School of Psychology
Ysgol Seicoleg

A Systematic Review of the Psychometric Properties of Cognitive Screens used in Paediatric Traumatic Brain Injury, and an Empirical Study Validating the Triangles Theory of Mind Task in Young Children with Emerging Behavioural and Social-Interpersonal Problems

Thesis submitted in partial fulfilment of the requirement for the degree of:

Doctorate of Clinical Psychology (DClinPsy)

South Wales Doctoral Programme in Clinical Psychology

Cardiff University

Treasa A.M. Murphy

Supervised by:

Prof Reg Morris, Dr Chris Hobson, Dr Cerith Waters

August 2021

Table of Contents

ACKNOWLEDGEMENTS **6**

PREFACE **7**

PAPER 1: SYSTEMATIC LITERATURE REVIEW 7
PAPER 2: EMPIRICAL STUDY 8

THE PSYCHOMETRIC PROPERTIES OF SCREENING TESTS FOR COGNITIVE IMPAIRMENT IN PAEDIATRIC TRAUMATIC BRAIN INJURY (TBI) POPULATIONS: A SYSTEMATIC REVIEW **10**

ABSTRACT **11**

RATIONALE 13
Global Incidence of Paediatric TBI *13*
The Impact of Paediatric TBI *14*
Rationale for Cognitive Screening Tests *15*
Cognitive Screening Tests and their Psychometric Properties *16*
Gaps in Current Knowledge and Rationale for SLR *17*
OBJECTIVES 17

METHODS **18**

STUDY PROCEDURE 18
STAGE A. PERFORM THE LITERATURE SEARCH 19
Step 1. Protocol and registration *19*
Step 2. Eligibility criteria *19*
Step 3. Perform a literature search *20*
Step 4. Study selection *21*
STAGE B. EVALUATE THE MEASUREMENT PROPERTIES 22
Data Extraction *23*
DATA EVALUATION 25
Risk of bias in individual studies *25*
Quality Assessment *26*
Synthesis of Results *27*

RESULTS **28**

STUDY SELECTION 28

STUDY CHARACTERISTICS 30
Cognitive Screening Tests *30*
Individual Study Characteristics *31*
RESULTS OF INDIVIDUAL STUDIES PER COGNITIVE SCREEN 41
Brain Injury (BI) Alert *41*
CNS Vital Signs *42*
Cognitive and Linguistic Scale (CALS) *43*
LANSE-C/A *44*
S-FAVRES *45*
The Pediatric Test of Brain Injury *46*
SYNTHESIS OF RESULTS 46
Structural Validity *46*
Internal Consistency *47*
Measurement Invariance *48*
Reliability *49*
Criterion Validity *50*
Hypotheses Testing *51*
Responsiveness *52*
DISCUSSION 53
SUMMARY OF EVIDENCE 54
Study Samples *54*
Cognitive Screening *54*
Content Validity *55*
Internal Structure *56*
Remaining Measurement Properties *57*
CLINICAL RECOMMENDATIONS 57
IMPLICATIONS FOR RESEARCH 58
LIMITATIONS 59
CONCLUSIONS 60
FUNDING 61
REFERENCES 62

VALIDATION OF THE TRIANGLES THEORY OF MIND TASK IN YOUNG SCHOOL-AGED CHILDREN WITH EARLY EMERGING BEHAVIOURAL AND SOCIAL-INTERPERSONAL PROBLEMS 70

ABSTRACT 71

THEORY OF MIND 72
TOM DEVELOPMENT 73
TOM DIFFICULTIES 75
HOW IS TOM MEASURED? 76
Traditional ToM Tasks 76
Triangles Task 77
RATIONALE FOR THE CURRENT STUDY 78
Hypotheses 79

METHODOLOGY 79

THE NEURODEVELOPMENTAL ASSESSMENT UNIT 79
ETHICAL APPROVAL 80
PARTICIPANT SAMPLE 80
MEASURES 81
Theory of Mind 81
Language 82
Executive Function 83
Social-Interpersonal Skills and Behavioural Problems 83
DATA ANALYSIS 84

RESULTS 85

DESCRIPTIVE STATISTICS 85
INTERRATER RELIABILITY 87
HYPOTHESES 1 AND 2 87
HYPOTHESIS 3 88
HYPOTHESIS 4 89
HYPOTHESIS 5 90
FOLLOW-UP ANALYSES 91
POST HOC POWER ANALYSIS 92

DISCUSSION 92

Interrater Reliability 93
Convergent Validity 94
Language 95
Executive Functioning 96
Social-Interpersonal Skills 97
Strengths and Limitations 97
Implications for Clinical Practice 99
Future Research 99

CONCLUSION 100

REFERENCES 101

APPENDICES 110

A. Journal of Neuropsychology Review: Manuscript Submission Guidelines .. 110
B. Full Electronic Database Searches 123
C. COSMIN Modified GRADE Approach for grading the quality of evidence . 128
D. Results of studies on measurement properties for each cognitive screen... 129
E. The British Journal of Developmental Psychology: Publication Guidelines 140
F. NDAU Referral Pack and Consent Forms 143
G. Ethical Approval of Empirical Project 150
H. TRIANGLES SCORING CRITERIA 151
I. SDQ Scoring 154

Acknowledgements

Firstly, I would like to thank the children and families from NDAU. Your contribution to research is invaluable. I hope my work is a merit to this.

Reg, you have stuck with me during every hurdle that has been thrown my way over the past few years. Thank you for your unwavering support and keeping me grounded. Chris and Cerith, you came to my rescue in my hour of need. Despite already being overloaded, you never let me feel like the burden I'm sure I was. Thank you for all of your support.

2017 Cohort, you have been the most special group of people to meet and share this eventful journey with. You not only helped me survive the past few years, but actually made it a joy.

Claire-Marie, you laughed with me during madness, and listened to me during despair. You made Wales my home away from home. Thank you.

My family, you have been the wind in my sails. Thank you for believing in me, loving me and feeding me. I know I haven't always been easy.

Arjun, it's difficult to put in to words everything I know you sacrificed for me over the last few years. Thank you for your patience and for valuing me as your wife. I love you and I hope it has all been worth it.

Finally, Ayan, my sweet boy. You don't know it yet but you have been a huge part of this journey too. Born in the middle of the chaos, you have been the gorgeous smile that made me step away. I hope I have done you proud.

Preface

This document comprises two distinct research papers, a systematic literature review and an empirical study, written for publication in peer reviewed journals. In its entirety it is an academic thesis submitted in partial fulfilment of the requirements for the degree of Doctorate of Clinical Psychology (DClinPsy).

Paper 1: Systematic Literature Review

Through a systematic literature review, the author sought to identify cognitive screening tests being used with school-aged children and adolescents who have acquired a traumatic brain injury (TBI), and assess and compare the quality of the instrument properties using a gold standard framework (Mokkink et al., 2010). As a leading cause of disability in child populations, TBI and its impact is important (WHO, 2006). There are both economic and health arguments for the use of cognitive screening to support early intervention and prevention for children most at risk of developing difficulties post injury (Morley et al., 2015; Quinlivan et al., 2015). A cognitive screen is defined as a brief, objective measure which is designed to be highly sensitive to cognitive difficulties in areas such as attention, memory, understanding, and reasoning (Burton & Tyson, 2015).

A 10-step process recommended by Mokkink et al. (2018) was followed in line with the quality framework, to ensure a scientifically robust and systematic procedure. Studies included in the review were focused on screen development or measurement properties. Five electronic databases were searched, and appropriate studies were identified through a strict screening process involving two independent reviewers. Instrument properties were assessed and compared across two levels: (1) Against a set of quality criteria for each property studied; and (2) In context of the overall study quality. A total of 2,601 papers were retrieved using the search terms, of which 14 papers were included in the final study. Within the 14 papers a total of six cognitive screens were found. Furthermore, 33 studies on measurement properties were reported within these papers. The six cognitive screens identified through the systematic literature review are tentatively recommended for clinical use, until further more robust research is conducted looking at content validity and internal structure. Findings were considered in context of the broader literature. Expert driven recommendations on the evaluation and selection of cognitive screening tests were provided.

Paper 2: Empirical Study

An empirical study was conducted to investigate whether a computerised Theory of Mind (ToM) test, the Triangles task, is an adequate test to use with young school aged children with emerging social and behavioural problems. This was the first time this test was validated in a sample of children predominantly aged 4 to 7 years presenting with these difficulties. ToM develops in early childhood and becomes more sophisticated over time (Peterson, Slaughter & Wellman, 2018). First order ToM refers to the ability to understand that others can hold different beliefs and thoughts to our own, and predict others' behaviour based on this understanding (Frith & Frith, 2005). It is a construct found across cultures and different clinical populations (Liu, Wellman, Tardif & Sabbagh, 2008). ToM is of particular interest due to its link with social and behavioural problems (Austin, Bondü & Elsner, 2020; Wells, et al., 2020). Therefore, it may provide clinicians with a measurable early risk factor for behavioural and social problems.

The Triangles task was compared with a group of established ToM tasks, to test the study hypotheses. In comparing the two tasks, the author sought to assess the degree to which the Triangles task accurately measures the construct first-order ToM, a process known as convergent validity. It was expected that the Triangles task would be less reliant on skills such as receptive language (the ability to understand language) and executive function (attention; working memory; inhibition control), constructs traditional tests are particularly sensitive to (Milligan, Astington & Dack, 2007). If this was the case it may support an argument for the use of the Triangles task within clinical settings.

A total of 55 children participated in this study. The Triangles task was observed to be an adequate test of first order ToM in this sample, when compared to traditional ToM tasks. There was partial evidence supporting hypotheses around language ability and executive function skills. The expected association between the Triangles ToM task and behavioural and social-interpersonal problems was not supported, but non-significant relationships were also found between social-interpersonal problems and traditional ToM tasks in this sample. As a result of this preliminary study, it was concluded that the Triangles task can be tentatively recommended for use as a measure of ToM in young children, although further research is needed. The clinical implications of the study findings and the feasibility of using this test are discussed. Additional research addressing some of the current study limitations is required.

THE PSYCHOMETRIC PROPERTIES OF SCREENING TESTS FOR COGNITIVE IMPAIRMENT IN PAEDIATRIC TRAUMATIC BRAIN INJURY (TBI) POPULATIONS: A SYSTEMATIC REVIEW

Author name: Treasa Murphy

Affiliation(s) of the author:

Doctorate of Clinical Psychology Programme, School of Psychology, Cardiff University, Cardiff, Wales, United Kingdom.

Cardiff and Vale University Health Board, National Health Service (NHS) Wales, Cardiff, Wales, United Kingdom.

Word count: 7,976 (excluding abstract, references, figures and tables)

Author Note

This paper was written for submission to the peer-reviewed journal Neuropsychology Review (Loring & Bowden, n.d.) (see Appendix A). In line with submission guidelines for the journal, this paper was prepared by the author using PRISMA-P (Preferred Reporting Items for Systematic reviews and Meta-Analyses Protocols) guidelines (Moher, Liberati, Tetzlaff & Altman, 2010) for systematic literature reviews and written in the style of the American Psychological Association Publication Manual (APA, 2020).

ABSTRACT

Structured Summary

Background: Traumatic brain injury (TBI) is one of the leading causes of disability in the paediatric population. Cognitive screening can support identification of those most at risk, requiring further assessment and intervention.

Objectives: To identify cognitive screening assessments that are used within paediatric traumatic brain injury (TBI) populations, and appraise and synthesise their psychometric properties.

Data Sources: MEDLINE, EMBASE, CINAHL, PsycINFO, PsycTests

Study Eligibility Criteria: Empirical studies on the development, use and accuracy of cognitive screening tests for paediatric TBI, published in peer-reviewed journals, were included.

Participants: The studied population was paediatric traumatic brain injury. School-aged children and adolescents aged 4 to 18 years were included.

Data extraction: Independent extraction of data from the final studies was conducted using pre-stipulated databases. Two independent researchers were involved in this process.

Study appraisal and synthesis methods: Quality appraisal of the studies was conducted using the COSMIN quality assessment tool and risk of bias checklist (Mokkink et al., 2018).

Results: Six cognitive screens were identified and assessed against the COSMIN quality criteria. The evidence varied greatly across 33 studies on psychometric properties.

Limitations: The methodology of studies was inconsistent, which limited conclusions and recommendations on the selection and use of cognitive screens.

Conclusions: All six cognitive screens are recommended for clinical use on a provisional basis. The evidence must be considered with caution until further research is conducted.

Implications of key findings: There is limited evidence for content validity, internal structure and other measurement properties, more robust research is needed following COSMIN criteria.

***Systematic Review Registration Number*:** CRD42021238163

Keywords: Paediatric; Traumatic Brain injury; Cognition; Screening; Assessment;

Declarations

Funding: The author was funded by Health Education and Improvement Wales as part of their training in Clinical Psychology.

Conflicts of interest/Competing interests: There are no conflicts of interest to be declared.

Availability of data and material: A string search strategy was developed using a concept-based approach to identify relevant papers for review. Databases were searched using access through Cardiff University online library.

Code availability: n/a

THE PSYCHOMETRIC PROPERTIES OF SCREENING TESTS FOR COGNITIVE IMPAIRMENT IN PAEDIATRIC TRAUMATIC BRAIN INJURY (TBI) POPULATIONS: A SYSTEMATIC REVIEW

Rationale

Definitions of traumatic brain injury (TBI) vary greatly across the literature as well as clinically (Chan, Thurairajah & Colantonio, 2015). It is a broad term which refers to an acquired injury to the head as a result of trauma from an external force and may result in damage to the cerebral cortex (Kirkwood, Yeates, Taylor, Randolph, McCrea & Anderson, 2008; Yeates & Taylor, 2005). TBIs are categorised as being in the mild, moderate to severe range. The majority of paediatric TBI cases are categorised as being in the mild range (Lee, 2007), while one third of cases admitted to hospital are estimated to be in the moderate to severe range (Thurman, 2016). Clinically the severity of an injury is diagnosed by neurology professionals who synthesise information collected using a number of methods, including, the Glasgow Coma Scale (GCS scores: mild = 13–15; moderate = 9–12; and severe ≤ 8), neuroimaging technology, and neuropsychological assessment (Braun et al., 2011; Keenan & Bratton, 2006). Keenan et al. (2006) described the complexity in diagnosing injury severity, where child development needs to be considered and possible gaps in information on pre-morbid abilities lie. The accuracy of using these diagnostic methods independent of each other has been criticised.

Global Incidence of Paediatric TBI

TBIs are a common occurrence and one of the leading causes of disability in the paediatric population (WHO, 2006). An epidemiological review by Dewan, Mummareddy, Wellons and Bonfield (2016) reported variable global estimates of paediatric TBI, ranging between 47 and 280 per 100,000 children. While incidence rates vary across countries, greater incidences are consistently reported in males versus females (Andersson, Sejdhage &

Wage, 2012; Faul, Wald, Xu & Coronado, 2010). Age appears to be a factor, with research indicating that very young children and adolescents are at a greater risk of a TBI (Dewan et al., 2016). In a UK study of clinical and demographic data from a paediatric intensive care unit, Parslow, Morris, Tasker, Forsyth and Hawley (2005) reported that children from families of low-socio economic status were also at greater risk. Primary causes of paediatric TBI included; falls, motor vehicle accidents and being struck by or against an object (Faul et al., 2010; Parslow et al., 2005).

The Impact of Paediatric TBI

There is a lack of consensus in the literature on prognosis after paediatric TBI. In their meta-analytic review of the literature Babikian and Asarnow (2009) highlighted significant variance in reporting across studies. Their findings suggest that injury trajectory can look different at an individual level. It must be noted that despite the quantity of available studies, the quality in order to conduct a meta-analysis was poor. In their systematic review of outcomes after mild TBI, Emery and colleagues (Emery et al., 2016) found neuropsychological difficulties to be more frequent in cases where children were hospitalised, where there was a history of mild TBIs or there were pre-morbid psychological difficulties. These findings stress the importance of understanding pre-morbid functioning and suggest a possible link with injury severity or the level of trauma experienced by the child. Inconsistencies across the literature are thought to be due to factors such as variance in outcome measures being used, poorly defined groups, the absence of control groups and an overall lack of scientific rigor in study design (Babikian et al., 2009; Emery et al., 2016).

Di Battista, Soo, Catroppa, and Anderson (2012) conducted a meta-analysis on studies looking at quality of life (QoL) after paediatric TBI. QoL was reported to be associated with the severity of the injury acquired as well as the longer it had been since injury. Similarly, Yeates, Taylor, Wade, Drotar, Stancin and Minich (2002) found the severity of the injury to be correlated with poorer neuropsychological outcomes when compared to orthopaedic controls. Recovery was observed to plateau over time and difficulties remained consistent 4 years post-injury. A longitudinal study of outcomes after severe TBI highlighted attention, memory and executive functioning as cognitive domains that may be affected (Van Heugten, Hendriksen, Rasquin, Dijcks, Jaeken & Vles, 2006). Consistent with findings by Yeates et al. improvement within the first two years of recovery was evident; however, follow-up suggested that difficulties at this point remained more consistent. Again, this study drew attention to the variability in outcomes after accounting for injury severity and the importance of understanding differences at an individual level, and thus highlights the need for early assessment to identify those most at risk of neuropsychological difficulties and in need of intervention.

Rationale for Cognitive Screening Tests

Measuring health related outcomes is a complex but essential process in healthcare settings (Pantaleon, 2019). Symptoms reported by children and their families often cannot be measured objectively and are multi-faceted. Clinicians therefore rely on valid and reliable measures in order to inform important clinical decisions. The literature on outcomes after TBI underlines the importance of cognitive screening in identifying those most at risk at an individual level, both in acute and post-acute stages of recovery. Research on screening tests has demonstrated their ability to enhance services by reducing costs (through early intervention and prevention), acting as an aid to clinical decision making and providing a low cost tool for repeated measures (Morley et al., 2015; Quinlivan et al., 2015). Bodies such as the American Academy of Neurology (AAN) have endorsed the use of screening in early detection of cognitive impairments (Petersen, Stevens, Ganguli, Tangalos, Cummings & DeKosky, 2001). Within a paediatric clinical pathway, cognitive screening can support clinicians in identifying and referring at-risk children and adolescents for further neuropsychological testing and potentially intervention. From a practical perspective, cognitive screens are designed to be quick to administer by multiple members of a clinical team and are less of a burden for patients to complete (Cordell et al., 2013). This helpful step in the clinical pathway can reduce costs and use of resources within services. It is important that these tools are effective in identifying those most at risk of neuropsychological difficulties and informing clinical decisions.

