125	range: 6–20, mean 11.4 (4.2)	urban	healthy	French, unnamed local dialects
(Gonen et al., 2019)	Preschool Self-Regulation Assessment. (PRSA)	EF in pre-schoolers: self-regulation	Turkey, USA	cross cultural validation, reliability	pre-school children	interview-based task	3 tasks: The Balance Beam, Tower Task and Pencil Tap tasks	Turkey: 471 children+ caregivers; USA: 286 children+ caregivers	Turkey: 5.23yrs (2.83 to 5.92); US 4.12 yrs (2.97 to 5.07)	urban	healthy	Turkish, English
(Xu et al., 2020)	computerized version of: Stop Signal task; Figure Matching task, Spatial Span task; Tower of Hanoi task	EF	China, Hong Kong and USA	cross cultural validation	school-age children and adolescents	self-administered computer-based tasks	4 tasks each measuring: inhibition, cognitive flexibility, working memory, planning	China: 453; Hong Kong 371; UK 487	HK: 12.2yrs, UK: 11.9yrs; China: 11.93yrs	urban and rural	healthy	Mandarin, Cantonese, English

Declarations

Conflict of Interest No conflict to report for any of the authors.

Financial and Non-Financial Interests The authors have no relevant financial or non-financial interests to declare.

References

Allen, A. B., Finestone, M., Eloff, I., Sipsma, H., Makin, J., Triplett, K., Ebersöhn, L., Sikkema, K., Briggs-Gowan, M., & Visser, M. (2014). The role of parenting in affecting the behavior and adaptive functioning of young children of HIV-infected mothers in South Africa. *AIDS and Behavior, 18*(3), 605–616.

Alloway, T. P. (2007). *Automated Working: Memory Assessment: Manual*. Pearson.

Amani, M., Gandomani, R. A., & Nesayan, A. (2018). The reliability and validity of behavior rating inventory of executive functions tool teacher's form among Iranian primary school students. *Iranian Rehabilitation Journal, 16*(1), 25–34. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040818221&partnerID=40&md5=79338f26243ab34a2ef9c96d9fa6364a

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Standards for educational and psychological testing*. American Educational Research Association.

Amini, M., Hassani Mehraban, A., Rostamzadeh, O., & Mehdizadeh, F. (2017). Psychometric properties of the Iranian-Children Participation Questionnaire (I-CPQ) when used with parents of preschool children with cerebral palsy. *Occupational Therapy in Health Care, 31*(4), 341–351. https://doi.org/10.1080/07380577.2017.1382753

Amini, M., Mehraban, A. H., Haghni, H., Asgharnezhad, A. A., Mahani, M. K., Hassani Mehraban, A., & Khayatzadeh Mahani, M. (2016). Development and validation of Iranian children's participation assessment scale. *Medical Journal of the Islamic Republic of Iran*, *30*(1), 333. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem2&NEWS=N&AN=27390703

Ardila, A. (1995). Directions of research in cross-cultural neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17*(1), 143–150. https://doi.org/10.1080/13803399508406589

Arruda, M. A., Arruda, R., & Anunciação, L. (2020). Psychometric properties and clinical utility of the executive function inventory for children and adolescents: A large multistage populational study including children with adhd. *Applied Neuropsychology: Child, 11*(1), 1–17. https://doi.org/10.1080/21622965.2020.1726353

Association, A. P., & Association, A. P. (2013). Diagnostic and statistical manual of mental disorders: DSM-5. *Arlington, VA.*

Bailey, C. E. (2007). Cognitive Accuracy and Intelligent Executive Function in the Brain and in Business. *Annals of the New York Academy of Sciences, 1118*(1), 122–141. https://doi.org/10.1196/annals.1412.011

Bakar, E. E., Taner, Y. I., Soysal, A. S., Karakas, S., & Turgay, A. (2011). Behavioral rating inventory and laboratory tests measure different aspects of executive functioning in boys: A validity study [Davranış derecelendirme envanteri ve laboratuvar testleri erkek çocuklarda yönetici fonksiyonların farklı yönlerini ölçmektedir]. *Klinik Psikofarmakoloji Bulteni, 21*(4), 302–316. https://doi.org/10.5455/BCP.20111004014003

Bakare, M. O., Munir, K. M., & Bello-Mojeed, M. A. (2014). Public health and research funding for childhood neurodevelopmental disorders in Sub-Saharan Africa: a time to balance priorities. *Healthcare in Low-Resource Settings, 2*(1). https://doi.org/10.4081/hls.2014.1559

Barkley, R. A. (2012). *Barkley deficits in executive functioning scale–children and adolescents (BDEFS-CA).* Guilford Press.