Cognitive Screening Tests and their Psychometric Properties

An initial scope of the literature found that cognitive screening tests were poorly defined across studies on paediatric TBI. Burton and Tyson (2015) outlined a gold standard criteria in their systematic review of cognitive screening assessments which included; tests that were easy and quick to administer; the assessment of a minimum of three cognitive domains; sensitivity ≥ 80% and specificity ≥ 60%. Greater weight was given to the sensitivity of a screening test due to the considered risks in missing individuals with cognitive impairments at the screening stage. Following consideration of these factors, for the purpose of this study cognitive screening assessments were defined as:

- brief in nature, taking approximately 30 minutes to administer
- an objective measure of a number of cognitive domains which may include; attention, memory, executive functioning, orientation, language, processing speed and perception
- developed to be sensitive to mild cognitive impairments

Currently, there are no existing quality assessment tools for studies on the psychometric properties of cognitive screens. Alternatively, Mokkink et al. (2010) developed an expert driven manualised checklist for clinicians, to support the selection of outcome measures (OMs). Ten psychometric properties were identified through their four round Delphi study, which fall under three domains; content validity; internal consistency and remaining measurement properties. Pre-defined standards and criteria were established for evaluation. In their systematic review of the literature since the publication of the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) guidelines, Gorst, Prinsen, Salcher-Konrad, Matvienko-Sikar, Williamson and Terwee (2020) reported a marked improvement in the methodological quality of assessment selection. The COSMIN checklist provides a standardised approach to assessing the methodological quality of studies looking at the psychometric properties of OMs (Terwee et al., 2018). Thus whilst cognitive screens and OMs are fundamentally different, their psychometric properties map on to one another, with the exception of responsiveness. Responsiveness refers to an OM's sensitivity to change over time, while the primary function of a screen is to be sensitive to mild cognitive impairments at one point in time.

Gaps in Current Knowledge and Rationale for SLR

To the best of the author's knowledge, there are currently no existing systematic literature reviews exploring the use and psychometric properties of cognitive screening tests for paediatric TBI.

Objectives

The objective of the systematic literature review was to identify cognitive screening assessments that are used within paediatric traumatic brain injury (TBI) populations, and appraise and synthesise their psychometric properties, in order to determine how cognitive screening is delivered and outcomes are assessed in a range of clinical contexts.

1. Which cognitive screening tests are being used with paediatric TBI populations?
2. Are there any screening tests which are acceptable against the COSMIN quality criteria?

METHODS

Study Procedure

COSMIN methodology for systematic reviews is considered the gold standard for assessing the development and psychometric properties of studies on OMs (Mokkink, Prinsen, Patrick, Alonso, Bouter, de Vet & Terwee, 2018; Prinsen, Mokkink, Bouter, Alonso, Patrick, De Vet & Terwee, 2018; Rosenkoetter & Tate, 2018; Terwee et al., 2018). The COSMIN 10 step procedure was followed in conducting the current study (see Table 1).

Table 1. COSMIN 10-Step Procedure

Stage A. Perform the literature search		
Step	**Procedure**	
1.	**Formulate the aim of the review**	
2.	**Formulate eligibility criteria**	
3.	**Perform a literature Search**	
4.	**Select abstracts and full-text articles**	
Stage B. Evaluate the measurement properties		
5.	**Evaluate content validity**	**Evaluate the overall quality of the screen (synthesis):** Evaluate the methodological quality of the included studies by using the **COSMIN risk of bias checklist**
6.	**Evaluate internal structure**	
	Structural validity	
	Internal consistency	
	Cross-cultural validity	

7.	**Evaluate the remaining measurement properties** • Reliability • Measurement error • Criterion validity • Hypotheses testing for construct validity • Responsiveness	Apply criteria for good measurement properties by using **quality criteria** Summarise the evidence and grade the quality of the evidence by using the **GRADE Approach**
Stage C. Select an Cognitive Screen		
8.	**Evaluate interpretability and feasibility**	
9.	**Formulate recommendations**	
10.	**Report the systematic review**	

COSMIN = COnsensus-based Standards for the selection of health Measurement INstruments

GRADE = Grading of Recommendations Assessment, Development and Evaluation principles

Stage A. Perform the Literature Search

Step 1. Protocol and registration

A review protocol was developed and registered with PROSPERO (National Institute for Health Research, 2021). The registration number is CRD42021238163.

Step 2. Eligibility criteria

Inclusion Criteria

The researcher conducting the systematic literature review included studies on cognitive screening tests for multi-domain cognitive impairment in paediatric TBI populations. Studies were included where the focus of the paper was on test development and/or psychometric properties reported. Studies that involved paediatric populations aged 4 to 18 years as their primary sample were included or where the cognitive assessment was developed to include this population. Where the sample population was mixed or unclear, the researcher discussed this with their supervisors and reached a consensus as to whether the study was appropriate. Studies were included when they were conducted in any clinical or research setting and at any time period post injury. Only studies published in peer-reviewed journals were included.

Exclusion Criteria

Studies looking solely at pre-school children were excluded, due to the developmental stage of this population, given their variable language abilities, and the complexity of assessing cognitive domains in this population. Cognitive screening studies that looked specifically at the diagnosis of Global Developmental Delay; Intellectual Disability (ID); Attention Deficit Hyperactivity Disorder (ADHD); Autism; and sports-concussion were also excluded. Studies that were not reported in English were excluded due to the lack of funding available for translation. Data only published in poster, conference or abstract format were excluded due to the lack of information on methodology provided. Studies where the screening test was used as an intervention OM, for example in a randomised controlled trial, but which did not focus on the development or psychometric properties of the test itself, were also excluded.

Step 3. Perform a literature search

Information sources

The following electronic databases were searched; MEDLINE; EMBASE; CINAHL; PsycINFO; and PsycTests. These databases were chosen based on their relevance to the research question, after an initial scope by the researcher and consultation with a subject matter expert in the university library. Reference lists of relevant papers were manually searched additionally. No date restrictions were applied. Contact details could not be found for study's authors who the researcher wanted to contact for further information on test development. This is discussed within the review limitations. The last search was conducted on March 26th 2021.

Search

A string search strategy was developed using a concept-based approach, using combinations of the following subject heading searches and keyword search terms: P?ediatric or child* or Adolescen* or youngster* or teen* or young person* or young people* or school-age*, and Cognitive screen* or cognit* assessments or cognit* screening or cognit* tests or cognit* measures. The search strategy was limited to study titles, but kept broad so that no important papers were missed. Terms were modified in order to search different databases. An initial scope by the researcher found that an alternate more focused search strategy, limited to title, abstract and key words returned a significant number of inappropriate studies. All electronic database searches are presented in full in the appendices (see Appendices B).

Step 4. Study selection

Searches generated using the broad review search strategy were imported into EndNote. Using EndNote, the body of papers retrieved were firstly de-duplicated. The researcher then systematically screened the remaining papers by title, abstract and full paper against the research questions and eligibility criteria outlined in the review proforma. A batch of papers (15%) were reviewed by an independent reviewer to ensure reliability within the systematic process. Any differences found were discussed and a final decision was made. A detailed study selection process is documented in the results section.

Stage B. Evaluate the Measurement Properties

Figure 1., based on the COSMIN manual (Mokkink et al., 2018, Pg. 25), outlines the general process of extracting and evaluating data on pre-stipulated measurement properties (Mokkink et al., 2018).

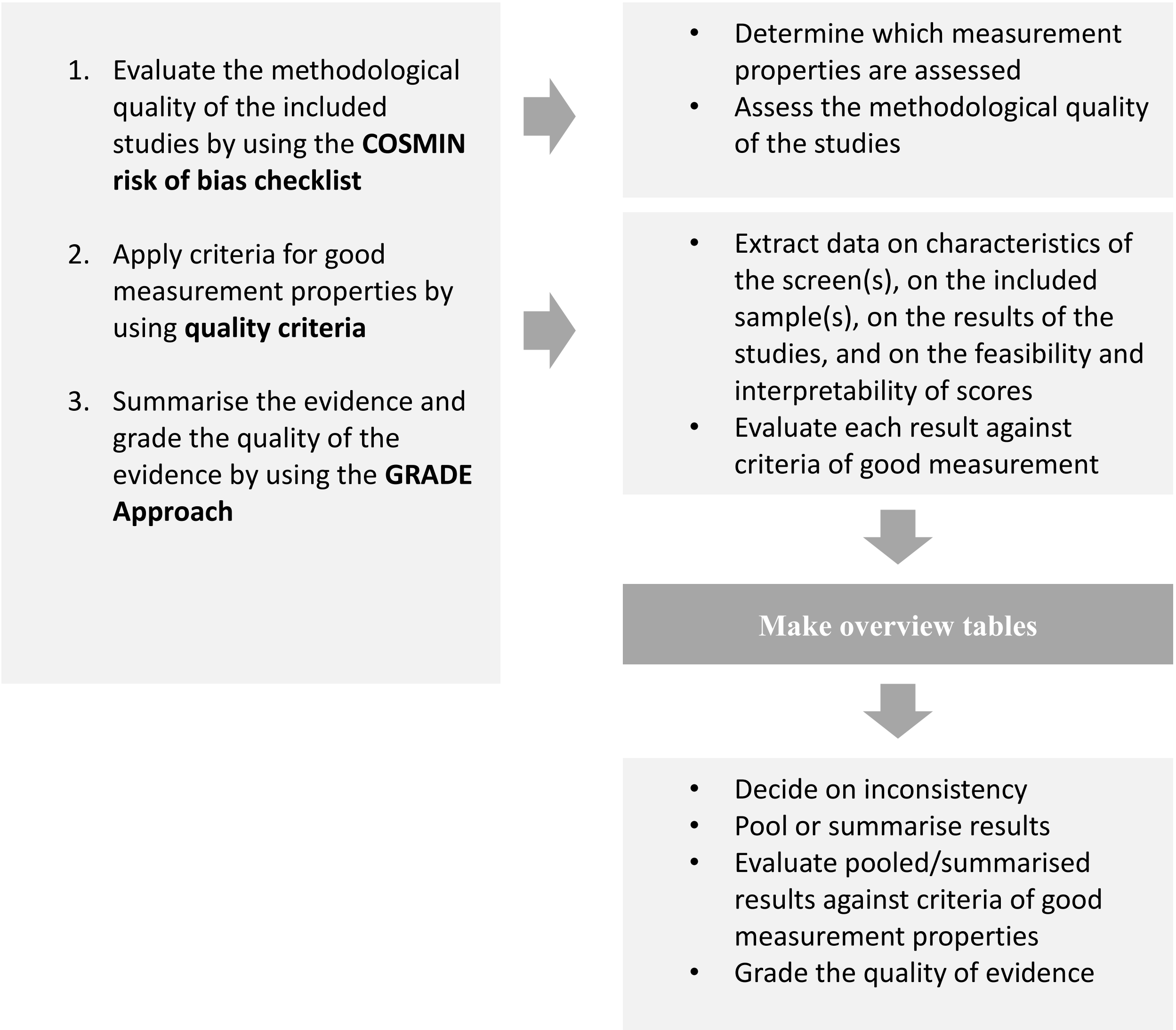

Figure 1. COSMIN general methodology for evaluation of measurement properties

GRADE = Grading of Recommendations Assessment, Development and Evaluation principles

Data Extraction

Data collection process

Data were extracted from reports independently and recorded in two Excel databases. The first database was developed by the author based on PRISMA-P and COSMIN guidelines for systematic literature reviews and reviewed by the research supervisor (RM). Data on study characteristics were extracted for descriptive analysis using this database. The second database was The Risk of Bias Checklist database, provided by COSMIN developers (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). This database provided the overall tool for analysis against the COSMIN quality criteria. Data were extracted and rated by the researcher (TM). A second DClinPsy student, familiar with the COSMIN procedures, acted as an independent reviewer (MJL) for the study. Any inconsistencies in ratings were discussed and a consensus reached. Where inconsistencies remained the researcher consulted with their supervisor.

Data items

Data were extracted on the characteristics of the cognitive screens included (construct; target population; mode of administration; administration time; subscales; response options; range of scores; language; available translations) and the study populations. Table 2 provides an overview of data sought based on COSMIN Risk of Bias Checklist with definitions from the COSMIN manual (Mokkink et al., 2018; Prinsen et al., 2018) and the order in which they are assessed. Items are assessed in order of importance as outlined by COSMIN methodology (Mokkink et al., 2018; Prinsen et al., 2018).

Table 2. Data items, COSMIN definitions and order of assessment

Content validity	***Definitions***
1. Cognitive Screen development	The overall process involved in developing the outcome measure, which may include feasibility and pilot studies
2. Content validity	The degree to which the content of a cognitive screen is an adequate reflection of the construct to be measured
Internal structure	
3. Structural validity	The degree to which the scores of a cognitive screen are an adequate reflection of the dimensionality of the construct to be measured
4. Internal consistency	The degree of the interrelatedness among the items
5. Cross-cultural validity\measurement invariance	The degree to which the performance of the items on a translated or culturally adapted cognitive screen are an adequate reflection of the performance of the items of the original version of the cognitive screen
Remaining measurement properties	
6. Reliability	The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. over time (test-retest); by different persons on the same occasion (interrater); or by the same persons on different occasions (intra-rater)

7. 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
8. Criterion validity	The degree to which the scores of a cognitive screen are an adequate reflection of a 'gold standard'
9. Hypotheses testing for construct validity	The degree to which the scores of a cognitive screen are consistent with 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 cognitive screen validly measures the construct to be measured
10. Responsiveness	The ability of a cognitive screen to detect change over time in the construct to be measured

Data Evaluation

Risk of bias in individual studies

Firstly, the risk of bias checklist was used to assess the methodological quality of each individual study reported in a paper. This was conducted at an outcome level (for each psychometric property). The checklist was used to guide the systematic process of extracting and analysing important factors against a set of standards (Mokkink et al., 2018). Depending on the measurement property, specific factors were considered during the assessment including: sample size; approach used for data analysis; stability of patients; test conditions; time intervals; and the clarity of reporting. Each factor was rated using the following scale:

very good (V); adequate (A); doubtful (D); inadequate (I); or not applicable (N). The methodological quality of the study was based on the lowest factor rating for a measurement property. For example, if the study is rated very good in all factors bar one, then the study is rated as doubtful overall.

Quality Assessment

The results of studies were evaluated against the COSMIN criteria for good measurement properties. Table 3 is adapted from the COSMIN criteria from which psychometric properties were analysed (Mokkink et al., 2018, Pg 28).

Table 3. COSMIN quality criteria

Measurement property	**Summary measures**
Structural validity	**CTT:** Confirmatory Factor Analysis: CFI (comparative fit index) or TLI (Tucker-Lewis index) or comparable measure or RMSEA (Root Mean Square Error of Approximation) **IRT/Rasch**: CFI or TLI or comparable measure OR RMSEA OR SRMR (Standardized Root Mean Residuals)
Internal consistency	Structural validity AND Cronbach's alpha(s) for each unidimensional scale or subscale
Reliability	Intra Class Correlation Coefficient (ICC) or weighted Kappa
Measurement Error	Smallest Detectable Change (SDC) or LoA (limits of agreement)
Hypotheses testing for construct validity	Analysis of Variance; Correlations

Cross-cultural validity\measurement invariance	Multiple group factor analysis (MGFA) OR Differential item functioning (DIF) for group factors
Criterion validity	Correlations of change; Area Under ROC Curve (AUC); Sensitivity/Specificity
Responsiveness	Analysis of Variance; Correlations

CTT = Classical Test Theory

IRT = Item Response Theory

ROC = Receiver Operating Characteristic

Each summary measure has clear criteria in order to rate the results as *sufficient (+), insufficient (-) or indeterminate (?)*. It is important children or adolescents with cognitive impairments are not missed by insensitive cognitive screening assessments, as they might then fail to receive necessary further assessment or interventions. Therefore, the nature of a screening tool needs to be highly sensitive to mild cognitive impairments, and hence reducing the risk of type II errors (false negatives). Through consultation with the literature the researcher set the criteria for sensitivity of screening assessments at ≥ 0.80. This criterion has been used in previous systematic literature reviews of cognitive screening assessments (Burton & Tyson, 2015). Although responsiveness is primarily a property of OMs, it was included in the analysis due to the use of cognitive screens as OMs in some clinical settings.

Synthesis of Results

Quantitative analysis of pooled results could not be conducted where there was significant variance across studies. Following COSMIN guidelines, the pooled study results for each measurement property per cognitive screen were summarised qualitatively and evaluated once more against the COSMIN quality criteria. Qualitative synthesis involved three steps: (1) Summarising the available evidence on each psychometric property for a screen; (2) Rating the summarised evidence following COSMIN guidelines as *sufficient (+), insufficient (–), inconsistent (±), or indeterminate (?)*; (3) Grading the overall quality of the evidence using the COSMIN GRADE approach (*high; moderate; low; very low*) (Mokkink et al., 2018). Grading of quality involved consideration of a number of factors: risk of bias; inconsistency across studies; imprecision (total sample); and indirectness (partially relevant population/conducted in the wrong context). Assessment of each factor can contribute to downgrading of the study quality. Synthesised data on the risk of bias was used to inform the graded approach on the methodological quality of evidence (see Appendices C) (Mokkink et al., 2018, Pg. 34). Overall ratings for psychometric properties were then considered in context of the quality of available evidence.

RESULTS

Study Selection

The search strategy used across the five electronic databases returned 3, 441 papers, and 10 additional records were found after manually searching the reference list of relevant papers. The full study selection process is outlined in the PRISMA flow diagram (see Figure 2) (Moher et al., 2010).