Barreto, L. C. R., Pulido, J. H. P., Torres, J. D. C., & Estupinan, G. P. F. (2018). Psychometric properties and standardization of the ENFEN test in rural and urban areas of Tunja city (Colombia). *Diversitas-Perspectivas En Psicologia, 14*(2), 339–350. https://doi.org/10.15332/s1794-9998.2018.0002.10

Bathelt, J., Holmes, J., Astle, D. E., Holmes, J., Gathercole, S., Astle, D., Manly, T., & Kievit, R. (2018). Data-Driven Subtyping of Executive Function-Related Behavioral Problems in Children. *Journal of the American Academy of Child and Adolescent Psychiatry, 57*(4), 252-262.e4. https://doi.org/10.1016/j.jaac.2018.01.014

Bennett, P. C., Ong, B., & Ponsford, J. (2005). Measuring executive dysfunction in an acute rehabilitation setting: Using the dysexecutive questionnaire (DEX). *Journal of the International Neuropsychological Society, 11*(4), 376–385. https://doi.org/10.1017/S1355617705050423

Berardi, A., Panuccio, F., Pilli, L., Tofani, M., Valente, D., & Galeoto, G. (2021). *Evaluation Instruments for Executive Functions in Children and Adolescents: a Systematic Review, 21*(5), 885–896. https://doi.org/10.1080/14737167.2021.1908889

Berg, C., & LaVesser, P. (2006). The preschool activity card sort. *OTJR: Occupation, Participation and Health, 26*(4), 143–151.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology, 39*(1), 15–22. https://doi.org/10.1080/00221309.1948.9918159

Bitta, M., Kariuki, S. M., Abubakar, A., & Newton, C. R. J. C. (2017). Burden of neurodevelopmental disorders in low and middle-income countries: A systematic review and meta-analysis. *Wellcome Open Research, 2*, 121. https://doi.org/10.12688/wellcomeopenres.13540.3

Burgess, P. & Shallice, T. (1997) *The Hayling and Brixton tests. Test manual.* Bury St Edmunds, UK: Thames Valley Test Company. ISBN: 978-1874261261.

Burkey, M. D., Murray, S. M., Bangirana, P., Familiar, I., Opoka, R. O., Nakasujja, N., et al. (2015). Executive function and attention deficit/hyperactivity disorder in Ugandan children with perinatal HIV exposure. *Global Mental Health (Cambridge, England), 2,* e4–e4. https://doi.org/10.1017/gmh.2015.2

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*(2), 141–154. https://doi.org/10.1037/0735-7028.31.2.141

Chernoff, M. C., Laughton, B., Ratswana, M., Familiar, I., Fairlie, L., Vhembo, T., Kamthunzi, P., Kabugho, E., Joyce, C., Zimmer, B., Ariansen, J. L., Jean-Philippe, P., & Boivin, M. J. (2018). Validity of Neuropsychological Testing in Young African Children Affected by HIV. *Journal of Pediatric Infectious Diseases, 13*(3), 185–201. https://doi.org/10.1055/s-0038-1637020

Chevignard, M. P., Catroppa, C., Galvin, J., & Anderson, V. (2010). Development and Evaluation of an Ecological Task to Assess Executive Functioning Post Childhood TBI: The Children's Cooking Task. *Brain Impairment, 11*(2), 125–143. https://doi.org/10.1375/BRIM.11.2.125

Clark, C., Prior, M., & Kinsella, G. (2002). The relationship between executive function abilities, adaptive behaviour, and academic achievement in children with externalising behaviour problems. *Journal of Child Psychology and Psychiatry, 43*(6), 785–796. https://doi.org/10.1111/1469-7610.00084

Cochrane Library. (2012). *LMIC Filters | Cochrane Effective Practice and Organisation of Care.* https://epoc.cochrane.org/lmic-filters

de Bustamante Carim, D., Miranda, M. C., & Bueno, O. F. A. (2012). Tradução e Adaptação para o Português do Behavior Rating Inventory of Executive Function—BRIEF = Translation and adaptation into Portuguese of the Behavior Rating Inventory of Executive Function—BRIEF. *Psicologia: Reflexão e Crítica, 25*(4), 653–661. https://doi.org/10.1590/S0102-79722012000400004

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system.* Pearson San Antonio. ISBN: 978-0158091310.

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

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diaz, M. A. R., & Anacona, C. A. R. (2017). Adaptation and evaluation of the psychometric properties of brief-p in colombian preschoolers. *Revista Ces Psicologia, 10*(1), 48–62. https://doi.org/10.21615/cesp.10.1.4

Doyle, A. E., Willcutt, E. G., Seidman, L. J., Biederman, J., Chouinard, V.-A., Silva, J., & Faraone, S. V. (2005). Attention-deficit/hyperactivity disorder endophenotypes. *Biological Psychiatry, 57*(11), 1324–1335.

Du, Y., Li, M., Jiang, W., Li, Y., & Coghill, D. R. (2018). Developing the Symptoms and Functional Impairment Rating Scale: A Multi-Dimensional ADHD Scale. *Psychiatry Investigation, 15*(1), 13–23. https://doi.org/10.4306/pi.2018.15.1.13

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School Readiness and Later Achievement. *Developmental Psychology, 43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology, 10*(JULY), 1–18. https://doi.org/10.3389/fpsyg.2019.01507

Ford, C. B., Kim, H. Y., Brown, L., Aber, J. L., & Sheridan, M. A. (2019). A cognitive assessment tool designed for data collection in the field in low- and middle-income countries. *Research in Comparative and International Education, 14*(1), 141–157. https://doi.org/10.1177/1745499919829217

Garcia-Barrera, M. A., Karr, J. E., Duran, V., Direnfeld, E., & Pineda, D. A. (2015). Cross-cultural validation of a behavioral screener

for executive functions: Guidelines for clinical use among Colombian children with and without ADHD. *Psychological Assessment, 27*(4), 1349–1363. https://doi.org/10.1037/pas0000117

Garcia-Barrera, M. A., Karr, J. E., Duran, V., Direnfeld, E., & Pineda, D. A. (2015). Cross-cultural validation of a behavioral screener for executive functions: Guidelines for clinical use among Colombian children with and without ADHD. *Psychological Assessment, 27*(4), 1349–1363. https://doi.org/10.1037/pas0000117

Gardiner, E., & Iarocci, G. (2018). Everyday executive function predicts adaptive and internalizing behavior among children with and without autism spectrum disorder. *Autism Research, 11*(2), 284–295. https://doi.org/10.1002/AUR.1877

Geisinger, K. F. (1994). Cross-Cultural Normative Assessment: Translation and Adaptation Issues Influencing the Normative Interpretation of Assessment Instruments. *Psychological Assessment, 6*(4), 304–312. https://doi.org/10.1037/1040-3590.6.4.304

Gilotty, L., Kenworthy, L., Sirian, L., Black, D. O., & Wagner, A. E. (2010). *Adaptive Skills and Executive Function in Autism Spectrum Disorders, 8*(4), 241–248. https://doi.org/10.1076/CHIN.8.4.241.13504

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behavior rating inventory of executive function: BRIEF*.

Gligorović, M., & Đurović, N. B. (2014). Inhibitory control and adaptive behaviour in children with mild intellectual disability. *Journal of Intellectual Disability Research, 58*(3), 233–242. https://doi.org/10.1111/jir.12000

Goldberg, M. R., Dill, C. A., Shin, J. Y., & Nguyen, V. N. (2009). Reliability and validity of the Vietnamese Vineland Adaptive Behavior Scales with preschool-age children. *Research in Developmental Disabilities, 30*(3), 592–602. https://doi.org/10.1016/j.ridd.2008.09.001

Gomez, I. N. B., Palomo, S. A. M., Vicuña, A. M. U., Bustamante, J. A. D., Eborde, J. M. E., Regala, K. A., Ruiz, G. M. M., & Sanchez, A. L. G. (2021). Performance-Based Executive Function Instruments Used by Occupational Therapists for Children: A Systematic Review of Measurement Properties. *Occupational Therapy International, 2021*. https://doi.org/10.1155/2021/6008442

Gonen, M., Guler-Yildiz, T., Ulker-Erdem, A., Garcia, A., Raikes, H., Acar, I. H., Ozkan-Yildiz, F., Karlidag, I., Ucus, S., & Davis, D. L. (2019). Examining the association between executive functions and developmental domains of low-income children in the United States and Turkey. *Psychological Reports, 122*(1), 155–179. https://doi.org/10.1177/0033294118756334

Green, R., Till, C., Al-Hakeem, H., Cribbie, R., Téllez-Rojo, M. M., Osorio, E., Hu, H., & Schnaas, L. (2019). Assessment of neuropsychological performance in Mexico City youth using the Cambridge Neuropsychological Test Automated Battery (CANTAB). *Journal of Clinical and Experimental Neuropsychology, 41*(3), 246–256. https://doi.org/10.1080/13803395.2018.1529229

Guerra, A., Hazin, I., Siebra, C., Rezende, M., Silvestre, I., Le Gall, D., & Roy, A. (2020). Assessing executive functions in Brazilian children: A critical review of available tools. *Applied Neuropsychology. Child*, 1–13. https://doi.org/10.1080/21622965.2020.1775598

Holding, P., Anum, A., van de Vijver, F. J. R., Vokhiwa, M., Bugase, N., Hossen, T., et al. (2018). Can we measure cognitive constructs consistently within and across cultures? Evidence from a test battery in Bangladesh, Ghana, and Tanzania. *Applied Neuropsychology Child, 7*(1), 1–13. https://doi.org/10.1080/21622965.2016.1206823