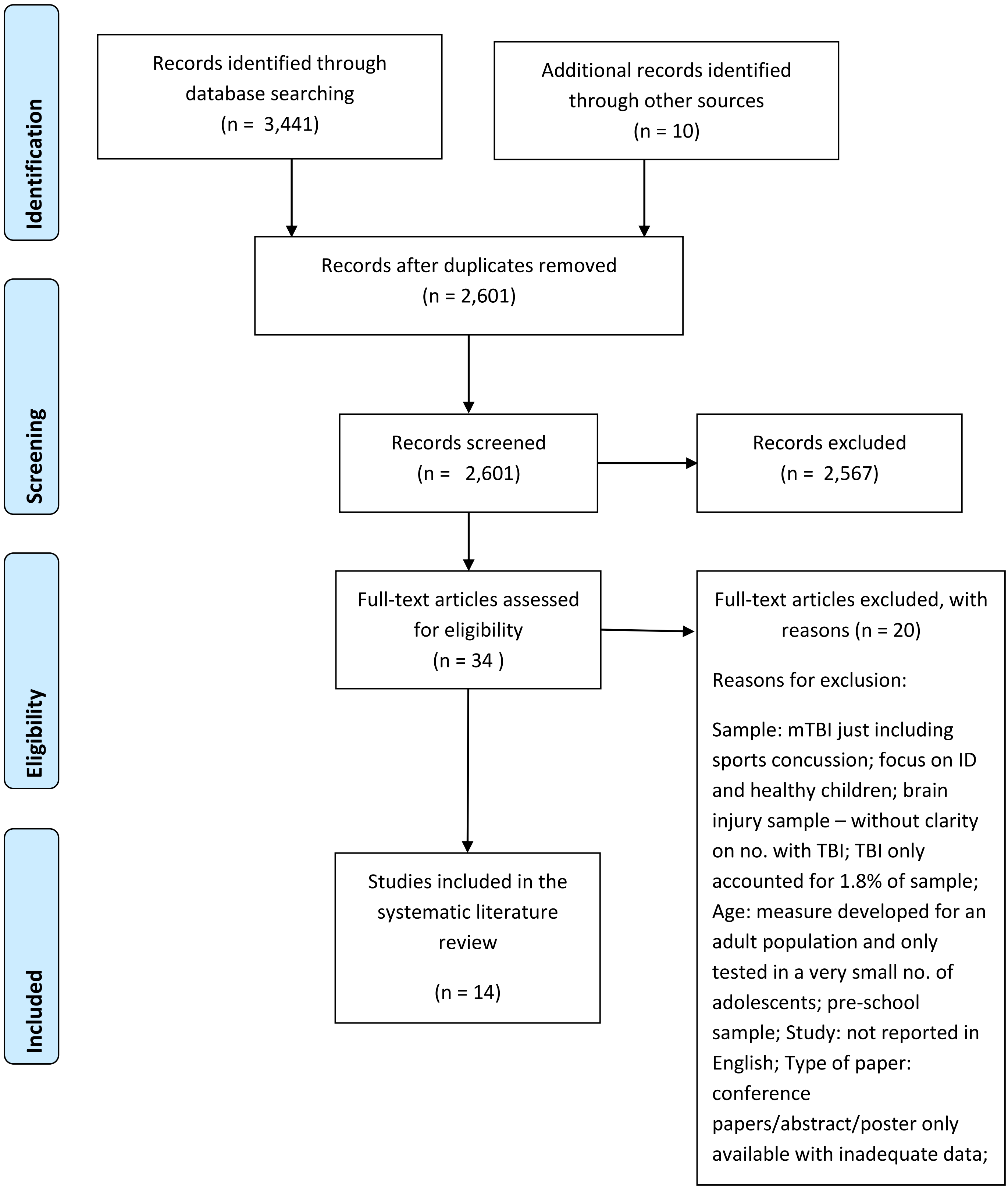

Figure 2. PRISMA flow diagram

After de-duplication, 2,601 papers were assessed against the eligibility criteria. Firstly, papers were screened by title and abstract and 2,567 papers were excluded. The reasons for exclusion of papers by title and abstract included: the use of adult samples; cognitive assessments not explicitly designed for screening; cognitive assessments explicitly designed for other neurological conditions (epilepsy, stroke, paediatric multiple sclerosis; sports concussion); and papers not reported in English. A final number of 34 papers were then screened by full article, of which 14 were included in the final review. The reasons for exclusion of full papers included; mTBI just including sports concussion; focus on ID and healthy children; screen developed for pre-school age or an adult population and only tested in a very small number of adolescents or number of adolescents not reported; language of study; conference papers/abstract/poster only available with inadequate data; brain injury sample without clarity on the number within the sample having a TBI or TBI only accounting for 1.8%.

Study Characteristics

In order to address the review research questions, the results were organised under four main headings. Cognitive screening tests and individual study characteristics provide data on which cognitive screening tests are being used with paediatric TBI populations. Results of individual studies per cognitive screen and synthesis of results provide data on the methodological quality of studies, as assessed using the COSMIN quality tool. These sections provide results on whether there are any screening tests which are acceptable against the COSMIN criteria.

Cognitive Screening Tests

A total of six cognitive screening tests were identified from the final 14 papers; Brain Injury Alert Screening Tool (BI Alert) (Rasquin, van Heugten, Winkens, Ritzen, Hendriksen & Vles, 2011); CNS Vital Signs (Gualtieri & Johnson, 2006); Cognitive and Linguistic Scale

(CALS) (Slomine, Eikenberg, Salorio, Suskauer, Trovato & Christensen, 2008); Lebby-Asbell neurocognitive screening examination for children and adolescents (LANSE-C/A) (Lebby, Pollock, Mouanoutoua & Lewey, 2015); Student version of the Functional Assessment of Verbal Reasoning and Executive Strategies (S-FAVRES) (MacDonald, 2015); and The Pediatric Test of Brain Injury (Hotz, Helm-Estabrooks & Nelson, 2001). The characteristics of the included screens are reported in Table 4.

Individual Study Characteristics

Table 5 presents the individual study characteristics under the six cognitive screens identified. Study characteristics were recorded against three main headings; population; disease characteristics and instrument administration.

Table 4. Characteristics of the included cognitive screens

Cognitive Screen	Construct(s)	Target population	Mode of administration	Administration Time	(Sub)scale (s)	Response options	Range of scores/scoring	Original language
Brain Injury Alert Screening Tool (Rasquin et al., 2011)	Cognitive, emotional and social problems	Paediatric Traumatic Brain Injury (6-18 years)	Interview with parents and teachers of children with TBI	5-60 mins (Avg. 15.4 mins for parents and 12.9 mins for teachers)	23 items that describe problems that can occur after brain injury, with 13 items covering the cognitive domain and 10 items covering the emotional and social domain	For each item the presence or absence is indicate and the level of severity is scored	Scoring 1 or 0 for presence/absence and the severity is scored (e.g. yes, the problem is present and it interferes with the development of the child; yes, the problem is present, but it is not interfering with the development of the child; the problem is not present; or the rater is not sure)	English
CNS Vital Signs (Gualtieri & Johnson, 2006)	Seven neuropsychologic al measures 1) verbal memory and 2) visual memory (composite memory); 3) psychomotor speed; 4) reaction time; 5) cognitive flexibility; 6)	Paediatric neurology patients including TBI	Computerised test, administered one-on-one by a trained professional using standardized instructions	25-30 mins	Overall Index Score; 7 neuropsychological measures with 18 subtests		VBM and VIM: 60-120; FTT: Avg number of taps left and right; SDC: As many correct responses in 120s. Stroop Test: Avg of two complex reaction time scores; SAT: no. of correct responses. CPT: correct responses,	English

Cognitive Screen	Construct(s)	Target population	Mode of administration	Administration Time	(Sub)scale (s)	Response options	Range of scores/scoring	Original language
	complex attention; 7) Neurocognition Index						commission errors, omission errors	
Cognitive and Linguistic Scale (CALS) (Slomine et al., 2008)	Cognitive and linguistic functioning	Children and adolescents (aged 2–19 years) with acquired or traumatic brain injury	Clinician rated assessment instrument (designed for serial administration by varying members of an interdisciplinary treatment team)	20-30 mins	20 items which assess - arousal, responsivity, emotional regulation, inhibition, attention, response time, orientation, memory, receptive language, expressive language, initiation, pragmatics, problem-solving, visuo-perceptual ability, visuospatial ability, self-monitoring and cognitive safety	Items are rated on the basis of performance on standardized administration, responses to semi structured interview questions, or via explicit behavioural observation	Scoring for each item ranges from 1 to 5; the total possible score ranges from 20 to 100. Higher scores reflect better performance	English
LANSE-C/A (Lebby et al., 2015)	1) Orientation 2) Attention 3) Language 4) Reasoning 5) Memory 6) Visual Perception and 7) Praxis	6-11year 11month old children /12-17 year 11month old Adolescents with TBI	Paper and pencil screen administered directly with the patient by an examiner	25-30 mins	7 subscales with 14 components; Verbal Auditory Memory; Number-Letter Sequencing; Judgment; Number-Sequencing Backward; Visual Memory; Orientation;	111/122 item measure	Correct items on each sub-scale are summed to yield a total sub-scale score, which is compared to an age-adjusted cut-off score; no full scale score	English

Cognitive Screen	Construct(s)	Target population	Mode of administration	Administration Time	(Sub)scale (s)	Response options	Range of scores/scoring	Original language
					Verbal Association; Number Sequencing Forward; Visual-Spatial Reasoning; Sentence Repetition; Receptive Language; Expressive Vocabulary; Object Use; Visual-Motor Integration;			
S-FAVRES (MacDonald, 2015)	Cognitive-communication skills	Adolescents (aged 12-19) with mild to severe ABI	Administered one-to-one by clinically experienced professionals or research assistants	20-113 mins	Accuracy, Rationale, Reasoning Subskills and Time - evaluates the interplay between complex comprehension, complex expression, social communication, verbal reasoning, problem-solving and executive functioning (4 Tasks)		Accuracy: Score for correct answer. Rationale: Score for reasons provided for choosing a particular answer. Time: Efficiency with which examinee completed the task. Analysis of reasoning sub-skills: Post-hoc analysis of the process the examinee engaged in to derive an answer. Strengths and weaknesses checklist: Qualitative scoring of behaviours.	

Cognitive Screen	Construct(s)	Target population	Mode of administration	Administration Time	(Sub)scale (s)	Response options	Range of scores/scoring	Original language
The Pediatric Test of Brain Injury (Hotz, Helm-Estabrooks & Nelson, 2001)	Cognitive-linguistic skills: attention, memory, language, reading, writing, metalinguistic, and metacognitive skills	Paediatric Brain Injury in school-aged children (6-16 years)	Administered one-to-one by an examinee	30-40 mins	10 tasks		Scores for accuracy of responses and behavioural descriptors	English

TBI = Traumatic Brain Injury, Avg = average, mins = minutes, CNS = Computeised Neurocognitive Assessment, VBM = Verbal Memory Test, VIM = Visual Memory Test, FTT = Finger tapping Test, SDC = Symbol Digit Coding, CPT = Continuous Performance Test, LANSE-C/A = The Lebby-Asbell Neurocognitive Screening Examination for Children and Adolescents, S-FAVRES = The Student Version of the Functional Assessment of Verbal Reasoning and Executive Strategies, ABI = Acquired Brain Injury

Table 5. Characteristics of the included study populations

		Population			Disease characteristics		Instrument administration		
Cognitive Screen	**Reference**	**N**	**Age Mean (SD, range) year**	**Gender**	**Disease**	**Disease duration mean (SD) year**	**Setting**	**Country**	**Language**
BI Alert	Rasquin et al. (2011)	Parents and teachers of children (N=42 children with TBI and N=29 controls); parents (TBI N=41, Controls n=29) and teachers (TBI N=36, controls N=27)	TBI Age M=10.6 (SD=3.6, range 6–16); Controls M=9.9 (SD=3.1)	TBI (Male/Female) 29/13; Control (Male/Female) 17/12;	TBI	TBI within the past 5 years; sustained at least 3 months before assessment	Interviews held in own home	Netherlands	English
CNS Vital Signs	Gualtieri & Johnson (2006)	Normative sample (N=1069); Test-re-test (normal N=40, neuropsychiatric N=59); Concurrent validity (neuropsychiatric	Normative sample age range 7-90 years (N= 25 <10 years; N=112 10–14years; N= 48 15–19years); Neuropsychiatric sample age range	Normative sample (28% males <10 years; 53.6% males 10–14 years; 56.3% males 15–19 years);	Neuropsychiatric sample incl. severe TBI			USA	English

		Population			Disease characteristics		Instrument administration		
Cognitive Screen	Reference	N	Age Mean (SD, range) year	Gender	Disease	Disease duration mean (SD) year	Setting	Country	Language
		patients incl. sTBI N=84 out of 144)	10–85 years, mean age 34.8;						
	Brooks & Sherman (2012)	Pediatric Neurology Patients (N=166) incl. TBI (N=44; 26.5% of neurology sample) and healthy controls (n=281)	7-19 years; neurology (M = 13.0 years; SD = 3.2) and controls (M = 13.2 years; SD = 3.2)	Male controls 154 (54.8), Neurology 79 (47.6); Female controls =124 (44.1) Neurology= 87 (52.4)	Neurology sample including complicated mild to severe TBI		Tertiary care pediatric hospital in Calgary, Alberta; Controls from CNS Vital Signs normative database	Canada	English
	Brooks et al. (2014)	mTBI in Youth (N=77) and orthopaedic controls (N=28);	Range: 8–17 years of age; mTBI mean age = 13.6 years (SD=2.6); controls mean age = 13.9 years (SD=2.1);	mTBI Males 58.2%; controls Males 50.0%	Mild TBI	Time between injury and testing: mTBI (M= 36.4 h) and OIC group (M=25.3 h)	Emergency department (ED)	Canada	English
	Gualtieri & Hervey (2015)	Study 1: N=3420; Normal; Study 2: N=3420; Normal; Study 3: Including TBI N=694;	S1 and S2: 4-90 years (M=39.6); S3 TBI: 7-85 years old (39.59);	S1 and S2: Males 51%; S3 TBI: Males 69.6%	Neurocognitive disorders including severe TBI			USA	English

			Population		Disease characteristics		Instrument administration		
Cognitive Screen	**Reference**	**N**	**Age** **Mean (SD, range) year**	**Gender**	**Disease**	**Disease duration mean (SD) year**	**Setting**	**Country**	**Language**
	Brooks et al. (2016)	TBI (N=77); orthopedic controls (N=28);	TBI mean age = 13.6 (SD = 2.6); OIC mean age = 13.9 years (SD = 2.1)	Males: TBI 58.4%; OIC 50.0%;	Mild TBI	Time of presenting at ED	Emergency department (ED)	Canada	English
	Plourde & Brooks (2017)	TBI (N=33); matched with 33 healthy controls	Children and Adolescents: 8-18 years; 8–12 years n=14 (42.4%); 13–18 years n = 19 (57.6%)	Male n=20 (60.6%); Female n=13 (39.4%);	Moderate to severe TBI; Moderate n=28 (84.8%); Severe n=5 (15.2%);	Within 6 months' post-injury; right before discharge from hospital;	TBI: Inpatient acute medical unit in a tertiary care hospital; Controls from CNS normative database;	Canada	English
	Brooks et al. (2019)	Neurology sample (N = 280); TBI (N=102; 36.4%)	M=14.0 (SD=2.9), range = 7.0–19.2	Male 46.1%; Female 53.9%;	Neurological diagnoses incl. TBI		Neuropsychology service at a tertiary pediatric hospital	Canada	English
CALS	Slomine et al. (2008)	N=100 children with acquired brain injury; incl. TBI (N=42);	M=10.5 years (SD=4.92), range 2–19 y	Male/female, % 56/44	Acquired Brain Injury, incl. TBI	32 days post-injury; Assessed at admission and discharge	Kennedy Krieger Institute's brain injury unit	USA	English

			Population		Disease characteristics		Instrument administration		
Cognitive Screen	**Reference**	**N**	**Age** **Mean (SD, range) year**	**Gender**	**Disease**	**Disease duration mean (SD) year**	**Setting**	**Country**	**Language**
LANSE-C/A	Lebby et al. (2015)	TBI (N=59); non-injured (N=190)	Children 6 – 11y 11m and adolescents 12 – 17y 11m	Children Female/Male % 40/60; Adolescents Female/Male % 35/65	Mild to severe TBI		Administered in an acute care facility	USA	English
	Kahn, Asbell & Donders (2015)	N=56	Range 12-17 year 11m	63% male	TBI	Median of 32 days post-injury	Inpatient rehabilitation setting	USA	English
S-FAVRES	MacDonald (2015)	N= 182 typically developing; ABI Group (N=59) of which incl. TBI (N=49; 83% of ABI sample)	Adolescents; age range 12–19 years	TD group: 59% (107) females and 41% (75) males; ABI group: 40% (23) females and 60% (34) males	Mild (n=10), moderate (n=2) and severe (n=32) Acquired Brain Injury (ABI) incl. TBI	Distribution of time post-injury for ABI: 0–6 months (11), 7–12 months (6), 13–24 months (13), 25–36 months (10) and >	Acute care hospitals, general hospitals, children's rehabilitation centres, schools and private practice clinics in the community	Canada and USA	English

			Population		Disease characteristics		Instrument administration		
Cognitive Screen	**Reference**	**N**	**Age** **Mean (SD, range) year**	**Gender**	**Disease**	**Disease duration mean (SD) year**	**Setting**	**Country**	**Language**
						36 months (17)			
The Pediatric Test of Brain Injury (Research Version)	Hotz, Helm-Estabrooks & Wolf Nelson (2001)	N=3 (Case illustrations)	Children 7-14 years	2 male; 1 female	TBI	9 days to 4 months post injury	Acute care and rehabilitation units at Jackson Memorial Medical Center/University of Miami School of Medicine	USA	English
	Hotz, Helm-Estabrooks, Wolf Nelson & Plante (2009)	N=2 (case illustrations)	14-15 years old	1 male; 1 female	TBI	1-2 months post-injury	Hospital	USA	English

BI = Brain Injury, TBI = Traumatic brain Injury, CNS = Computeised Neurocognitive Assessment, sTBI = Severe traumatic Brain Injury, mTBI = Mild Traumatic Brain Injury, OIC = Orthopaedic Controls, LANSE-C/A = The Lebby-Asbell Neurocognitive Screening Examination for Children and Adolescents, S-FAVRES = The Student Version of the Functional Assessment of Verbal Reasoning and Executive Strategies, ABI = Acquired Brain Injury

Results of Individual Studies per Cognitive Screen

Results on the methodological quality and individual quality assessment of studies are reported under the relevant screen below (see Appendices D). Within the 14 research papers retrieved for the systematic review, a total of 33 different studies of psychometric properties were reported.

Brain Injury (BI) Alert

There was one development and validity paper retrieved for the BI Alert screen (Rasquin et al., 2011). In their study, Rasquin et al. reported on screen development and content validity, which involved studies with parents and professionals on screen relevance, comprehensiveness and comprehensibility (see Appendix C). These studies were rated as *sufficient* overall. The evidence was of *moderate* quality due to poor reporting on interview methods used in selecting screen items, as well as quantitative methods being used to assess content validity.

A two factor solution was reported using exploratory factor analysis (EFA) (cognition and emotion/social behaviour), this study was of *adequate* quality due to the type of analysis used. Internal consistency was rated as *indeterminate* due to inconsistency in findings from teacher and parent reports with Cronbach's α 0.68 for parents and 0.82 for teachers. Using a multi-group confirmatory factor analysis (MGCFA) there were no important differences found for group characteristics including; age, sex, years of education and family situation. Evidence for measurement invariance was of *doubtful* quality due the small sample size (n = 71).

Despite authors reporting reasonable screen reliability (r = 0.46 - 0.82), test-re-test reliability was rated as *indeterminate* in a study of *doubtful* quality due to Pearson's correlation being used instead of intra-class correlation coefficients (ICC). There was *sufficient* evidence for convergent validity when testing the hypothesis that scores on the BI

Alert would correlate with parent and teacher scores on the child-behaviour-checklist (CBCL). This study was of *adequate* quality against COSMIN standards.

CNS Vital Signs

There were seven papers retrieved for the CNS Vital Signs (Brooks, Daya, Khan, Carlson, Mikrogianakis & Barlow, 2016; Brooks, Khan, Daya, Mikrogianakis & Barlow, 2014; Brooks, Plourde, Fay-McClymont, MacAllister & Sherman, 2019; Brooks & Sherman, 2012; Gualtieri & Hervey, 2015; Gualtieri & Johnson, 2006; Plourde & Brooks, 2017). The quality of screen development was *doubtful* as there was no indication of a qualitative process used to identify relevant items. Studies on content validity were rated as *inconsistent* overall, due to clear theoretical groundings being provided but no evidence of pilot studies involving patients and professionals. Items were instead based on tests widely used by neuropsychologists and therefore an assumption made by the developers that they are reliable, valid and comprehensive.

There were two studies reporting on structural validity, one using a large normative sample and one with a paediatric neurology sample. Both studies reported a three factor solution, the first using confirmatory factor analysis (CFA) and the second using EFA. This evidence was rated as *sufficient* based on consistency of findings and the quality of evidence from the earlier study by Gualtieri and Hervey (2015). There were no available studies reporting internal consistency of the screen. Four studies on measurement invariance were rated as *sufficient*. While two of the studies were *indeterminate*, the second two studies used a robust design and analysis providing *sufficient* evidence that there were no important differences found for group characteristics.