Holding, P., & Kitsao-Wekulo, P. (2009). Is assessing participation in daily activities a suitable approach for measuring the impact of disease on child development in African children? *Journal of Child and Adolescent Mental Health, 21*(2), 127–138. https://doi.org/10.2989/JCAMH.2009.21.2.4.1012

Huppert, F. A., Brayne, C., Gill, C., Paykel, E. S., & Beardsall, L. (1995). CAMCOG–a concise neuropsychological test to assist dementia diagnosis: Socio-demographic determinants in an elderly population sample. *British Journal of Clinical Psychology, 34*(4), 529–541.

Injoque-Ricle, I., Calero, A. D., Alloway, T. P., & Burin, D. I. (2011). Assessing working memory in Spanish-speaking children: Automated Working Memory Assessment battery adaptation. *Learning and Individual Differences, 21*(1), 78–84. https://doi.org/10.1016/j.lindif.2010.09.012

Isquith, P. K., Gioia, G. A., & Espy, K. A. (2004). Executive Function in Preschool Children: Examination Through Everyday Behavior. *Developmental Neuropsychology, 26*(1), 403–422. https://doi.org/10.1207/s15326942dn2601_3

Jobe, J. B., & Mingay, D. J. (1989). Cognitive research improves questionnaires. *American Journal of Public Health, 79*(8), 1053–1055.

Kashala, E., Elgen, I., Sommerfelt, K., Tylleskar, T., & Lundervold, A. (2005). Cognition in African children with attention-deficit hyperactivity disorder. *Pediatric Neurology, 33*(5), 357–364. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med6&NEWS=N&AN=16243224

Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & Ustün, T. B. (2007). Age of onset of mental disorders: A review of recent literature. *Current Opinion in Psychiatry, 20*(4), 359–364. https://doi.org/10.1097/YCO.0b013e32816ebc8c

Korkman, M. (1998). NEPSY. A developmental neurop-sychological assessment. *Test Materials and Manual*.

Korzeniowski, C., & Ison, M. (2019). Executive functioning scale for schoolchildren: An analysis of psychometric properties [Escala de Funcionamiento Ejecutivo para Escolares: Análisis de las Propiedades Psicométricas]. *Psicologia Educativa, 25*(2), 147–157. https://doi.org/10.5093/psed2019a4

Kusi-Mensah, K., Nuamah, N. D., Wemakor, S., Agorinya, J., Seidu, R., Martyn-Dickens, C., & Bateman, A. (2021). Assessment tools for executive function and adaptive function following brain pathology among children in developing country contexts: a scoping review of current tools. *Neuropsychology Review, 169*(7), 1–24. https://doi.org/10.1007/S11065-021-09529-W

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ (Clinical Research Ed.), 339*. https://doi.org/10.1136/bmj.b2700

Low Kapalu, C. M., Hall, J. J., & Wallace, D. P. (2018). Neuropsychological functioning of youth receiving intensive interdisciplinary pain treatment. *Journal of Pediatric Psychology, 43*(8), 870–881. https://doi.org/10.1093/jpepsy/jsy034

Luria, A. R. (1973). *Psychophysiology of the frontal lobes*. Academic Press.

Malek, A., Hekmati, I., Amiri, S., Pirzadeh, J., & Gholizadeh, H. (2013). The standardization of Victoria Stroop color-word test among Iranian bilingual adolescents. *Archives of Iranian Medicine, 16*(7), 380–384. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84879381324&partnerID=40&md5=30ac419347b132927ea9a155b6bf3468

Malkawi, S. H., Hamed, R. T., Abu-Dahab, S. M. N., AlHeresh, R. A., & Holm, M. B. (2015). Development of the Arabic Version of the Preschool Activity Card Sort (A-PACS). *Child: Care, Health and Development, 41*(4), 559–568. https://doi.org/10.1111/cch.12209

Mashhadi, A., Maleki, Z. H., Hasani, J., & Rasoolzadeh Tabatabaei, S. K. (2020). Psychometric properties of Persian version of the Barkley deficits in executive functioning scale–children and adolescents. *Applied Neuropsychology: Child*. https://doi.org/10.1080/21622965.2020.1726352

Matson, J. L., Rivet, T. T., Fodstad, J. C., Dempsey, T., & Boisjoli, J. A. (2009). Examination of adaptive behavior differences in adults with autism spectrum disorders and intellectual disability. *Research in Developmental Disabilities, 30*(6), 1317–1325. https://doi.org/10.1016/j.ridd.2009.05.008

Merikangas, K. R. K., Nakamura, E. F. E., & Kessler, R. C. (2009). Epidemiology of mental disorders in children and adolescents. *Dialogues in Clinical Neuroscience, 11*(1), 7–20. http://www.ncbi.nlm.nih.gov/pubmed/19432384.