Test-retest reliability was rated as *indeterminate* in a study of *doubtful* quality due to Pearson correlation being used for analysis. Studies of criterion validity had an overall rating of *indeterminate* due to the level of inconsistency across studies. One study included a paediatric sample with mild TBI screened in the emergency department (sensitivity = 0.969) and one in a sample with moderate to severe TBI in an acute inpatient unit (sensitivity = 0.60), where time after injury was not clear (Brooks et al., 2016; Plourde & Brooks, 2017). The overall rating from the four studies on hypotheses testing for construct validity was *indeterminate* due to inconsistencies across findings. All four studies used a known-groups approach (two groups expected to have contrasting scores) and one assessed convergent validity. All four construct validity studies were of *very good* quality against COSMIN standards.

Cognitive and Linguistic Scale (CALS)

There was one paper returned for the CALS, reporting on a number of psychometric property studies (Slomine, Eikenberg, Salorio, Suskauer, Trovato & Christensen, 2008). The standards reported for screen development were of *doubtful* quality as the method of data collection and item selection was not clearly described, which highlighted a risk of bias in reporting. Content validity studies were *inconsistent* for relevance and comprehensiveness, as only professionals were consulted and the methods used were not clearly described. Comprehensibility of the screen was *indeterminate* as no patients were consulted. The quality of this evidence was *moderate* due to risk of bias in reporting standards.

Slomine et al. reported a two factor solution from their EFA (basic responding, higher-level cognitive skills), this result was rated as *indeterminate* due to the type of analysis used (EFA rather than CFA). Cronbach's α met the criteria for internal consistency of >0.70 (α = 0.96), but the overall result was *indeterminate* due to tests on structural validity being indeterminate.

Test-re-test reliability was *sufficient* (ICC = 0.99) in a study of *doubtful* quality due to the small sample size (n = 9). Results for hypotheses testing of convergent validity were *sufficient*, with over 75% in accordance with the hypothesis of there being strong correlations between the CALS and the WeeFIM. This was a study of *very good* quality against COSMIN standards (n = 100). Similar results were found for screen responsiveness when comparing results from admission and discharge across the two instruments.

LANSE-C/A

There were two papers returned for the LANSE-C/A (Kahn, Asbell & Donders, 2015; Lebby, Pollock, Mouanoutoua & Lewey, 2015), one of which only focused on the LANSE-A. Quality of screen development was *inadequate* due to a poor description of the constructs to be measured in the available papers. The development process involved administration of the screen to normal subjects to determine means and standard deviations for each subtest and age group, but there was no description of pilot studies consulting patients on content validity. The overall rating for screen relevance was *inconsistent,* due to only some aspects of the test being clearly described such as the context and target population. The researcher felt it was assumable, from the use of field experts and cross-reference with standardised neuropsychological tests measuring similar constructs, that the selected items were relevant; however, the construct was not clearly defined. The quality of evidence for screen relevance was *low* based on the available studies.

One study of *very good* methodological quality was *sufficient* for criterion validity with sensitivity of 94.62% reported for the failure of two sub-tests. The second criterion validity study was excluded based on COSMIN guidelines of no gold standard for comparison. Hypothesis testing for known-groups validity was deemed *sufficient* for construct validity, with over 75% of results in accordance with the hypothesis in a study of *very good* quality. Results of a responsiveness study looking at whether the LANSE-A would predict variance in the WeeFIM at discharge were *insufficient* to meet COSMIN criteria. The study was of *adequate* quality as there was no indication of previous studies looking at construct validity of the LANSE-A in comparison with the WeeFIM.

S-FAVRES

There was one standardisation paper returned for the S-FAVRES, which reported on a number of studies of psychometric properties (MacDonald, 2015). The paper provides clear descriptions of the construct, context and target population which the screen was designed for; however, the screen development study was of *doubtful* quality due to a lack of information on the overall methodology used in selecting items. Both patients and professionals were consulted in studies of content validity, where a quantitative approach to data collection was used. The evidence for relevance of measurement items was *sufficient* in a study of *moderate* quality. The rating for comprehensiveness was *indeterminate* and *inconsistent* for comprehensibility. Both studies were of *doubtful* quality due to a lack of information around the methodology used.

There were no studies found for structural validity of the S-FAVRES. The rating for internal consistency was *indeterminate* as the test did not meet the criteria of Cronbach's α >0.70 for all-subscales or for structural validity. There were no important differences found for group characteristics, however this study was of *inadequate* quality due to the number of subjects per group.

Results for test-retest reliability were *insufficient*, with ICC ranging from 0.28 to 0.80 across test sub-scales. This study was of *doubtful* quality as there was no description of attempts to assess patient stability over time and there was significant variance in the time interval for re-test. Criterion validity was *sufficient* in a study of very good quality (AUC = 0.85). There were two hypotheses tested for construct validity, one using comparison with another measure (The BRIEF) and the second using know-groups. The convergent validity study was of *doubtful* quality and had *insufficient* findings. The known-groups validity study was of *very good* quality and *sufficient* results were reported, as >75% of results were in accordance with the hypothesis.

The Pediatric Test of Brain Injury

There were two papers returned for The Pediatric Test of Brain Injury (Hotz, Helm-Estabrooks & Nelson, 2001; Hotz, Helm-Estabrooks, Nelson & Plante, 2009). These studies were based on the research version of the test and were limited to screen development. Both papers reported a clear rationale and detailed description for screen constructs, target population and context. The authors mention pilot testing being conducted in a sample representative of the target population but the methodology was not reported. The overall rating for item relevance was *inconsistent* as test development had strong theoretical underpinnings but there was no indication that patients were consulted in elicitation of items. Ratings of comprehensiveness and comprehensibility were *indeterminate* as there was no indication of patients or professionals being consulted on content validity.

Synthesis of results

There was significant variance across available studies per screen in terms of design, severity of injury, the use of mixed samples, and different age groups. In consideration of this, the researcher conducted a qualitative synthesis following the COSMIN guidance on summarised ratings of evidence and graded overall quality per measurement property for each screen (Mokkink et al., 2018).

Structural Validity

There were four reported studies on structural validity across three screens; BI Alert (1), CNS Vital Signs (2) and CALS (1). A summary of these studies in terms of their rating of summarised evidence and the graded overall quality of the evidence per screen is reported in Table 6. CNS Vital signs was the only measure with *sufficient* pooled results to meet the quality criteria. The evidence was of *moderate* quality, due to one *very good* study with a normative sample downgraded due to some inconsistencies in the findings. Findings for the

BI Alert and CALS had overall ratings of *indeterminate* due to the type of analysis used, EFA instead of CFA.

Table 6. Summary of Findings for Sructural Validity

Structural validity	**Summary**	**Rating of summarised evidence**	**Graded overall quality of evidence**
BI Alert	2 factor solution	Indeterminate (based on EFA)	Moderate (one study of adequate quality available)
CNS Vital Signs	3 factor solution	Sufficient	Moderate (Multiple studies of at least adequate quality; downgraded due to inconsistencies in analyses used with paediatric neurology sample)
CALS	2 factor solution	Indeterminate (based on EFA)	Moderate (one study of adequate quality available)

Internal Consistency

Three studies of internal consistency across three screens were reported; BI Alert (1), CALS (1) and S-FAVRES (1). A summary of these results are presented in Table 7. Despite the CALS meeting the criteria for Cronbach's $\alpha > 0.70$, the overall rating for all three screens was *indeterminate*. The reason for these ratings was not meeting the criteria of 'at least low level evidence for sufficient structural validity' (Mokkink et al., 2018).

Table 7. Summary of Findings for Internal Consistency

Internal consistency	Summary	Rating of summarised evidence	Graded overall quality of evidence
BI Alert	Inconsistency across parent ratings and teacher ratings: parents α = 0.68; teachers α = 0.82	Indeterminate (Criteria for at least low evidence for sufficient structural validity not met)	Moderate (One study of very good quality available; downgraded due to inconsistent results)
CALS	Cronb. alpha ≥ 0.70; α = 0.96	Indeterminate (At least low evidence for sufficient structural validity not met)	High (one very good study available)
S-FAVRES	Did not meet criteria of Cronb. Aplha >0.70 for all subscales	Indeterminate (criteria for at least low evidence of structural validity not met)	High (one very good study available)

Measurement Invariance

Six studies of measurement invariance were found for three screens; BI Alert (1), CNS Vital Signs (4) and S-FAVRES (1). Pooled results across all three screens had *sufficient* ratings overall (see Table 8). Methodological quality of the evidence found for CNS Vital Signs was *high* quality. Evidence for the BI Alert and S-FAVRES was of *low* quality, studies were assessed to have an *extremely serious* risk of bias due to inadequate sample sizes.

Table 8. Summary of Findings for Measurement Invariance

Measurement invariance	**Summary**	**Rating of summarised evidence**	**Graded overall quality of evidence**
BI Alert	No important differences found for group characteristics (age, sex, years of education, family situation etc)	Sufficient (use of MGCFA)	Very Low (Only one study of inadequate quality due to sample size)
CNS Vital Signs	No important differences found based on group characteristics	Sufficient (two studies using MGCFA)	High (multiple studies of at least adequate quality)
S-FAVRES	No important differences found between group factors	Sufficient	Very Low (one study of inadequate quality)

Reliability

Four studies of reliability were found for four of the screens; BI Alert (1), CNS Vital Signs (1), CALS (1) and S-FAVRES (1) (see Table 9). Results for the BI Alert and CNS Vital signs were *indeterminate* in *low* quality studies. Results for S-FAVRES were *insufficient* in context of a *low* quality study. Overall rating for the CALS was *sufficient* but in a study of *very low* methodological quality. There was a significant risk of bias across all studies on reliability, mainly due to the type of analysis used and in one due to the study design.

Table 9. Summary of Findings for Reliability

Reliability	Summary	Rating of summarised evidence	Graded overall quality of evidence
BI Alert	0.46 – 0.82	Indeterminate	Low (Only one study of doubtful quality available due to analysis used (Pearson's rather than ICC)
CNS Vital Signs	r = 0.314-0.874	Indeterminate	Low (Only one study of doubtful quality available; ICC/KAPPA not reported; Pearson's/Spearman's CC reported)
CALS	ICC = 0.99	Sufficient	Very Low (Only one study of doubtful quality due to time interval not being reported; and downgraded further due to small sample size;)
S-FAVRES	ICC = 0.28-0.80	Insufficient (ICC or weighted Kappa < 0.70)	Low (one study of doubtful quality due to sample size)

Criterion Validity

Five studies of criterion validity were found for three of the screens; CNS Vital Signs (3), LANSE-C/A (1) and S-FAVRES (1). Overall results were rated as *sufficient* for LANSE-C/A and S-FAVRES, both in studies of *high* methodological quality. There was no indication of risk of bias across these studies. Overall rating for the CNS Vital Signs was *indeterminate*, in context of evidence of *moderate* quality. The quality of evidence for these studies was downgraded due to inconsistencies in the findings and differing subgroups in terms of injury severity which made comparison of results difficult.

Hypotheses Testing

Nine studies were found on hypotheses testing for construct validity. These studies were across five out of the six screens; BI Alert (1), CNS Vital Signs (4), CALS (1), LANSE-C/A (1) and S-FAVRES (2) (see Table 10). Four of the studies looked at convergent validity, comparing the screen with another measurement instrument, and five of the studies used a known-groups approach. Overall ratings for the CALS, LANSE-C/A and S-FAVRES were *sufficient* in studies assessed as providing *high* quality evidence. Overall rating for the BI Alert was also *sufficient*, in a study of *moderate* quality. This study was downgraded due to a risk of bias around the reporting of analyses used. Overall rating for the CNS Vital Signs was *indeterminate* due to inconsistencies across study designs and outcomes which made it difficult to compare and come to an overall conclusion. Evidence for these studies was of *high* methodological quality.

Table 10. Summary of Findings for Hypotheses Testing

Hypotheses testing	**Summary**	**Rating of summarised evidence**	**Graded overall quality of evidence**
BI Alert	Result is in accordance with hypothesis for convergent validity (compared with CBCL and TRF)	Sufficient	Moderate (One study of adequate quality)
CNS Vital Signs	Inconsistent results across studies using population of different severity	Indeterminate	High (multiple studies of at least adequate quality)
CALS	Results are in accordance with the hypothesis - strong	Sufficient	High (one study of very good quality and >100 participants)

Hypotheses testing	Summary	Rating of summarised evidence	Graded overall quality of evidence
	correlations between CALS and WeeFIM		
LANSE-C/A	Different methods used in each study (subgroup vs other measurement); 13 of the 14 subtests reached statistical significance at the P < .001 level when comparing sub-groups; appears WeeFIM was a poor fit	Sufficient	High (At least one study of very good quality)
S-FAVRES	Those with ABI obtained statistically lower scores than TD group; 75% of results were in accordance with hypothesis	Sufficient	High (at least one study of very good quality)

Responsiveness

Two studies of responsiveness were found for two of the five screens; CALS (1) and the LANSE-C/A (1) (see Table 11). Both studies compared the screen with another measurement instrument. Results for the CALS were *sufficient* in a *high* quality study. Results for the LANSE-C/A were *insufficient* in a *low* quality study. This study was considered to have a serious risk of bias given queries around the adequacy of the measurement instrument used for comparison, and was downgraded further due to the sample size of <100.

Table 11. Summary of Findings for Responsiveness

Responsiveness	Summary	Rating of summarised evidence	Graded overall quality of evidence
CALS	Results are in accordance with the hypothesis - strong correlations between CALS and WeeFIM between admission and discharge	Sufficient	High (one study of very good quality and >100 participants)
LANSE-C/A	Did not meet 75% criteria for convergent validity – possibly poor instrument comparison	Insufficient	Low (One study of adequate quality; downgraded due to small sample size)

DISCUSSION

The aims of the review were to identify cognitive screening assessments that are used within paediatric TBI and appraise and synthesise their psychometric properties against the COSMIN criteria. There were six measures identified that have been developed to screen for cognitive impairments within this population. All six screens were developed and administered in the English language. Levels of evaluation conducted across measurement properties varied, with one of the biggest gaps being around screen development and structural validity which provide the foundation for interpreting other measurement properties. Below is a summary of the overall outcomes from the systematic literature review and its limitations.

Summary of evidence

Study Samples

A greater number of males than females in TBI samples reflect the literature on paediatric TBI, suggesting that samples may be representative of this population in terms of gender (Andersson, Sejdhage & Wage, 2012). Previous systematic reviews of cognitive screening tests included only studies where the target population made up at least 50% of the overall sample being studied (Burton and Tyson, 2015). The research team decided to include papers where there were smaller sub-samples of paediatric TBI for two main reasons (1) where the screen was developed specifically for paediatric neurology samples and grounded in relevant empirical data and (2) due to the lack of available studies on the test with larger samples.

Cognitive Screening

Only one of the six measures, the LANSE-C/A, met all three factors used to define cognitive screening tests; brief, sensitive to mild cognitive impairment and covering multi cognitive domains (Burton & Tyson, 2015). The LANSE-C/A was reported to take 20-30 minutes to complete; covering a broad range of cognitive domains (Orientation; Attention; Language; Reasoning; Memory; Visual Perception; and Praxis); and had *sufficient* results in a study of *high* quality for sensitivity (Lebby et al., 2015). The significant variance across administration timings for the BI Alert and S-FAVRES may be of concern for clinicians as cognitive screens are meant to be brief in nature and less resource intensive when compared to larger neuropsychological batteries (Cordell et al., 2013). Further explanation on the variance of administration timings would inform the feasibility of using these screens in clinical settings.

Out of the six measures only the LANSE-C/A and S-FAVRES had studies with *sufficient* evidence of *high* methodological quality for sensitivity (Lebby et al., 2015;

MacDonald, 2015). Results for the CNS Vital Signs were *indeterminate* due to inconsistency in some findings, possibly due to differences in study designs with differing levels of injury severity and time after injury (Brooks et al., 2016; Plourde & Brooks, 2017). Sensitivity is a key factor in ensuring that paediatrics in need of further assessment and intervention are not missed in the screening process (Burton & Tyson, 2015).

There was variance across the screening tests in relation to the constructs they were developed to measure, despite all measures aiming to screen for multi-domain cognitive impairments. This made it difficult to directly compare the cognitive screens. There was some overlap, with four of the screens assessing aspects of memory (verbal; visual; auditory), and five of them assessing aspects of executive function skills (cognitive flexibility; attention; inhibition), alongside a range of other cognitive domains. The broad range of cognitive skills assessed by all six measures was a strength. However, the proposed constructs and their relevant subscales can only be considered in context of the quality of studies on content and structural validity. Study findings were summarised across measures under the three broad areas covered by COSMIN – content validity, internal structure, and remaining measurement properties.

Content Validity

Content validity is considered the most important factor within the COSMIN methodology (Mokkink et al., 2018; Terwee et al., 2018). Content validity refers to both the development process of the cognitive screen and the extent to which the items or subscales adequately reflect the cognitive domains being measured. The overall rating of studies on screen development were *doubtful* for five of the six cognitive screens identified (BI Alert; CNS Vital Signs; CALS; S-FAVRES; and The Pediatric Test of Brain Injury), and *inadequate* for the LANCE-C/A. The weaknesses of these studies were predominantly due to underreporting of methodology on the item selection process, context and target population, and poor definitions of the constructs to be measured. The BI-Alert and S-FAVRES both had content validity studies of *sufficient* quality based on evidence of studies with parents and professionals (Rasquin et al., 2011). These studies were of *moderate* quality due to the methodology used, highlighting the need for caution around interpretability of the results. Findings on content validity for the Pediatric Test of Brain Injury were inconsistent. The test has strong theoretical underpinnings, without indication of an assessment of relevance, comprehensiveness and comprehensibility from the perspective of patients and professionals.

Reporting on screen development and content validity was generally inconsistent, with some studies failing to report the details of pilot studies and the overall process of item selection. These differences raise questions as to whether there was publication bias. Some studies allude to a process involving professionals or patients but do not describe the methods used and how this informed further changes. In some cases, there is an assumption made by developers that the screen is adequately measuring the proposed construct that they are basing their hypotheses on. This underreporting of methodology and results may reflect a lack of scientific rigor. It appears clinicians would have to use their clinical judgement in choosing the most appropriate tool to screen for cognitive difficulties and the domains they cover.

Internal Structure

CNS Vital Signs was the only screen to have *sufficient* results for structural validity and the overall evidence was of moderate quality due to inconsistencies in results across studies (Gualtieri & Hervey, 2015). The biggest downfall across studies of structural validity was the use of exploratory factor analysis, which meant that findings could not be interpreted as conclusive. The lack of sufficient studies on structural validity also meant the author was unable to interpret results on internal consistency, as it is not clear whether the screen is an adequate reflection of the cognitive domains it proposes to measure. Against the COSMIN criteria, the use of Pearson's/Spearman's correlation rather than ICC meant that studies of reliability for the BI Alert and CNS Vital Signs were indeterminate (Mokkink et al., 2018). While Pearson's correlation ensures that the order of results correlate across time, they do not account for an overall change in scores. ICC is a more robust test to use for this type of analysis and is required to meet COSMIN quality criteria. Further research is needed on internal structure before a conclusive interpretation can be made.

Remaining Measurement Properties

Evidence on test reliability was inconclusive as the available studies were of *low* or *very low* quality, due to the analyses used and small sample sizes. As previously mentioned, the LANSE-C/A and S-FAVRES both had *sufficient* evidence for criterion validity, both meeting the 'gold-standard' for sensitivity to cognitive impairment. The results were from studies of *high* methodological quality suggesting that the evidence is trustworthy. There were several studies reporting hypothesis testing for construct validity. There is strong evidence to suggest that three of the six measures are valid in measuring the constructs they are designed to assess. The CALS had high level evidence of *sufficient* convergent validity when comparing the cognitive screen with the WeeFIM (Ottenbacher et al., 1996). Whilst the LANSE-C/A and S-FAVRES had high level evidence of *sufficient* known-groups validity. The CALS was the only cognitive screen with high level evidence of *sufficient* responsiveness, indicating its ability to measure change over time. Responsiveness of a measure is particularly important when used as a repeated measure in clinical settings.

Clinical Recommendations

In light of the available evidence-base, the six cognitive screening tests were categorised under the following criteria, extracted from the COSMIN manual on guidelines for recommending OMs (Mokkink et al., 2018, Pg. 45):

a) Evidence of sufficient content validity and at least low quality evidence for sufficient internal consistency
b) Cognitive screens not in category A or C
c) Cognitive screens with high quality evidence for an insufficient measurement property

Cognitive screening tests in category A are to be recommended for use in clinical and research settings, while category C are not to be recommended. In the case where no cognitive screens fall under category A, category B measures should be highlighted as having the potential to be appropriate for clinical utility and therefore recommended on a provisional basis. None of the six cognitive screens met the criteria for category A or C. The BI-Alert was the only screen with evidence of sufficient content validity in a study of moderate quality, while evidence on the other five screens were inconsistent or inadequate. The BI-Alert, CALS and S-FAVRES had indeterminate results for internal consistency, while no studies were found for the remaining three. None of the six cognitive screens had "high quality evidence for an insufficient measurement property". All six cognitive screens should be recommended under category B, as provisionally appropriate for clinical utility on the basis that more robust research will be conducted. Tests may be distinguished from each other to suit specific clinical needs based on a number of factors; sensitivity to cognitive impairment, the cognitive domains the screen assesses, and the administration time.

Implications for Research

Some of the basic recommendations for studies on the psychometric properties of cognitive screening tests are around the quality of the reporting. The quality of the evidence was downgraded potentially due to omission of important details or poor definitions of constructs and processes. Arguably, these details have been omitted purposely and downgrading due to risk of bias may be in fact accurate.

Further research into the sensitivity of cognitive screening measures for paediatric TBI would be of benefit. This is particularly important in giving clinicians confidence in their use, by meeting the core criteria of a screen in reducing the risk of type-II errors. Understanding factors that may impact on administration time may shed light on the significant variance seen in studies for cognitive screens. Again, this would inform questions around the feasibility of their use in clinical settings. Research on cognitive screening tests would benefit from following the COSMIN process in their development and studies of psychometric properties (Gorst et al., 2020). Focusing firstly on content validity and internal consistency in order to provide a strong foundation to build further evidence on.

Limitations

There are a number of limitations within the systematic review which must be considered when interpreting study findings. Generally, the methodological quality of available evidence on cognitive screens for paediatric TBI was inconsistent. This reflected findings from the wider literature on paediatric TBI (Babikian and Asarnow, 2009). As a result, it was difficult for the researcher to come to definitive conclusions and recommendations around the selection and use of these cognitive screens in clinical practice. It was not appropriate to conduct a meta-analysis due to the differences across study methodologies. Due to a lack of available studies, the overall ratings and grading for some screens was based on the quality of only one available study.

The eligibility criteria and search strategy may have resulted in important studies being missed. The selection of papers published only in English due to the capacity of the author is an example of this. Retrieval of all identified research was incomplete. The researcher was unable to locate contact details of Lebby and Asbell, the developers of the LANSE-A/C, for further information on test development and content validity which is believed to be outlined in their Book “The Source for TBI – Children and Adolescents”

(Lebby & Asbell, 2007). This book was not available through the university library or their university contacts and the researcher did not have the budget to buy the book. This meant that the results from analysis on the LANSE-A/C may be limited due to a lack of available information rather than poor test development. This highlights the need for caution when interpreting the findings, as there are potential gaps due to missing data.

COSMIN guidelines recommend having a team of researchers with a strong knowledge of assessment development and psychometric properties. This study was conducted by a doctoral research student and supported by academic and clinical researchers with expertise in the field. While every effort was taken at each step of the process, the review is based on the interpretation of data by a small team. Further critique of the review process and interpretation would strengthen the overall quality of the systematic literature review. It is important to consider the potential risk of publication bias, due to underreporting of studies with statistically insignificant findings; particularly in the case where there is a conflict of interest when test developers are involved in the research or the research is being funded.

Conclusions

There are a number of cognitive screening tests for paediatric TBI available for use in clinical practice and research settings. The six identified screens are recommended for clinical use given the existing evidence, on the basis that they are used with caution and clinical judgement until further research is conducted. There was no evidence of particular concern which would lead to a recommendation of exclusion from clinical use. COSMIN checklist is a relatively new expert driven quality tool for assessing the development and psychometric properties of outcome measures. It is a thorough tool with strict standards and criteria. Further research on existing cognitive screening tests would benefit from using the COSMIN process as a basis for developing a high quality tool with robust psychometric properties. It is difficult to make clinical recommendations on the use of existing tools due to the scarcity of studies and the quality of available evidence.

Funding

There was a small research budget available to support the study. Databases were systematically searched and data extracted independently by the researcher (TM).

REFERENCES

American Psychological Association (APA). (2020). *Publication Manual of the American Psychological Association.* (7th Edition). Washington: American Psychological Association.

Andersson, E. E., Sejdhage, R., & Wage, V. (2012). Mild traumatic brain injuries in children between 0–16 years of age: A survey of activities and places when an accident occurs. *Developmental Neurorehabilitation, 15(1),* pp. 26-30. https://doi.org/10.3109/17518423.2011.633570

Babikian, T., & Asarnow, R. (2009). Neurocognitive outcomes and recovery after pediatric TBI: meta-analytic review of the literature. *Neuropsychology, 23(3),* pp. 283-296. https://psycnet.apa.org/doi/10.1037/a0015268

Braun, M., Tupper, D., Kaufmann, P., McCrea, M., Postal, K., Westerveld, M., Wills, K. & Deer, T. (2011). Neuropsychological assessment: A valuable tool in the diagnosis and management of neurological, neurodevelopmental, medical, and psychiatric disorders. *Cognitive and Behavioral Neurology, 24(3),* pp. 107-114. doi: 10.1097/WNN.0b013e3182351289

Brooks, B. L., Daya, H., Khan, S., Carlson, H. L., Mikrogianakis, A., & Barlow, K. M. (2016). Cognition in the emergency department as a predictor of recovery after pediatric mild traumatic brain injury. *Journal of the International Neuropsychological Society*, *22*(4), pp. 379-387. https://doi.org/10.1017/S1355617715001368

Brooks, B. L., Khan, S., Daya, H., Mikrogianakis, A., & Barlow, K. M. (2014). Neurocognition in the emergency department after a mild traumatic brain injury in youth. *Journal of Neurotrauma*, *31*(20), pp. 1744-1749. https://doi.org/10.1089/neu.2014.3356

Brooks, B. L., Plourde, V., Fay-McClymont, T. B., MacAllister, W. S., & Sherman, E. M. (2019). Factor structure of the CNS Vital Signs computerized cognitive battery in youth with neurological diagnoses. *Child Neuropsychology*, *25*(7), pp. 980-991. https://doi.org/10.1080/09297049.2019.1569609

Brooks, B. L., & Sherman, E. M. (2012). Computerized neuropsychological testing to rapidly evaluate cognition in pediatric patients with neurologic disorders. *Journal of Child Neurology, 27*(8), pp. 982-991. https://doi.org/10.1177/0883073811430863

Burton, L., & Tyson, S. F. (2015). Screening for cognitive impairment after stroke: a systematic review of psychometric properties and clinical utility. *Journal of Rehabilitation Medicine*, 47(3), pp. 193-203. doi: 10.2340/16501977-1930.

Chan, V., Thurairajah, P., & Colantonio, A. (2015). Defining pediatric traumatic brain injury using International Classification of Diseases Version 10 Codes: a systematic review. *BMC neurology*, *15*(1), pp. 1-26.

Cordell, C. B., Borson, S., Boustani, M., Chodosh, J., Reuben, D., Verghese, J., Thies W., Fried, L.B. & Medicare Detection of Cognitive Impairment Workgroup. (2013). Alzheimer's Association recommendations for operationalizing the detection of cognitive impairment during the Medicare Annual Wellness Visit in a primary care setting. *Alzheimer's & Dementia*, *9*(2), pp. 141-150. https://doi.org/10.1016/j.jalz.2012.09.011

Di Battista, A., Soo, C., Catroppa, C., & Anderson, V. (2012). Quality of life in children and adolescents post-TBI: a systematic review and meta-analysis. *Journal of Neurotrauma*, *29(9)*, pp. 1717-1727. https://doi.org/10.1089/neu.2011.2157

Dewan, M. C., Mummareddy, N., Wellons III, J. C., & Bonfield, C. M. (2016). Epidemiology of global pediatric traumatic brain injury: qualitative review. *World neurosurgery*, 91, pp. 497-509. https://doi.org/10.1016/j.wneu.2016.03.045

Emery, C. A., Barlow, K. M., Brooks, B. L., Max, J. E., Villavicencio-Requis, A., Gnanakumar, V., ... & Yeates, K. O. (2016). A systematic review of psychiatric, psychological, and behavioural outcomes following mild traumatic brain injury in children and adolescents. *The Canadian Journal of Psychiatry*, 61(5), pp. 259-269. https://doi.org/10.1177%2F0706743716643741

Gorst, S. L., Prinsen, C. A., Salcher-Konrad, M., Matvienko-Sikar, K., Williamson, P. R. & Terwee, C. B. (2020). Methods used in the selection of instruments for outcomes included in core outcome sets have improved since the publication of the COSMIN/COMET guideline. *Journal of Clinical Epidemiology*, pp. 64-75. https://doi.org/10.1016/j.jclinepi.2020.05.021

Gualtieri, C. T., & Hervey, A. S. (2015). The structure and meaning of a computerized neurocognitive test battery. *Frontiers*, *4*(2), pp. 11-21.

Gualtieri, C. T., & Johnson, L. G. (2006). Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Archives of Clinical Neuropsychology*, *21*(7), pp. 623-643. https://doi.org/10.1016/j.acn.2006.05.007

Hotz, G., Helm-Estabrooks, N., & Nelson, N. W. (2001). Development of the pediatric test of brain injury. *The Journal of head trauma rehabilitation*, *16*(5), pp. 426-440.

Hotz, G. A., Helm-Estabrooks, N., Nelson, N. W., & Plante, E. (2009). The pediatric test of brain injury: Development and interpretation. *Topics in Language Disorders*, *29*(3), pp. 207-223. doi: 10.1097/TLD.0b013e3181b53198

Kahn, D. A., Asbell, S. J., & Donders, J. (2015). Clinical utility of the LANSE-A in adolescents with traumatic brain injury. *Rehabilitation psychology*, *60*(2), pp. 187-192. https://psycnet.apa.org/doi/10.1037/rep0000034

Keenan, H. T., & Bratton, S. L. (2006). Epidemiology and outcomes of pediatric traumatic brain injury. Developmental neuroscience, 28(4-5), pp. 256-263. https://doi.org/10.1159/000094152

Lebby, P. C. & Asbell, S. J. (2007). The source for traumatic brain injury children and adolescents. *East Moline, Illinois: LinguiSystems Inc*, pp. 40-65.

Lebby, P. C., Pollock, M., Mouanoutoua, A., & Lewey, J. H. (2015). Performance of children and adolescents with brain injury on the Lebby-Asbell neurocognitive screening examination for children and adolescents. *Journal of child neurology*, *30*(10), pp. 1255-1262. https://doi.org/10.1177/0883073814556886

Lee, L. K. (2007). Controversies in the sequelae of pediatric mild traumatic brain injury. *Pediatric Emergency Care*, 23(8), pp. 580-583. doi: 10.1097/PEC.0b013e31813444ea

Loring, D. W. & Bowden, S. C. (n.d.). Neuropsychology Review. *Springer*. Retrieved January 12th, 2021, from https://www.springer.com/journal/11065.

MacDonald, S. (2015). Assessment of higher level cognitive-communication functions in adolescents with ABI: Standardization of the student version of the functional assessment of verbal reasoning and executive strategies (S-FAVRES). *Brain Injury*, *30*(3), pp. 295-310. https://doi.org/10.3109/02699052.2015.1091947

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International Journal of Surgery*, *8*(5), pp. 336-341. https://doi.org/10.1371/journal.pmed.1000097

Mokkink, L. B., Prinsen, C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C., & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of patient reported outcome measures (PROMs). *User manual*, 78, pp. 1-78.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M. & De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of life research*, *19*(4), pp. 539-549. https://doi.org/10.1007/s11136-010-9606-8

Morley, J. E., Morris, J. C., Berg-Weger, M., Borson, S., Carpenter, B. D., del Campo, N., Dubois, B., Fargo, K., Fitten, J., Flaherty, J.H., Ganguli, M., Grossberg, G.T., Malmstrom, T.K., Petersen, R.D., Rodriguez, C., Saykin, A.J., Tangalos. E.G., Verghese, J., Wilcock, G., Winblad, B., Woo, J. & Vellas, B. (2015). Brain health: the importance of recognizing cognitive impairment: an IAGG consensus conference. *Journal of the American Medical Directors Association*, *16*(9), pp. 731 739. https://doi.org/10.1016/j.jamda.2015.06.017

National Center for Injury Prevention and Control, Division of Injury Response. (2010). *Traumatic brain injury in the United States; emergency department visits, hospitalizations, and deaths.* (2002-2006). United States: Centers for Disease Control (CDC).

National Institute for Health Research (NIHR) (March 19th 2021). The psychometric properties of screening tests for cognitive impairment in paediatric traumatic brain injury (TBI) populations: a systematic review. *PROSPERO: International prospective register of systematic reviews.* Retrieved April 4th, 2021, from https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021238163.

Ottenbacher, K. J., Taylor, E. T., Msall, M. E., Braun, S., Lane, S. J., Granger, C. V., Lyons, N. & Duffy, L. C. (1996). The stability and equivalence reliability of the functional independence measure for children (WeeFIM)®. *Developmental Medicine & Child Neurology*, *38*(10), pp. 907-916. https://doi.org/10.1111/j.1469-8749.1996.tb15047.x

Pantaleon, L. (2019). Why measuring outcomes is important in health care. *Journal of veterinary internal medicine*, *33*(2), pp. 356-362. https://doi.org/10.1111/jvim.15458

Parslow, R. C., Morris, K. P., Tasker, R. C., Forsyth, R. J., & Hawley, C. (2005). Epidemiology of traumatic brain injury in children receiving intensive care in the UK. *Archives of disease in childhood*, 90(11), pp. 1182-1187. http://dx.doi.org/10.1136/adc.2005.072405

Petersen, R. C., Stevens, J. C., Ganguli, M., Tangalos, E. G., Cummings, J. L. & DeKosky, S. T. (2001). Practice parameter: early detection of dementia: mild cognitive impairment (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, *56*(9), pp. 1133-1142. https://doi.org/10.1212/WNL.56.9.1133

Plourde, V., & Brooks, B. L. (2017). Is computerized cognitive testing useful in children and adolescents with moderate-to-severe traumatic brain injury?. *Journal of the*

International Neuropsychological Society, *23*(4), pp. 304-313. doi:10.1017/S1355617717000066

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

Quinlivan, A., Thakkar, V., Stevens, W., Morrisroe, K., Prior, D., Rabusa, C., Youssef, P., Gabbay, E., Roddy, J., Walker, J.G., Zochling, J., Sahhar, J., Nash, P., Lester, S., Rischmueller, M., Proudman, S.M. & Nikpour, M. (2015). Cost savings with a new screening algorithm for pulmonary arterial hypertension in systemic sclerosis. *Internal Medicine Journal*, *45*(11), pp. 1134-1140. https://doi.org/10.1111/imj.12890

Rasquin, S., van Heugten, C., Winkens, I., Ritzen, W., Hendriksen, J., & Vles, H. (2011). Development and validity of the Brain Injury Alert (BI Alert) screening tool for cognitive, emotional and social problems after paediatric acquired brain injury. *Brain Injury*, 25(7-8), pp. 777-786. https://doi.org/10.3109/02699052.2011.580311

Rosenkoetter, U., & Tate, R. L. (2018). Assessing features of psychometric assessment instruments: A comparison of the COSMIN checklist with other critical appraisal tools. *Brain Impairment*, *19*(1), pp. 103-118. doi:10.1017/BrImp.2017.29

Slomine, B., Eikenberg, J., Salorio, C., Suskauer, S., Trovato, M., & Christensen, J. (2008). Preliminary evaluation of the Cognitive and Linguistic Scale: a measure to assess recovery in inpatient rehabilitation following pediatric brain injury. *The Journal of*

Head Trauma Rehabilitation, *23*(5), pp. 286-293. doi: 10.1097/01.HTR.0000336841.53338.2f

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

Thurman, D. J. (2016). The epidemiology of traumatic brain injury in children and youths: a review of research since 1990. *Journal of Child Neurology, 31(1),* pp. 20-27. https://doi.org/10.1177/0883073814544363

Van Heugten, C. M., Hendriksen, J., Rasquin, S., Dijcks, B., Jaeken, D., & Vles, J. H. S. (2006). Long-term neuropsychological performance in a cohort of children and adolescents after severe paediatric traumatic brain injury. *Brain Injury, 20(9),* pp. 895 -903. https://doi.org/10.1080/02699050600832015

World Health Organization (WHO). (2006). Neurological disorders: public health challenges. World Health Organization.

Yeates, K.O. & Taylor, H. G. (2005). Neurobehavioural outcomes of mild head injury in children and adolescents. *Pediatric Rehabilitation, 8(1),* pp. 5-16. https://doi.org/10.1080/13638490400011199

Yeates, K. O., Taylor, H. G., Wade, S. L., Drotar, D., Stancin, T., & Minich, N. (2002). A prospective study of short-and long-term neuropsychological outcomes after traumatic brain injury in children. *Neuropsychology, 16(4),* pp. 514. https://psycnet.apa.org/doi/10.1037/0894-4105.16.4.514

VALIDATION OF THE TRIANGLES THEORY OF MIND TASK IN YOUNG SCHOOL-AGED CHILDREN WITH EARLY EMERGING BEHAVIOURAL AND SOCIAL-INTERPERSONAL PROBLEMS

Author name: Treasa Murphy

Affiliation(s) of the author:

Doctorate of Clinical Psychology Programme, School of Psychology, Cardiff University, Cardiff, Wales, United Kingdom.