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018a). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research, 27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C., & Terwee, C. B. (2018b). *COSMIN manual for systematic reviews of PROMs COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual.* www.cosmin.nl

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., de Vet, H. C. W. W., Prinsen, C. A. C., Mokkink, L. B., & de Vet, H. C. W. W. (2014). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties. *Quality of Life Research, 63*(7), 32. https://doi.org/10.1016/j.jclinepi.2010.02.006

Morgan, P. L., Li, H., Farkas, G., Cook, M., Pun, W. H., & Hillemeier, M. M. (2017). Executive functioning deficits increase kindergarten children's risk for reading and mathematics difficulties in first grade. *Contemporary Educational Psychology, 50*, 23–32. https://doi.org/10.1016/j.cedpsych.2016.01.004

Munir, S. Z., Zaman, S., & McConachie, H. (1999). Development of an Independent Behaviour Assessment Scale for Bangladesh. *Journal of Applied Research in Intellectual Disabilities, 12*(3), 241–252. https://doi.org/10.1111/j.1468-3148.1999.tb00080.x

Nampijja, M., Apule, B., Lule, S., Akurut, H., Muhangi, L., Elliott, A. M., & Alcock, K. J. (2010). Adaptation of western measures of cognition for assessing 5-year-old semi-urban Ugandan children. *British Journal of Educational Psychology, 80*(1), 15–30. https://doi.org/10.1348/000709909X460600

Nell, V. (1999). Luria in Uzbekistan: The Vicissitudes of Cross-Cultural Neuropsychology. *Neuropsychology Review, 9*(1), 45–52. https://doi.org/10.1023/A:1025643004782.

Nyongesa, M. K., Ssewanyana, D., Mutua, A. M., Chongwo, E., Scerif, G., Newton, C. R. J. C., & Abubakar, A. (2019). Assessing executive function in adolescence: A scoping review of existing measures and their psychometric robustness. *Frontiers in Psychology, 10*, 311. https://doi.org/10.3389/fpsyg.2019.00311

Omigbodun, O. O., & Bella, T. T. (2005). Obstetric risk factors and subsequent mental health problems in a child psychiatry clinic population in Nigeria. *Tropical Journal of Obstetrics and Gynaecology, 21*(1), 15–20. https://doi.org/10.4314/tjog.v21i1.14456

Perna, R., Loughan, A. R., & Talka, K. (2012). *Executive Functioning and Adaptive Living Skills after Acquired Brain Injury., 19*(4), 263–271. https://doi.org/10.1080/09084282.2012.670147

Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia, 20*(3), 249–262. https://doi.org/10.1016/0028-3932(82)90100-2

Pineda, D. A., Puerta, I. C., Aguirre, D. C., Garcia-Barrera, M. A., & Kamphaus, R. W. (2007). The role of neuropsychologic tests in the diagnosis of attention deficit hyperactivity disorder. *Pediatric Neurology*, *36*(6), 373–381.

Pluck, G., Amraoui, D., & Fornell-Villalobos, I. (2019). Brief communication: Reliability of the D-KEFS Tower Test in samples of children and adolescents in Ecuador. *Applied Neuropsychology: Child,* 1–7. https://doi.org/10.1080/21622965.2019.1629922

Portellano, J. A., Martínez, R., & Zumárraga, L. (2009). *ENFEN: Evaluación Neuropsicológica de las funciones ejecutivas en niños*. TEA Ediciones.

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Pugliese, C. E., Anthony, L., Strang, J. F., et al. (2015). Increasing adaptive behavior skill deficits from childhood to adolescence in autism spectrum disorder: Role of Executive Function. *J Autism Dev Disord, 45,* 1579–1587. https://doi.org/10.1007/s10803-014-2309-1

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*(1), 33–65. https://doi.org/10.1016/j.acn.2004.02.005

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. (Les problems.). [The psychological examination in cases of traumatic encepholopathy. Problems.]. *Archives De Psychologie, 28*, 215–285.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children: Manual*. American Guidamq Service, Inc. https://www.sciencedirect.com/science/article/pii/002244059490037X?via%3Dihub

Richard's, M. M., Introzzi, I., Zamora, E., & Vernucci, S. (2017a). Analysis of internal and external validity criteria for a computerized visual search task: A pilot study. *Applied Neuropsychology: Child, 6*(2), 110–119. https://doi.org/10.1080/21622965.2015.1083433

Richard's, M. M., Vernucci, S., Stelzer, F., Introzzi, I., & Guàrdia-Olmos, J. (2018). Exploratory data analysis of executive functions in children: A new assessment battery. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues, 1*–8. https://doi.org/10.1007/s12144-018-9860-4

Richard's, M. M., Vernucci, S., Zamora, E., Juric, L. C., Introzzi, I., & Guardia, J. (2017b). Empirical contributions for contrasting groups validity of the Tareas de Autorregulación Cognitiva (TAC) battery. *Interdisciplinaria, 34*(1), 173–192. https://doi.org/10.16888/interd.2017.34.1.11

Rosenberg, L., Jarus, T., & Bart, O. (2010). *Development and Initial Validation of the Children Participation Questionnaire (CPQ)., 32*(20), 1633–1644. https://doi.org/10.3109/09638281003611086

Rosete, H. S. (2018). Normative data on a neuropsychological screening instrument for school-aged adolescents in Chiang Mai, Thailand [ProQuest Information & Learning]. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 79*(1–B(E)). https://ezp.lib.cam.ac.uk/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2017-54456-296&site=ehostlive&scope=site

Rosetti, M. F., Ulloa, R. E., Reyes-Zamorano, E., Palacios-Cruz, L., de la Peña, F., & Hudson, R. (2018). A novel experimental paradigm to evaluate children and adolescents diagnosed with attention-deficit/hyperactivity disorder: Comparison with two standard neuropsychological methods. *Journal of Clinical and Experimental Neuropsychology, 40*(6), 576–585. https://doi.org/10.1080/13803395.2017.1393501

Rosetti, M. F., Ulloa, R. E., Vargas-Vargas, I. L., Reyes-Zamorano, E., Palacios-Cruz, L., de la Peña, F., Larralde, H., & Hudson, R. (2016). Evaluation of children with ADHD on the Ball-Search Field Task. *Scientific Reports, 6*(1), 1–10. https://doi.org/10.1038/srep19664

Ruffieux, N., Njamnshi, A. K., Mayer, E., Sztajzel, R., Eta, S. C., Doh, R. F., Kengne, A.-M., Ngamaleu, R. N., Chanal, J., Verdon, V., & Hauert, C.-A. (2010). Neuropsychology in cameroon: First

normative data for cognitive tests among school-aged children. *Child Neuropsychology, 16*(1), 1–19. https://doi.org/10.1080/09297040902802932

Sabat, C., Arango, P., Tassé, M. J., & Tenorio, M. (2020). Different abilities needed at home and school: The relation between executive function and adaptive behaviour in adolescents with Down syndrome. *Scientific Reports, 10*(1), 1683. https://doi.org/10.1038/s41598-020-58409-5

Sallum, I., da Mata, F. G., Cheib, N. F., Mathias, C. W., Miranda, D. M., & Malloy-Diniz, L. F. (2017). Development of a version of the self-ordered pointing task: a working memory task for Brazilian preschoolers. *Clinical Neuropsychologist, 31*(2), 459–470. https://doi.org/10.1080/13854046.2016.1275818

Sartori, R. F., Valentini, N. C., Nobre, G. C., & Fonseca, R. P. (2020). Motor and verbal inhibitory control: development and validity of the go/No-Go app test for children with development coordination disorder. *Applied Neuropsychology: Child,* 1–10. https://doi.org/10.1080/21622965.2020.1726178

Schonfeld, A. M., Paley, B., Frankel, F., & O'Connor, M. J. (2007). *Executive Functioning Predicts Social Skills following Prenatal Alcohol Exposure, 12*(6), 439–452. https://doi.org/10.1080/09297040600611338

Schünemann, H. J., Oxman, A. D., Brozek, J., Glasziou, P., Jaeschke, R., Vist, G. E., Williams, J. W., Kunz, R., Craig, J., Montori, V. M., Bossuyt, P., Guyatt, G. H., & Group, G. W. (2008). GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ : British Medical Journal, 336*(7654). https://doi.org/10.1136/bmj.a139

Selvam, S., Thomas, T., Shetty, P., Thennarasu, K., Raman, V., Khanna, D., Mehra, R., Kurpad, A. V., & Srinivasan, K. (2018). Development of norms for executive functions in typically-developing Indian urban preschool children and its association with nutritional status.*Child Neuropsychology : A Journal on Normal and Abnormal Development in Childhood and Adolescence*,*24*(2), 226–246.