Cardiff and Vale University Health Board, National Health Service (NHS) Wales, Cardiff, Wales, United Kingdom.

Word count: 7,165 (excluding abstract, references, figures and tables)

Author Note

This paper was written for submission to the peer-reviewed journal The British Journal of Developmental Psychology (British Psychological Society, n.d.) (see Appendices E). In line with submission guidelines for the journal, this paper was prepared by the author in the style of the American Psychological Association Publication Manual (APA, 2020). The word count of 8,000 words is based on guidance from the Cardiff DClinPsy programme on thesis submission.

ABSTRACT

We examined the validity of a computerised theory of mind (ToM) task (Triangles) in children (aged 4-7 years) identified as presenting with early emerging behavioural and/or social-interpersonal problems. Comparison with traditional ToM tests were conducted to assess: convergent validity; whether Triangles was less demanding on skills such as language and executive function; and whether performance was associated with social-interpersonal skills. N = 55 children (mean age = 6.28 years; 42% female) participated in the study. Correlation coefficients assessed the strength of associations and Fisher's Z transformations and z-tests were used to test significance of difference between associations. Significant positive associations were observed for convergent validity. Partial support was found for hypotheses on language and executive functioning. Contrary to hypotheses, children with more severe conduct problems performed better on the Triangles task. Future research would benefit from using more generalised tests of language ability, larger samples and longitudinal study designs.

Keywords: *Theory of Mind; School-age; Behavioural problems; Social Problems*

Data availability statement

Data was provided for the current study by the Neurodevelopmental Assessment Unit (NDAU), Cardiff University. The data are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Funding: The author was funded by Health Education and Improvement Wales.

Conflicts of Interest: There are no conflicts of interest to be declared.

VALIDATION OF THE TRIANGLES THEORY OF MIND TASK IN YOUNG SCHOOL-AGED CHILDREN WITH EARLY EMERGING BEHAVIOURAL AND SOCIAL-INTERPERSONAL PROBLEMS

Theory of Mind (ToM) is a construct of social cognition relevant to human development (Quesque & Rossetti, 2020). Research on ToM spans over 30 years and has highlighted the complexity of this construct. ToM might be an important transdiagnostic risk factor for social and behaviour problems across various disorders (Wells, Hunnikin, Ash & Van Goozen, 2020). Therefore, research into measuring ToM is important, as it might contribute to improving early detection and intervention for vulnerable children. Traditional ToM tasks have been criticised for their demands on cognitive processes such as language and executive function (EF) skills (McAlister & Peterson, 2013). Questions have been raised as to whether Traditional ToM tasks are measuring the construct in its purest form and whether they are suitable for children with language and EF difficulties (Beaudoin, Leblanc, Gagner & Beauchamp, 2020). The Triangles task (Abell et al., 2000; Castelli, Frith, Happé & Frith, 2002) was developed to address some of these limitations. The aim of this study was to explore the validity of the Triangles task in young-school-aged children with early emerging behavioural and social-interpersonal problems.

Theory of Mind

ToM refers to understanding that others may hold different beliefs to ourselves, the ability to make inferences about others' beliefs, feelings and desires, and in turn understand and predict their behaviour (Abell, Happe & Frith, 2000). ToM is established as a multi-faceted phenomenon. Cognitive and affective ToM have been distinguished from each other in the literature (Baldimtsi, Nicolopoulou & Tsimpli, 2021; Sebastian et al., 2012). Cognitive ToM is the ability to understand another person's thought processes, while affective ToM is the ability to understand how someone else is feeling. Kalbe et al. (2010) further supported this distinction through brain imaging research with adults. They reported activity in different neural regions during thought focused ToM tasks (right dorsolateral prefrontal cortex) compared to emotion focused ToM tasks. First Order ToM is the ability to understand the perspective of another, by attributing thoughts to their actions or predicting their actions based on this understanding (Frith & Frith, 2005). Second Order ToM is a more complex skill, which involves understanding what someone thinks a third person's perspective is.

ToM Development

ToM is a universal construct with a similar trajectory across cultures (Liu, Wellman, Tardif & Sabbagh, 2008; Wellman, Cross & Watson, 2001). It appears to have a developmental progression, whereby accuracy on ToM tasks improves with age (Wellman, 2011). Gender has also been reported as a significant factor, with girls performing better than boys in middle-childhood (Calero, Salles, Semelman & Sigman, 2013). Studies involving children as young as 6 months old have demonstrated the emergence of social intuition. Through joint attention and social referencing, infants make attempts to understand others (Mireault, Crockenberg, Sparrow, Pettinato, Woodard & Malzac, 2014). While a child is not yet able to verbalise their understanding, the appropriate use of social cues suggests an implicit understanding. In their meta-analysis of ToM development, Wellman et al. (2001) concluded that first order ToM is likely to develop between 3 to 4 years of age. Second-order ToM understanding has been estimated to emerge later, at around 6 to 7 years of age, in studies of typically developing children (Dumontheil, Apperly & Blakemore, 2010; Peterson, Slaughter & Wellman, 2018).

Over and above age, language ability is strongly associated with ToM accuracy (Milligan, Astington & Dack, 2007). Language is an umbrella term spanning verbal and non-verbal processes of communication. The association between ToM and language ability has been found across samples of both clinical and typically developing children (Abbeduto, Short-Meyerson, Benson & Dolish, 2004; Bailey & Im-Bolter, 2020; Ebert, 2020). Possible explanations for this relationship may be the verbal demands of ToM tasks, the need to use language processes to think about the minds of others, or alternatively language may provide more tangible representations of abstract concepts such as mental states (Ebert, 2020; Low & Simpson, 2012). Theorists suggest that language is an integral part of ToM development, noting the importance of semantics (attachment of meaning to words) and syntactic language (sentence formation) in order to give meaning and make sense of abstract concepts such as beliefs (Bailey & Im-Bolter, 2020). In their meta-analysis of 104 papers (n=8,891) Milligan et al. (2007) explored whether the type of language test used or the type of false belief task influenced the language-ToM association in children under 7 years of age. They found significant associations between language ability and ToM performance across all five language domains (general language; semantics; receptive vocabulary; syntax; and memory for complements). Furthermore, general language tests (Test of Early Language Development; Test for Auditory Comprehension of Language) were more strongly related to ToM performance compared to specific measures of receptive language (British Picture Vocabulary Scale; Peabody Picture Vocabulary Test). The type of false belief task did not influence whether there was a relationship between language and ToM performance. These findings suggest that there may be multiple aspects of linguistics associated with performance in false belief tasks, explaining the strength of general language tests during early childhood.

More recent studies have demonstrated the continuation of the ToM-language relationship in to middle-childhood. Ebert (2020) conducted a longitudinal study of the ToM-language association in a group of children from preschool to adolescence (n=231; age = 3years 6 months to 13 years 7 months). Reflecting earlier studies, language (receptive grammar; receptive vocabulary; text comprehension) was a predictor of ToM, and these findings were moderately consistent over time. A large proportion of the ToM-language research has been conducted using false-belief tasks, potentially highlighting their particular sensitivity to language ability. While the exact nature of the ToM-language relationship is not fully understood, it seems reasonable to predict that ToM tasks, particularly false-belief tasks, will positively correlate with tests of language ability in children during early to middle childhood.

ToM Difficulties

Navigating social relationships and the wider social system often requires ToM (Bohl, 2015). Difficulties can lead to what are deemed to be inappropriate social behaviours, social and emotional problems, or in more extreme cases conduct problems (Austin, Bondü & Elsner, 2020; Wells, et al., 2020). Conversely, ToM ability has been positively associated with improved prosocial behaviours in children (Imuta, Henry, Slaughter, Selcuk & Ruffman, 2016). It is widely accepted that ToM difficulties are characteristic of autism spectrum disorders (ASD: Baron-Cohen, 2000). In profiling ToM difficulties in children with ASD, Rosello and colleagues (2020) identified a subgroup of children with poorer ToM skills. These children had significantly more difficulties with daily functioning, social and communication skills. Associations between age and ToM seen in typically developing children are not as strong in children with high functioning ASD, potentially signifying greater individuality in their developmental trajectories (Bal, Yerys, Sokoloff, Celano, Kenworthy, Giedd & Wallace, 2013). As in the general population, the positive association between language ability and ToM has been found in ASD samples (Leno et al., 2021). Given that ToM impairment is theorised to be a crucial risk factor for the development of social-interpersonal difficulties including ASD, it is important that there are pure methods of measuring ToM that do not simply reflect other abilities such as language.

How is ToM measured?

Traditional ToM Tasks

Traditional ToM tasks, such as the Sally-Anne test (Baron-Cohen, Leslie & Frith, 1985), the Coco-pops test (Baron-Cohen, 1991), and the Smarties Test (Perner, Frith, Leslie & Leekam, 1989), continue to be used in some clinical and research settings today (Aslıer, Aslıer, Kirkim & Güneri, 2020; Dhadwal, Najdowski & Tarbox, 2021; Senju, 2012). They are first order false-belief tasks, where an individual predicts the behaviour of another based on the false information they hold in contrast to their own knowledge. These tests have been validated in samples of typically developing children, as well as clinical samples (Grant, Grayson & Boucher, 2001).

Traditional false-belief tests have been criticised for being too simplistic in measuring what we now know to be a complex phenomenon (Abell et al., 2000; Beaudoin et al., 2020). A child either passes or fails, rather than sitting on a spectrum of ability. Also, traditional tasks are particularly reliant on receptive language ability and perhaps some other skills such as EF. The Sally-Anne test, for example, requires receptive language, attention and working memory to follow a narrative and problem-solve (Baron-Cohen et al., 1985). The task involves two dolls, one with a basket and one with a box. The child watches one doll place an object in the basket and leave. While they are gone the second doll moves the object to the box. The child is then asked to judge where the first doll would look for the object on returning. The EF skills of cognitive inhibition and working memory are required to keep in mind the information, and inhibit one's own belief in order to predict another's behaviour based on a false belief (Devine & Hughes, 2014). This means that young children, or children with atypical development whose language and EF skills may be limited, would struggle with these tasks. Therefore, poor performance on traditional ToM tasks could be attributed to impairments in language and EF as opposed to deficits in social cognition (McAlister &

Peterson, 2013). In their comparison of atypical and typically developing children, Kampis, Fogd and Kovács (2017) drew attention to non-verbal cognitive processes of ToM that may contribute to individual differences in assessment, including sustained attention, working memory and problem-solving. Critical appraisal of the early ToM literature led to the development of a multitude of ToM tasks, developed to reflect the multi-component construct and incorporating a developmental perspective (Devine & Hughes, 2013; Wellman, & Liu, 2004).

Triangles Task

The Triangles Task, also known as the Animated Shapes Task, is a first-order ToM test consisting of a set of short silent animations (Abell et al., 2000; Castelli et al., 2002). The animations involve geometric shapes moving around a computer screen in one of two conditions: (1) goal directed (GD) and (2) ToM. The test requires participants to view the shapes as having independent minds and describe the type of interactions they witness. It was designed to address limitations of more traditional tasks by increasing replicability and efficiency; reflecting a spectrum of ToM ability rather than viewing it as a binary construct (Livingston, Shah & Happé, 2019). The silent nature of the tasks and in the moment descriptions aim to reduce the demands on receptive language and EF skills. Participants' verbal descriptions are scored across three domains: appropriateness (accuracy); intentionality (the use of mental state language); and length (the number of descriptions provided) (Castelli et al., 2002). The subjectivity of this coding process has been criticised, leading researchers to develop objective scores of accuracy in the individual's response to multiple choice questions (White, Coniston, Rogers & Frith, 2011). This method however has its drawbacks, as it has inadvertently reverted to a language dominant assessment. The Triangles task has been established as a reliable task, with strong to almost perfect interrater agreement across studies (κ 0.92 – 0.96), and good re-test reliability in typically developing school-aged samples (9-11 years) and adults (Castelli et al., 2002; Shahrivar, Tehrani-Doost, Banaraki & Mohammadzadeh, 2020). It has been validated in large samples of children with a range of clinical presentations (ASD; Epilepsy; Schizophrenia) and ages (7-17 years) (Warrier & Baron-Cohen, 2018). Significantly worse accuracy scores have been found across samples of both children and adolescents with ASD using the Triangles task, while this was not the case for intentionality and length (Abell et al., 2000; Salter, Seigal, Claxton, Lawrence & Skuse, 2008); thus whilst references to mental states were used, their accuracy was poor. Despite some of its limitations, the Triangles task is appealing for use in younger clinical samples due to the intended reduced demands on language and EF skills.

Rationale for the Current Study

To summarise, ToM is of interest due to its links with social and behavioural problems. It is a transdiagnostic construct which with appropriate tests can be a measurable early risk factor for these difficulties, as well as a potential intervention target. Therefore, establishing valid and reliable methods to identify high risk children during early childhood (age 4-7) when ToM is at a crucial stage of development, is a priority for research and clinical practice. Traditional ToM tasks are sensitive to cognitive processes such as receptive language and EF skills, indicating that they may not be the purest measure of ToM. More importantly, these potential confounding factors are often impaired in young children with emerging behavioural and social-interpersonal problems. The Triangles task was designed to address some of the limitations of traditional tasks, particularly in terms of performance being less impacted by language and EF skills. The current study aims to validate the Triangles task in a sample of children predominantly aged 4-7 years, referred to a neurodevelopmental assessment research unit due to early emerging behavioural and social-interpersonal problems. The Triangles task has not yet been validated in this age group. Therefore, the aims of this project are: 1) to explore whether the Triangles task is less sensitive than traditional tasks to cognitive processes such as receptive language and EF skills (attention; working memory; inhibition control); and 2) to assess whether performance on the Triangles task has expected associations with social-interpersonal and behavioural problems (defined in this study as peer problems, prosocial behaviour problems, and conduct problems) in children during early-middle childhood.

Hypotheses

Based on the literature it is expected that:

1. Performance on the Triangles task will be positively associated with performance on traditional ToM tasks.

2. Accuracy, in the Triangles task, will be a better predictor of ToM difficulties in traditional tasks than intentionality and length.

3. Performance on the Triangles task will be less strongly associated with receptive language ability than traditional ToM tasks.

4. Performance on the Triangles task will be less strongly associated with executive function skills than traditional ToM tasks.

5. Performance on the Triangles task will be associated with behavioural problems and social-interpersonal skills in children (specifically, better ToM performance on the Triangles task will be associated with improved prosocial behaviour, fewer conduct problems, and fewer problems with peer relationships).

METHODOLOGY

The Neurodevelopmental Assessment Unit

The Neurodevelopmental Assessment Unit (NDAU; https://www.cardiff.ac.uk/research/explore/research-units/neurodevelopment-assessment-unit) is a trans-diagnostic assessment unit based in Cardiff University. NDAU's primary function is research, with the aim of informing early intervention and prevention in clinical practice. Early school aged children, predominantly four to seven years of age, are referred to NDAU where there are concerns around a child's development. These concerns may include cognitive, emotional and behavioural difficulties, and come under the umbrella of atypical neurodevelopment. Data for the current study had already been collected through NDAU, see Appendices F for NDAU full referral pack and consent forms. Following a referral, children were invited to the assessment unit to complete a battery of assessments, which were administered by a team of researchers over two days. Parents and teachers of the children were also invited to complete questionnaire and interview measures. The relevant measures for this study are detailed below.

Ethical Approval

The School of Psychology, Research and Ethics Committee at Cardiff University approved the following study: A Feasibility Study of a Neurodevelopmental Disorders Assessment Unit (EC.16.10.11.4592GRA5) (see Appendices G). The current study procedures were included in this approval.

Participant Sample

The participant sample included 55 children (42% female) aged between 4 and 8 years old (mean = 6.28; SD = 1.16). Children were selected on the basis that they were referred to NDAU for assessment due to emerging behavioural and social-interpersonal problems. They were included in the study if they completed the standard battery of assessments administered by the NDAU team, inclusive of the Triangles task.

Measures

Theory of Mind

The Triangles Task

The Triangles task is a test of first-order ToM (Abell et al., 2000; Castelli et al., 2002), which takes approximately 6 minutes to complete. Participants were asked to watch six video clips of two triangles, one large and one small, interacting on the screen of a laptop. They were prompted in real time to describe what they were seeing. It is intended that the shapes depict two characters. There are two conditions (1) goal directed (G-D) and (2) theory of mind (ToM). Of the six clips, two were G-D and four were ToM (one of which is a practice). The clips were shown to each participant in one of three orders. The two G-D conditions involve the triangles fighting and chasing each other. It was expected that these clips would evoke descriptions of an interaction between the two shapes. The four ToM clips involve; a surprise; one triangle seducing/persuading the other; the big triangle coaxing the little triangle outside; and mocking. It was expected that these clips would evoke descriptions of mental states such as the intention to trick someone.

Participants' verbal descriptions of the silent animations were video and audio recorded during the assessment. The audio clips were transcribed and scored by the researcher against the criteria outlined by Castelli, Happé, Frith and Frith (2000) and Abell et al. (2000). Each item was scored against three distinct scales (1) accuracy (0-2) (2) intentionality (0-5) and (3) length (0-4) (see Appendix H). Scores for the three ToM clips (excluding the practice clip) were summed to give a Total ToM score across the three dimensions. The sum of accuracy scores for the two G-D clips gave a total performance score. To ensure accuracy, the audio clips were double transcribed by an independent transcriber and an independent reviewer double scored a subset (20%) of participants to establish interrater reliability, the details of this process are outlined in the results section.

Traditional ToM Tasks

The traditional ToM tasks are made up of four separate tests of ToM, which take approximately 10 minutes to complete. The Smarties Identity Test (Perner et al., 1989; Wellman & Lui, 2004), the Cheerios task (Baron-Cohen, 1991) and the Sally-Anne task (Baron-Cohen et al., 1985) are all first-order false-belief tests. Participants were required to listen to a story supported visually by props for each test and asked to predict what a character would do based on the false information they hold. The fourth test, the second-order false belief test (Coull, Leekam & Bennett, 2006), involved participants following a narrative of two dolls hiding a teddy bear, and required them to demonstrate an understanding of what one character believes a second character would do based on the false-belief they hold.

Independently the four tasks were scored as pass or fail. Together they were categorised as “age expected” or “below age expected”. Categorisation, rather than a sum of scores, was used as it is a more meaningful representation of performance, compared to an artificial scale combining different tasks. Participants were grouped based on the following criteria; three correct first order tests or two correct first order and a correct second order test = age expected; two or less correct = below age expected. These cut-offs were based on the literature, where it is expected that first-order ToM would be developed in the study sample minimum age of 4 years old (Wellman, 2011).

Language

The British Picture Vocabulary Scale (BPVS) is a measure of receptive language ability (Dunn, Dunn, Whetton & Burley, 1997). It is designed for children aged 3 to 16 years and is suitable for non-verbal children. It was administered directly with participants, and took approximately 10 minutes to complete. For each item, participants were shown a set of four black and white pictures. They were asked to listen to a word and choose the picture that best depicted the word. The words span a range of categories including animals, actions and emotions. The BPVS provides standardised scores with 85 to 115 being within the age expected range. The BPVS is reported to have good reliability at 0.80 (3 to 7 years) and 0.81 (4 to 9 years) (Gathercole, Willis, Baddeley & Emslie, 1994).