Selvam, S., Thomas, T., Shetty, P., Zhu, J., Raman, V., Khanna, D., Mehra, R., Kurpad, A. V, & Srinivasan, K. (2016). Norms for developmental milestones using VABS-II and association with anthropometric measures among apparently healthy urban Indian preschool children. *Psychological Assessment, 28*(12), 1634–1645. https://doi.org/10.1037/pas0000295

Semrud-Clikeman, M., Romero, R. A. A., Prado, E. L., Shapiro, E. G., Bangirana, P., & John, C. C. (2017). Selecting measures for the neurodevelopmental assessment of children in low- and middle-income countries. In *Child Neuropsychology* (Vol. 23, Issue 7, pp. 761–802). Routledge. https://doi.org/10.1080/09297049.2016.1216536

Senturk, N., Yeniceri, N., Alp, I. E., & Altan-Atalay, A. (2014). An exploratory study on the Junior Brixton Spatial Rule Attainment Test in 6- to 8-year-olds. *Journal of Psychoeducational Assessment, 32*(2), 123–132. https://doi.org/10.1177/0734282913490917

Shallice, T., Marzocchi, G. M., Coser, S., Del Savio, M., Meuter, R. F., & Rumiati, R. I. (2002). Executive function profile of children with attention deficit hyperactivity disorder. *Developmental Neuropsychology, 21*(1), 43–71.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Altman, D. G., Booth, A., Chan, A. W., Chang, S., Clifford, T., Dickersin, K., Egger, M., Gøtzsche, P. C., Grimshaw, J. M., Groves, T., Helfand, M., & Whitlock, E. (2015). Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: Elaboration and explanation. *BMJ (online), 349*(January), 1–25. https://doi.org/10.1136/bmj.g7647

Simblett, S. K., Badham, R., Greening, K., Adlam, A., Ring, H., & Bateman, A. (2012). Validating independent ratings of executive functioning following acquired brain injury using Rasch analysis. *Neuropsychological Rehabilitation, 22*(6), 874–889. https://doi.org/10.1080/09602011.2012.703956

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22*(2), 173–187. https://doi.org/10.1016/J.ECRESQ.2007.01.002

Sobeh, J., & Spijkers, W. (2012). Development of attention functions in 5-to 11-year-old Arab children as measured by the German Test Battery of Attention Performance (KITAP): A pilot study from Syria. *Child Neuropsychology, 18*(2), 144–167. https://doi.org/10.1080/09297049.2011.594426

Sparrow, S. S., & Cicchetti, D. V. (1985). Diagnostic uses of the vineland adaptive behavior scales. *Journal of Pediatric Psychology, 10*(2), 215–225. https://doi.org/10.1093/jpepsy/10.2.215

Sparrow, S. S., Cicchetti, D. V, Balla, D. A., & Doll, E. A. (2005). *Vineland adaptive behavior scales: Survey forms manual* (2nd ed.). Pearson.

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology, 59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643.

Stuss, D. T. (1983). Frontal lobe lesions and behavior localization. In A. Kertesz (Ed.), *Localization in neuropsychology* (pp. 429–454). Academic Press.

Stuss, D. T., & Alexander, M. P. (2000). Executive functions and the frontal lobes: A conceptual view. *Psychological Research Psychologische Forschung, 63*(3–4), 289–298. https://doi.org/10.1007/s004269900007

Tascu, A., Iliescu, D., & Avram, E. (2012). Adapting and implementing the Wisconsin Card Sorting Test in Romania. A look at the clinical data. In M. and C., M. Anitei, & C. Vasile (Eds.), *PSIWORLD 2011* (vol. 33, pp. 1022–1026). https://doi.org/10.1016/j.sbspro.2012.01.277

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018a). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research, 27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Terwee, Caroline B, Prinsen, C. A., Chiarotto, A., Cw De Vet, H., Bouter, L. M., Marjan, J. A., Donald, W., Patrick, L., Mokkink, L. B., & Terwee, C. B. (2018b). COSMIN methodology for assessing the content validity of PROMs: User manual. *Circulation*, *120*(9), 0–70.

Thorell, L. B., Lazarević, N., Milovanović, I., Bugarski Ignjatović, V., Lazarevic, N., Milovanovic, I., & Ignjatovic, V. B. (2020). Psychometric properties of the Teenage Executive Functioning Inventory (TEXI): A freely available questionnaire for assessing deficits in working memory and inhibition among adolescents. *Child Neuropsychology : A Journal on Normal and Abnormal Development in Childhood and Adolescence, 26*(6), 1–8. https://doi.org/10.1080/09297049.2020.1726885

Thorell, L. B., & Nyberg, L. (2008). The Childhood Executive Functioning Inventory (CHEXI): A new rating instrument for parents and teachers. *Developmental Neuropsychology, 33*(4), 536–552. https://doi.org/10.1080/87565640802101516

Tol, W. A., Komproe, I. H., Jordans, M. J. D., & Susanty, D. (2011). Developing a function impairment measure for children affected by political violence: A mixed methods approach in Indonesia. *International Journal for Quality in Health Care, 23*(4), 375–383. https://doi.org/10.1093/intqhc/mzr032