Executive Function

NIH Toolbox Dimensional Change Card Sort (DCCS) is a measure of EF (Zelazo, Anderson, Richler, Wallner-Allen, Beaumont & Weintraub, 2013). The computerised test requires skills such as attention, working memory, inhibition control and cognitive flexibility. Each child was asked to complete a series of matching tasks under three conditions (1) matching shapes (2) matching colours and (3) random switching between matching shapes and colours. The test requires the child to attend to stimuli on a computer screen (pictures of rabbits and boats), retain the instructions and task rules in order to match the correct items, to inhibit the incorrect matching information and to be able to switch between rules. The test provides age-corrected scores which can be used to profile children as below average (<85) average (85-114) and above average (115+) in ability.

Social-Interpersonal Skills and Behavioural Problems

The Strengths and Difficulties Questionnaire (SDQ) is a measure of emotional, social and behavioural skills in children aged between 4 and 17 years (Goodman, 1997). It has 25 questions across five subscales (1) Emotional Symptoms (2) Conduct Problems (3) Hyperactivity/Inattention (4) Peer Relationship Problems and (5) Prosocial. The SDQ was completed by both parents and teachers. For each statement they were asked to rate how their child presents on a scale of; ‘Not True’, ‘Somewhat True’ and ‘Certainly True’. Individual scales have a total sum of 10. Scores can be used to categorise children as falling within the ‘close to average’, ‘slightly raised/slightly lowered’, ‘high/low’ or ‘very high/very low’ range (see Appendices I).

Data Analysis

Data were analysed using Statistical Package for the Social Science software (IBM SPSS 26). Kappa statistics were used to assess interrater reliability for each Triangles scale score between two independent assessors. Where outliers were found in the data, the empirical rule of correcting scores to three standard deviations from the norm was used (Field, 2013). This was applied to one case for the BPVS and two cases for the DCCS. Descriptive statistics, QQ Plots and tests of normality were used to test parametric assumptions.

Bivariate Kendall's correlation coefficients were used to test for associations between study variables. A non-parametric test was chosen due to the dataset including ordinal data (traditional ToM tasks). Bonferroni correction was considered in correlation tests of multiple independent variables, due to the increased risk of Type I errors. However, in consideration of the planned hypotheses and the implications for risk of Type II errors in a preliminary validation study it was not used (Armstrong, 2014). In tests where variable scores were age adjusted, age was not included as a covariate.

Due to significant associations with medium to large effect sizes between the Triangles scales (particularly accuracy and intentionality) regression analysis was deemed a poor fit when testing for independent predictors. Instead, to test for significance of difference between dependent correlations, Kendall's τ coefficients were converted to Fisher's Z scores using a two-step process outlined by Walker (2003) and compared statistically using asymptotic z-tests (Lee & Preacher, 2013; Steiger, 1980). The associations between variables were accounted for in this process. As the sample size in this study was fixed, post-hoc power analyses were conducted using G*Power software to inform further research in this area.

RESULTS

Descriptive Statistics

Descriptive statistics on sample demographics, ToM, language, executive functioning, social-interpersonal skills and behavioural problems are presented in Table 1.

Table 1. Descriptive Statistics

Total Sample (n = 55)		**N (%)**
Gender (n = 54)		
	Male	31 (56.4%)
	Female	23 (42.6%)
		Mean (Range)
Age (months) (n = 55)		75.30 (52-99)
		N (%)
Traditional ToM Tasks (n = 54)		
	Age Expected	21 (38.2)
	Below Age Expected	33 (60.00)
		Mean (SD)
Triangles Task (ToM conditions) (n = 55)	ToM Accuracy	2.29 (1.63)

	Intentionality	7.78 (2.61)
	Length	10.31 (2.21)
BPVS (standard score) (n = 52)		94.98 (11.64)
DCCS (n = 49)		
	Age Corrected Scores	94.71 (15.03)
	Average N (%)	35 (63.60)
	Below Average N (%)	11 (20.00)
	Above Average N (%)	3 (5.5)
SDQ (Parent) (n=55)		
	Prosocial Behaviour	6.56 (2.34)
	Conduct Problems	4.95 (2.92)
	Peer Relationships	3.73 (2.35)
SDQ (Teacher) (n = 54)		
	Pro-Social Behaviour	4.81 (2.62)
	Conduct Problems	2.93 (2.66)
	Peer Relationships	3.56 (2.45)

Bivariate Kendall's correlations were conducted to test for correlations between study variables. Significant positive associations were found between the three Triangles scales. Age was also significantly positively correlated with the Triangles accuracy and intentionality scales, while gender was not (see Table 2).

Table 2. Bivariate Kendall's τ Coefficients

	Accuracy	Intentionality	Length	Age	Gender
Accuracy		0.68**	0.27**	0.26**	0.10
Intentionality			0.24*	0.29**	0.10
Length				0.13	0.09

** significance at $p < 0.01$

* significance at $p < 0.05$

Interrater Reliability

Initially, blind independent scoring for 20% of the Triangles data was conducted. There was good agreement per item on the accuracy scale (κ = 0.62), moderate agreement on the intentionality scale (κ = 0.51) and fair agreement for length (κ = 0.40). The independent assessors reviewed inconsistencies in these scores and came to an agreement. A further 18% of assessments were scored independently to assess reliability following preliminary discussions and agreement. The second round of independent scoring for assessments retrieved a Kappa score of 0.90 for accuracy, 0.91 for intentionality and 0.85 for length.

Hypotheses 1 and 2

In support of hypothesis 1, that performance on the Triangles task would be positively associated with performance on traditional ToM tasks, significant positive associations were found. Significant associations were found for the accuracy ($\tau = 0.37$) and intentionality ($\tau = 0.30$) scales at the $p < 0.01$ level. A significant association with Triangles length ($\tau = 0.25$) at the $p < 0.05$ level was also found. There were no significant differences found between the magnitude of the associations on the triangles subscales and performance on traditional ToM tests as presented in Table 3. Counter to the study hypothesis, accuracy was not found to be a better predictor of ToM difficulties in traditional tasks when compared to intentionality and length.

Table 3. Significance of difference between Triangles scales and traditional ToM task associations

Correlations (n = 54)	tau (τ)	converted r	Fisher's Z	z statistic	p value
Trad and Accuracy	0.37	0.55	0.62		
Trad and Intentionality	0.30	0.46	0.50		
Trad and Length	0.25	0.38	0.40		
Significance of differences between associations					
Trad and Accuracy / Trad and Intentionality				1.55	0.06
Trad and Accuracy / Trad and Length				1.34	0.09
Trad and Intentionality / Trad and Length				0.58	0.28

Trad = traditional Theory of Mind tasks

ToM = Theory of Mind

Hypothesis 3

Partial support was found for Hypothesis 3 that performance on the Triangles task would be less strongly associated with receptive language ability than traditional ToM tasks. In support of the hypothesis, a significant positive correlation was observed between BPVS standardised scores (measuring receptive language ability) and traditional ToM tasks ($\tau = 0.25$; $p < 0.05$), in contrast, there were no significant associations between BPVS scores and performance on the Triangles test; accuracy ($\tau = 0.14$; $p = 0.08$); intentionality ($\tau = 0.14$; $p = 0.08$); and length ($\tau = -0.10$; $p = 0.17$). Further, a significant difference was observed when the magnitude of the association between receptive language ability and traditional ToM task performance and receptive language ability and the length subscale of the Triangles were compared ($z = 3.67$; $p < 0.01$). However, observed differences in the magnitude of the associations between receptive language ability and traditional ToM task performance and receptive language ability and other indices of performance on the Triangles task (accuracy and intentionality sub-scales) were not statistically significant (see Table 4).

Table 4. Significance of difference between BPVS scores, traditional ToM tasks and Triangles scales associations

Correlations (n = 52)	**tau (τ)**	**converted r**	**Fisher's Z**	**z statistic**	**p value**
BPVS and Trad	0.25	0.39	0.41		
BPVS and Accuracy	0.14	0.22	0.23		
BPVS and Intentionality	0.14	0.22	0.23		
BPVS and Length	-0.10	-0.16	-0.16		
Significance of difference between associations					
BPVS and Trad / BPVS and Accuracy				1.34	0.09
BPVS and Trad / BPVS and Intentionality				1.23	0.11
BPVS and Trad / BPVS and Length				3.67	**0.00

BPVS = British Picture Vocabulary Scale

ToM = Theory of Mind

Trad = traditional Theory of Mind tasks

**significance at $p < 0.01$

Hypothesis 4

There were no significant associations between DCCS scores (measuring global EF skills) and traditional ToM task performance ($\tau = 0.159$; $p = 0.10$). Similarly, there were no significant associations between EF ability and performance on the Triangles test; accuracy ($\tau = 0.14$; $p = 0.09$); intentionality ($\tau = - 0.01$; $p = 0.46$); and length ($\tau = - 0.02$; $p = 0.45$). Further statistical analyses comparing the magnitude of the associations between EF ability and performance on traditional ToM tests and EF ability and performance on the triangles test found significant differences for the intentionality and length subscales of the Triangles test (see Table 5). Concurrent with the study hypothesis that performance on the Triangles task would be less strongly associated with EF skills than traditional ToM tasks, performance on the Triangles task (intentionality and length subscales) was less strongly associated with EF skills than traditional ToM tasks.

Table 5. Significance of difference between DCCS scores, traditional ToM tasks and Triangles scales associations

Correlations (n = 49)	tau (τ)	converted r	Fisher's Z	z statistic	p value
DCCS and Trad	0.16	0.25	0.25		
DCCS and Accuracy	0.14	0.22	0.23		
DCCS and Intentionality	-0.01	-0.02	-0.02		
DCCS and Length	-0.02	-0.02	-0.02		
Significance of difference between associations					
DCCS and Trad / DCCS and Accuracy				0.22	0.41
DCCS and Trad / DCCS and Intentionality				1.80	*0.03
DCCS and Trad / DCCS and Length				1.68	*0.04

DCCS = Dimensional Change Card Sort

Trad = traditional Theory of Mind tasks

*significance at $p < 0.05$

Hypothesis 5

There were no significant associations between parent or teacher reports of children's social-interpersonal skills and behavioural problems on the SDQ and Triangles accuracy and intentionality scale scores (Table 6).

Table 6. Bivariate Kendall's Correlations between Triangles and SDQ scales

	Triangles Accuracy	Triangles Intentionality	Triangles Length
Parent SDQ Conduct	0.00	0.03	0.30**
Parent SDQ Peer Relationships	-0.05	-0.10	0.17
Parent SDQ Prosocial	0.05	0.06	-0.18*
Teacher SDQ Conduct	-0.08	-0.05	0.06
Teacher SDQ Peer Relationships	-0.10	-0.08	0.03
Teacher SDQ Prosocial	0.03	0.03	0.12

SDQ = Strengths and Difficulties Questionnaire

**significance at $p < 0.01$

*significance at $p < 0.05$

Counter to the study hypothesis, a significant positive association was found between Triangles length and parent reports of children's conduct problems. Whilst a significant negative association was found between Triangles length and parent reports of prosocial behaviour. These findings suggest that performance on the Triangles task (length sub-scale) is associated with more parent reported conduct problems and worse parent reported prosocial behaviour.

Follow-up Analyses

The author was interested in whether the unexpected results for hypothesis five were due to differences at the ToM task or sample level. Bivariate Kendall's τ coefficients were run to assess the relationships between traditional ToM tasks and scores on the SDQ. There were no significant correlations found between parent or teacher SDQ scales and traditional ToM tasks (see Table 7).

Table 7. Bivariate Kendall's Correlations between Traditional ToM tasks and SDQ scales

	Traditional ToM Tasks
Parent SDQ Conduct	0.06
Parent SDQ Peer Relationships	-0.06
Parent SDQ Prosocial	0.01
Teacher SDQ Conduct	-0.07
Teacher SDQ Peer Relationships	-0.15
Teacher SDQ Prosocial	0.08

SDQ = Strengths and Difficulties Questionnaire

Post Hoc Power Analysis

A post-hoc power analysis was conducted to provide context for the interpretation of study findings, and also to inform future research. Power analyses were run based on Cohen's 'rule of thumb' for small (0.1), medium (0.3) and large (0.5) effect sizes in measures of associations ($n = 54$; $\alpha = 0.05$) (Cohen, 1992). The analyses retrieved a power score of 0.17 for small, 0.72 for medium and 0.98 for large effect sizes. This suggests that the current study was lacking in power to identify small to medium effects.

DISCUSSION

This study sought to validate the Triangles task, a ToM test, for the first-time in a sample of school-aged children with emerging social-interpersonal and behavioural problems. Triangles was considered to be a replicable and efficient test, with the scope to reflect the complexity of ToM abilities (Livingston et al. 2019). Based on a review of the existing literature, it was expected that the Triangles task would be a valid measure of ToM when compared to traditional ToM tasks (Castelli et al., 2002; Sharivar et al., 2020). It was also expected that Triangles would be less sensitive to receptive language and EF skills, a limitation highlighted within traditional false belief tasks (McAlister & Peterson, 2013), thus potentially making it a better test of ToM ability in children aged 4-7 years with emerging social-interpersonal and behavioural problems.

Interrater Reliability

Although not specifically included in the formal study hypotheses, the degree of interrater reliability in this younger child sample should be firstly considered when interpreting the main findings from the study. Interrater agreement using blind independent scoring was lower than that reported in studies of children (9-11 years) recruited from mainstream schools and adult samples using the triangles task (Castelli et al., 2002; Sharivar et al., 2020). This discrepancy may reflect the complexity of presentations and interpretation of data from young school-aged children with emerging social-interpersonal and behavioural problems. The language used by the sample in their descriptions of the animations, did not map on to the more sophisticated examples provided in the scoring criteria, for example ‘seduce’, ‘mock’ or ‘persuade’. The lack of child-friendly examples meant that scoring was left up to the assessors’ interpretation and clinical judgement. Discrepancies in scoring may also reflect the subjective nature of this process in the Triangles assessment, which has previously been criticised (White et al., 2011). On some occasions, the children’s descriptions and non-verbal behaviours (e.g. laughing) were suggestive of ToM understanding, but if their verbatim answers did not closely match the scoring criteria, they were scored lower. This made it particularly difficult for the assessors to remain objective and not draw assumptions from the data. The above issues raise questions as to whether important information is missed within the Triangles task scoring process when used with young children.

Convergent Validity

To test convergent validity, the Triangles task was compared to performance on a group of four widely used ToM tasks (Baron-Cohen, 1991; Baron-Cohen et al., 1985; Coull et al., 2006; Perner et al., 1989; Wellman & Lui, 2004). Performance on traditional ToM tasks was scored as age expected or below age expected against set criteria. In line with the study hypothesis 1, that performance on the Triangles task would be positively associated with performance on traditional ToM tasks, performance on Triangles was significantly positively associated with performance on traditional tasks. These findings suggest that the Triangles task may be an adequate test of first order ToM in young-school age children with early emerging social-interpersonal and behavioural problems.

Whilst the magnitude of association between Triangles accuracy scale and traditional ToM tasks was greater than that of intentionality and length, there was no significance of difference found between the scales in contrast from what was expected (hypothesis 2). Study findings might be seen as in contrast to some outcomes in the literature, where Triangles intentionality and length sub-scales were not significantly associated with expected group differences when comparing children with ASD and typically developing children (Abell at al., 2000; Salter et al., 2008). Abell et al. found children with ASD used significantly more inaccurate mental state language when compared to children with intellectual disabilities and typically developing children, explaining some of the differences in the sensitivity of the intentionality and length scales to ToM difficulties. One of the major differences between these studies and the current study is the age of the sample, and the assumed developmental stage of the children. It could be argued that the children in the current sample were less aware of the task requirements, because of their age and/or at risk status, and therefore used only spontaneous mental state language where accurate, rather than having an awareness of test expectations but failing to understand higher level processes. Awareness of task expectations and attempts to use mental state language, even though inaccurate, may be a reflection of a number of factors including the age of a sample, developmental stages, or the use of compensatory strategies.

Language

The author investigated whether performance on the Triangles task was less sensitive to language ability, specifically receptive language, in comparison to traditional ToM tasks. Emerging trends from the data were in line with the hypothesis 3 that performance on the Triangles task would be less strongly associated with receptive language ability than traditional ToM tasks. Traditional ToM tasks were significantly associated with receptive language scores while the Triangles scales were not. There was a significant difference between the magnitude of associations between receptive language and traditional ToM task performance and receptive language and performance of the Triangles 'length' subscale, although this was not the case for the accuracy and intentionality scales. There are a number of factors to consider when interpreting these findings.

The BPVS was chosen to measure language ability, as the demands on receptive language was a marked difference between traditional ToM tasks and the Triangles task. The small effect size for the BPVS and traditional tasks association (and relatedly therefore the lack of a significant difference between the magnitude of this association compared to the non-significant language associations between the accuracy and intentionality scales of the Triangles task), may be explained by the use of a specific language measure rather than a more general language measure (Milligan et al., 2007). The use of a more general measure may have yielded more conclusive results, in line with the literature on the ToM-language association (Ebert, 2020).

A significant effect of language ability has been demonstrated consistently for false belief tasks within both clinical and typically developing samples (Bailey & Im-Bolter, 2020;

Ebert, 2020). However, past research which considered the Triangles task has been less consistent; specifically, Salter at al. (2008) observed no significant associations between Triangles accuracy and intentionality sub-scales and language in a group of children with ASD similar to the current study in a sample with emotional/behavioural problems, but did find ToM-language associations in the Triangles task in the typically developing comparator sample. This highlights the possibility of the Triangles task being less strongly associated with language ability in certain clinical samples, perhaps children with social and behavioural problems. Observed differences found when using the Triangles task may reflect greater individuality in the development of ToM ability in some clinical samples, and possibly the need to find other means of understanding and interpreting the intentions of others outside of language. The Triangles tasks appears to be sensitive to these differences in comparison to traditional tasks, which would support the argument that the Triangles task is a purer measure of ToM ability. Further research is needed to understand the association between performance on the Triangles task and children's language ability relative to traditional ToM tasks.

Executive Functioning

Partial evidence to support hypothesis 4, that performance on the Triangles task would be less strongly associated with EF skills than traditional ToM tasks, was found. Statistical analyses comparing the magnitude of associations between EF and ToM tasks, identified significant differences between traditional tasks and two of the Triangles task scales (intentionality and length), this was not the case for the accuracy scale. Whilst the differences between the associations were significant, none of the associations were at a significant level. Given the small non-significant effect sizes found for the observed associations, it is inconclusive whether traditional ToM tasks are biased by EF skills. Further studies with larger sample sizes may provide a better test of whether the Triangles task is in fact a purer measure of first order ToM (McAlister & Peterson, 2013).