Tombokan-Runtukahu, J., & Nitko, A. J. (1992). Translation, cultural adjustment, and validation of a measure of adaptive behavior. *Research in Developmental Disabilities, 13*(5), 481–501. https://doi.org/10.1016/0891-4222(92)90004-P

Trevisan, B. T., Dias, N. M., Berberian, A. A., & Seabra, A. G. (2017). Childhood executive functioning inventory: Adaptation and psychometric properties of the Brazilian version [Childhood executive functioning inventory: Adaptação e propriedades psicométricas da versão Brasileira]. *Psico-USF, 22*(1), 63–74. https://doi.org/10.1590/1413-82712017220106

Tseng, W. L., & Gau, S. S. F. (2013). Executive function as a mediator in the link between attention-deficit/ hyperactivity disorder and social problems. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 54*(9), 996–1004. https://doi.org/10.1111/jcpp.12072

Van de Weijer-Bergsma, E., Kroesbergen, E. H., Prast, E. J., & Van Luit, J. E. H. (2014). Validity and reliability of an online visual–spatial working memory task for self-reliant administration in school-aged children. *Behavior Research Methods, 47*(3), 708–719. https://doi.org/10.3758/S13428-014-0469-8

Venter, A. (2000). 'Cross-cultural' Issues in Child Development and Assessment within the South African Context. *Southern African Journal of Child and Adolescent Mental Health, 12*(2), 162–177. https://doi.org/10.1080/16826108.2000.9632381

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Ware, A. L., Crocker, N., O'Brien, J. W., Deweese, B. N., Roesch, S. C., Coles, C. D., Kable, J. A., May, P. A., Kalberg, W. O., Sowell, E. R., Jones, K. L., Riley, E. P., & Mattson, S. N. (2012). Executive Function Predicts Adaptive Behavior in Children with Histories of Heavy Prenatal Alcohol Exposure and Attention-Deficit/ Hyperactivity Disorder. *Alcoholism: Clinical and Experimental Research*, *36*(8), 1431–1441. https://doi.org/10.1111/J.1530-0277.2011.01718.X

WHO. (2008). *Country statistics | Statistics and Monitoring | UNICEF*. https://www.unicef.org/statistics/index_countrystats.html

Willoughby, M. T. T., Piper, B., Kwayumba, D., & McCune, M. (2019). Measuring executive function skills in young children in Kenya. *Child Neuropsychology, 25*(4), 425–444. https://doi.org/10.1080/09297049.2018.1486395

Wilson, B. A., Alderman, N., Burgess, P. W., Emslie, H., & Evans, J. J. (1996). *BADS: Behavioural assessment of the dysexecutive syndrome*. Pearson.

Wong, A., Rodriguez, M., Quevedo, L., Fernandez de Cossio, L., Borges, A., Reyes, A., Corral, R., Blanco, F., & Alvarez, M. (2012). Questionnaire of executive function for dancers: an ecological approach. *Assessment, 19*(3), 383–387. https://doi.org/10.1177/1073191110397483

World Bank Group. (2019). *World Bank Country and Lending Groups*. World Bank Group. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

World Health Organization. (2001). *International classification of functioning, disability and health: ICF*. World Health Organization.

Xu, C., Ellefson, M. R., Ng, F.F.-Y., Wang, Q., & Hughes, C. (2020). An East-West contrast in executive function: Measurement invariance of computerized tasks in school-aged children and adolescents. *Journal of Experimental Child Psychology, 199*, 104929. https://doi.org/10.1016/j.jecp.2020.104929

Yang, Z.-Y., Sun, S.-F., Lui, S. S. Y., Shi, H.-S., Xie, D.-J., Xie, W.-L., Wang, Y., Cheung, E. F. C., Shum, D. H. K., & Chan, R. C. K. (2018). An attempt at revisiting the factor structure of the Dysexecutive Questionnaire in the Chinese setting. *Psych Journal, 7*(1), 25–30. https://doi.org/10.1002/pchj.198

Zarrabi, M., Shahrivar, Z., Tehrani Doost, M., Khademi, M., & Zargari Nejad, G. (2015). Concurrent validity of the behavior rating inventory of executive function in children with attention deficit hyperactivity disorder. *Iranian Journal of Psychiatry and Behavioral Sciences, 9*(1), e213–e213. https://doi.org/10.17795/ijpbs213

Zimmermann, P., Fimm, B., & Gondan, M. (2002). *Testbatterie zur aufmerksamkeitsprüfung für kinder:(KiTAP);[das schloß der geister]*. Psytest.

Zorza, J. P., Marino, J., & Mesas, A. A. (2016). Executive Functions as Predictors of School Performance and Social Relationships: Primary and Secondary School Students. *The Spanish Journal of Psychology, 19*, 1–10. https://doi.org/10.1017/SJP.2016.23