Social-Interpersonal Skills

It was expected that performance on the Triangles task would be associated with social-interpersonal skills and behavioural problems (Rosello et al., 2020). In contrast to the study hypothesis 5, Triangles length was the only subscale significantly associated with parent reports of conduct problems and prosocial behaviour. Furthermore, these significant associations were in an unexpected direction. The results indicate that children who used more clauses to describe the Triangles animations were rated by their parents as presenting with greater conduct problems and lower prosocial behaviour. This was not the case for teacher reports. Further analyses were conducted on the associations between traditional ToM tasks and the SDQ scales, to explore whether these unexpected outcomes were influenced by the task itself or the study sample. No significant correlations were observed in these analyses, indicating that the unexpected results were not due to the ToM task. These findings are in contrast to the literature using the Triangles task, where ToM difficulties in children (7-11 years) were a strong predictor of behaviour problems (Wells et al., 2020). In spite of the unexpected findings, the study results do not contradict the argument that ToM performance might still be an early transdiagnostic risk factor for later social-interpersonal problems, rather than a direct indicator of social-interpersonal and behavioural problems in young school-age children (Rosello et al., 2020). Further research is needed to understand the significance of the relationship between ToM as measured by the Triangles task and the development of social-interpersonal skills and behavioural problems over time.

Strengths and Limitations

Strengths of the current study included the wide variety of measures used in order to consider the validity of the Triangles Task in younger children, including language, EF, parent and teacher reported behavioural and social measures and established traditional ToM tasks. Furthermore, the subjective nature of the Triangles scoring is documented in the literature (White et al., 2011). In light of this, the author attempted to make the scoring and analysis process more robust by double transcribing the data to ensure accuracy, using two independent assessors for interrater reliability, and conducting a second round of independent scoring to improve accuracy prior to hypothesis testing. The triangles task has been validated for the first time in young children with emerging behavioural and social-interpersonal skills in this study, with promising results in terms of convergent validity. This study, while preliminary, adds to the literature trying to understand the complex relationships between ToM ability, language and EF skills. Nevertheless, whilst the study has many strengths, it also has limitations.

The study had a number of limitations. First, this is a preliminary cross-sectional validation study, where data was collected at one point in time. To fully understand the significance of emerging trends, a longitudinal study is required. Correlation analyses were used to assess the significance of relationships between study variables. Whilst this type of analysis provides an indication of associations, we cannot infer causation. Second, it would have been helpful to have a measure of expressive language ability due to the verbal nature of the Triangles task. The use of receptive language measures may be a limitation of the study, as previous research in early childhood has found that tests of general language abilities, covering multiple aspects of the language construct are better at capturing the complex relationship between language and ToM (Milligan et al., 2007). Additionally, due to Covid-19 there were restrictions on the number of participants recruited for this study. This affected the overall power of some of the statistical tests, thereby limiting the conclusions that can be drawn from the findings. Additional research looking at the demands of the Triangles task on skills such as language and EF with larger samples is required, in order to understand whether it is a better test of ToM ability in young children with emerging behavioural and social-interpersonal skills.

Implications for Clinical Practice

This was the first study to consider the validity of the Triangles task with a sample of children aged 4-7 years with emerging social-interpersonal and behavioural problems. Initial interrater discrepancies suggest the need to have more set guidance for consistent scoring in clinical use, particularly in younger samples where child-friendly examples of descriptions are needed. It is evident that descriptions can be interpreted in slightly different ways depending on how assessors have understood the scoring criteria. Study findings tentatively suggest that the Triangles task is a valid test of children's ToM skills. It is recommended that clinicians pay particular attention to the accuracy scale, while further investigation is needed in to the scoring and interpretation of the intentionality and length scales.

There are a number of factors, from a feasibility perspective, which clinicians may want to take in to account such as; the clarity of the child's speech over audio tapes, the ability to see the child's body language or lip movements to support interpretation; the length of time it takes to transcribe and score the data; and the subjectivity of scoring, meaning it may be an unreliable task to use when comparing scores over time, or across clinicians. Conversely, the Triangles task tentatively appears to be an adequate test of ToM and better represents the spectrum of first order ToM abilities when compared to traditional ToM tasks.

Future Research

There are a number of questions left unanswered in the current study which could be addressed in future research. With regards to the Triangles scales, further research into the use of inaccurate mental state language in some groups would inform the sensitivity of the intentionality scale in predicting ToM performance. Group characteristics such as age, developmental stage or clinical presentations could be explored. The use of general language tests may further our understanding of the non-significant trends observed in the current study. Longitudinal studies looking at the ToM-language relationship over time are also needed. A follow-up study to test whether early ToM difficulties assessed using Triangles were an early indicator of later social-interpersonal and behavioural problems is recommended. This would support the argument for testing this construct as an early risk factor within a neurodevelopmental assessment service.

CONCLUSION

In summary, this article reports a novel study validating the Triangles task in a sample of young school-aged children with emerging social-interpersonal and behavioural problems. This is the first time the Triangles task has been validated within a group of children predominantly aged 4 to 7 years with emerging social-interpersonal and behavioural problems. The Triangles task appears to be an adequate alternative to traditional ToM tasks in this age group in terms of correlating with traditional ToM tasks but having less covariance with language and executive functioning. However, ToM performance as measured by the Triangles Task, like the traditional ToM tasks, did not appear to be associated with social-interpersonal skills in this sample. Further research is needed in larger samples, with a longitudinal element, to build on the current study in terms of establishing whether the Triangles task is an efficient, reliable and valid alternative to traditional ToM tasks in research and clinical settings.

References

Abbeduto, L., Short-Meyerson, K., Benson, G. & Dolish, J. (2004). Relationship between theory of mind and language ability in children and adolescents with intellectual disability. *Journal of Intellectual Disability Research*, *48*(2), pp. 150-159. https://doi.org/10.1111/j.1365-2788.2004.00524.x

Abell, F., Happe, F. & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), pp. 1-16. https://doi.org/10.1016/S0885-2014(00)00014-9

American Psychological Association (APA). (2020). *Publication Manual of the American Psychological Association*. (7th Edition). Washington: American Psychological Association.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), pp. 502-508. https://doi.org/10.1111/opo.12131

Aslıer, M., Aslıer, N. G. Y., Kirkim, G. & Güneri, E. A. (2020). The influence of age and language on developmental trajectory of theory of mind in children with cochlear implants. *International Journal of Pediatric Otorhinolaryngology*, *135*, pp. 1-7. https://doi.org/10.1016/j.ijporl.2020.110127

Austin, G., Bondü, R. & Elsner, B. (2020). Executive Function, Theory of Mind, and Conduct Problem Symptoms in Middle Childhood. *Frontiers in psychology*, *11(539),* pp. 1-12. https://doi.org/10.3389/fpsyg.2020.00539

Bailey, K., & Im-Bolter, N. (2020). Theory of Mind and Language in Childhood Epilepsy. *Mind, Brain, and Education*, *14*(2), pp. 146-154. https://doi.org/10.1111/mbe.12230

Bal, E., Yerys, B. E., Sokoloff, J. L., Celano, M. J., Kenworthy, L., Giedd, J. N. & Wallace, G. L. (2013). Do social attribution skills improve with age in children with high functioning autism spectrum disorders?. *Research in Autism Spectrum Disorders*, *7*(1), pp. 9-16. https://doi.org/10.1016/j.rasd.2012.07.004

Baldimtsi, E., Nicolopoulou, A. & Tsimpli, I. M. (2021). Cognitive and Affective Aspects of Theory of Mind in Greek-Speaking Children with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *51*(4), pp. 1142-1156. https://doi.org/10.1007/s10803-020-04595-0

Baron-Cohen, S. (1991). Do people with autism understand what causes emotion?. *Child development*, *62*(2), pp. 385-395. https://doi.org/10.1111/j.1467-8624.1991.tb01539.x

Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen-year review. In S. Baron Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience* (pp. 3–20). Oxford University Press.

Baron-Cohen, S., Leslie, A. M. & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), pp. 37-46. https://doi.org/10.1016/0010-0277(85)90022-8

Beaudoin, C., Leblanc, É., Gagner, C. & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, *10*, pp. 1-23. https://doi.org/10.3389/fpsyg.2019.02905

Bohl, V. (2015). We read minds to shape relationships. *Philosophical Psychology*, *28*(5), pp. 674-694. https://doi.org/10.1080/09515089.2014.893607

British Psychological Society. (n.d.). *British Journal of Developmental Psychology.* (eJournal). Oxford: Wiley.

Calero, C. I., Salles, A., Semelman, M. & Sigman, M. (2013). Age and gender dependent development of Theory of Mind in 6-to 8-years old children. *Frontiers in human neuroscience*, *7*, pp. 281. https://doi.org/10.3389/fnhum.2013.00281

Castelli, F., Frith, C., Happé, F. & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*(8), pp. 1839-1849. https://doi.org/10.1093/brain/awf189

Castelli, F., Happé, F., Frith, U. & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage*, *12*(3), pp. 314-325. https://doi.org/10.1006/nimg.2000.0612

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), pp. 155-159. https://psycnet.apa.org/doi/10.1037/0033-2909.112.1.155

Coull, G. J., Leekam, S. R. & Bennett, M. (2006). Simplifying second-order belief attribution: What facilitates children’s performance on measures of conceptual understanding? *Social Development*, *15*(2), pp. 260-275. https://doi.org/10.1111/j.1467-9507.2006.00340.x

Dhadwal, A. K., Najdowski, A. C. & Tarbox, J. (2021). A Systematic Replication of Teaching Children With Autism and Other Developmental Disabilities Correct Responding to False-Belief Tasks. *Behavior Analysis in Practice*, 1-9. https://doi.org/10.1007/s40617-020-00531-x

Devine, R. T. & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child development, 84*(3), pp. 989-1003.

https://doi.org/10.1111/cdev.12017

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child development*, *85*(5), pp. 1777-1794. https://doi.org/10.1111/cdev.12237

Dumontheil, I., Apperly, I. A. & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental science*, *13*(2), pp. 331-338. https://doi.org/10.1111/j.1467-7687.2009.00888.x

Dunn, L. M., Dunn, L. M., Whetton, C. & Burley, J. (1997). British picture vocabulary scale II. Windsor. *UK: Nfer-Nelson*.

Ebert, S. (2020). Theory of mind, language, and reading: Developmental relations from early childhood to early adolescence. *Journal of experimental child psychology*, *191*, pp. 1-20. https://doi.org/10.1016/j.jecp.2019.104739

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. (4th Edition) Sage Publications, Inc.

Frith, C. & Frith, U. (2005). Theory of mind. *Current biology*, *15*(17), R644-R645. doi:10.1016/j.cub.2005.08.041

Gathercole, S. E., Willis, C. S., Baddeley, A. D. & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*(2), 103-127. https://doi.org/10.1080/09658219408258940

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Child Psychology & Psychiatry & Allied Disciplines, 38*(5), pp. 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Grant, C. M., Grayson, A. & Boucher, J. (2001). Using tests of false belief with children with autism: how valid and reliable are they?. *Autism*, *5*(2), pp. 135-145. https://doi.org/10.1177/1362361301005002004

Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental psychology*, *52*(8), pp. 1192-1205. https://psycnet.apa.org/doi/10.1037/dev0000140

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., Brande, M., Shamay-Tsooryg, S., Onur, A.O. & Kessler, J. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex*, *46*(6), pp. 769-780. https://doi.org/10.1016/j.cortex.2009.07.010

Kampis, D., Fogd, D. & Kovács, Á. M. (2017). Nonverbal components of Theory of Mind in typical and atypical development. *Infant Behavior and Development*, *48*, pp. 54-62. https://doi.org/10.1016/j.infbeh.2016.11.001

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org

Leno, V. C., Chandler, S., Whte, P., Yorke, I., Charman, T., Jones, C. R., Happe, F., Baird, G., Pickels, A. & Siminoff, E. (2020). Associations between theory of mind and conduct problems in autistic and non-autistic youth. *Autism Research*, *14*, pp. 276-288. https://doi.org/10.1002/aur.2346

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, *44*(2), pp. 523-531. https://psycnet.apa.org/doi/10.1037/0012-1649.44.2.523

Livingston, L. A., Shah, P., & Happé, F. (2019). *Further developing the Frith-Happé animations: A quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults.* (Unpublished doctoral dissertation). King's College London, London.

Low, J., & Simpson, S. (2012). Effects of labeling on preschoolers' explicit false belief performance: outcomes of cognitive flexibility or inhibitory control?. *Child Development*, *83*(3), pp. 1072-1084.

https://doi.org/10.1111/j.1467-8624.2012.01738.x

McAlister, A. R. & Peterson, C. C. (2013). Siblings, theory of mind, and executive functioning in children aged 3–6 years: New longitudinal evidence. *Child development*, *84*(4), pp. 1442-1458. https://doi.org/10.1111/cdev.12043

Milligan, K., Astington, J. W. & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, *78*(2), pp. 622-646.

https://doi.org/10.1111/j.1467-8624.2007.01018.x

Mireault, G. C., Crockenberg, S. C., Sparrow, J. E., Pettinato, C. A., Woodard, K. C., & Malzac, K. (2014). Social looking, social referencing and humor perception in 6-and 12 month-old infants. *Infant Behavior and Development*, *37*(4), pp. 536-545. https://doi.org/10.1016/j.infbeh.2014.06.004

Perner, J., Frith, U., Leslie, A. M. & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child development*, pp. 689-700. https://doi.org/10.2307/1130734

Peterson, C. C., Slaughter, V. & Wellman, H. M. (2018). Nimble negotiators: How theory of mind (ToM) interconnects with persuasion skills in children with and without ToM delay. *Developmental Psychology*, *54*(3), pp. 494 509. https://doi.org/10.1037/dev0000451

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), pp. 384-396. https://doi.org/10.1177%2F1745691619896607

Rosello, B., Berenguer, C., Baixauli, I., García, R. & Miranda, A. (2020). Theory of mind profiles in children with Autism Spectrum Disorder: adaptive/social skills and pragmatic competence. *Frontiers in Psychology*, *11*, pp. 1-14. https://doi.org/10.3389/fpsyg.2020.567401

Salter, G., Seigal, A., Claxton, M., Lawrence, K. & Skuse, D. (2008). Can autistic children read the mind of an animated triangle?. *Autism*, *12*(4), pp. 349-371. https://doi.org/10.1177/1362361308091654

Sebastian, C. L., McCrory, E. J. P., Cecil, C. A. M., Lockwood, P. L., De Brito, S. A., Fontaine, N. M. G. & Viding, E. (2012). Neural Responses to Affective and Cognitive Theory of Mind in Children with Conduct Problems and Varying Levels of Callous-Unemotional Traits. *Archives of General Psychiatry*, *69*(8), pp. 814-822. doi:10.1001/archgenpsychiatry.2011.2070

Senju, A. (2012). Spontaneous theory of mind and its absence in autism spectrum disorders. *The Neuroscientist*, *18*(2), pp. 108-113. https://doi.org/10.1177/1073858410397208

Shahrivar, Z., Tehrani-Doost, M., Banaraki, A. K. & Mohammadzadeh, A. (2020). Normative Data and Reliability of the Moving Shapes Paradigm. *Iranian Journal of Psychiatry*, *15*(2), pp. 143-149. PMCID: PMC7215251.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, *87*(2), pp. 245-251.

Walker, D. A. (2003). JMASM9: converting Kendall's tau for correlational or meta-analytic analyses. *Journal of Modern Applied Statistical Methods*, *2*(2), pp. 525-530. Doi: 10.22237/jmasm/1067646360

Warrier, V. & Baron-Cohen, S. (2018). Genetic contribution to 'theory of mind' in adolescence. *Scientific reports*, *8*(1), pp. 1-9.

https://doi.org/10.1038/s41598-018- 21737-8

Wellman, H. M. (2011). Developing a theory of mind. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 258–284). Wiley-

Blackwell.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, *72*(3), pp. 655-684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, *75*(2), pp. 523-541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wells, A. E., Hunnikin, L. M., Ash, D. P. & Van Goozen, S. H. (2020). Children with behavioural problems misinterpret the emotions and intentions of others. *Journal of abnormal child psychology*, *48*(2), pp. 213-221.

https://doi.org/10.1007/s10802-019-00594-7

White, S. J., Coniston, D., Rogers, R. & Frith, U. (2011). Developing the Frith-Happé animations: A quick and objective test of Theory of Mind for adults with autism. *Autism Research*, *4*(2), pp. 149-154. https://doi.org/10.1002/aur.174

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L. & Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development*, *78*(4), pp. 16-33. https://doi.org/10.1111/mono.12032

Appendices

A. Journal of Neuropsychology Review: Manuscript Submission Guidelines

Manuscripts submitted to Neuropsychology Review should conform to the style of the American Psychological Association Publication Manual (6th edition: 2010). Neuropsychology Review is an EQUATOR adopter. The EQUATOR network represents a collaboration of researchers and journal editors who aspire to improve accuracy and transparency in research by promoting better reporting standards. Because Neuropsychology Review publishes review articles, the EQUATOR elements most relevant are the PRISMA guidelines for preparation and reporting of systematic reviews and meta-analyses (http://www.equator-network.org/reporting-guidelines/prisma/).

While narrative reviews will still be considered for publication when appropriate, Neuropsychology Review encourages publication of systematic reviews of treatment, intervention and diagnostic validity studies as well as systematic reviews of research relating to scientific questions in all aspects of clinical neuropsychology and behavioral neuroscience. Systematic reviews are enhanced by inclusion of a carefully conducted meta-analysis whenever appropriate. Authors of systematic reviews and meta-analyses submitted to Neuropsychology Review should prepare their manuscripts according to the PRISMA guidelines and include a PRISMA checklist (http://prisma-statement.org/PRISMA Statement/Checklist.aspx) with manuscript submission. When completing the checklist, authors should consider whether their manuscript requires editing to address all of the reporting requirements.

Neuropsychology Review discourages use of numerical rating scales that assign a single number to rank the quality of studies included in the review. Instead authors should separately rate or classify individual study quality and risk of bias using established criteria such as those included in the critical appraisal checklists (e.g., randomized controlled trials or diagnostic validity studies (http://www.cebm.net/critical-appraisal/)). For treatment and intervention studies key risk-of-bias criteria include, but may not be limited to, adequacy of randomization, pre-treatment equality of groups, blinding of patients, therapist or person undertaking outcome evaluation, adequacy of follow-up and objectivity in outcome measurement. For diagnostic validity studies, risk-of-bias criteria include representativeness of sampling, full information on the test-to-be-evaluated (the index test) and diagnostic group status (the reference standard) and independent, blinded acquisition of reference and index test information. Other risk of bias criteria may be important in some contexts including commercial or other conflict of interest.

Prior to undertaking their systematic review, authors are encouraged to read the PRISMA Explanation and Elaboration paper (http://www.ncbi.nlm.nih.gov/pubmed/19621070). For authors not familiar with preparation of systematic reviews or the PRISMA guidelines, there are extensive information resources available on the PRISMA website (http://www.prisma-statement.org/).

Authors are encouraged to register their systematic review protocol early in the review process (e.g., PROSPERO), and use the PRISMA extension specifically written for reporting a systematic review protocol (i.e., , PRISMA-P (http://www.equator-network.org/reporting-guidelines/prisma-protocols/).

Authors of narrative reviews that are not based on systematic literature searching should justify in their cover letter and in the body of their manuscript why a systematic review was not feasible or appropriate. Likewise, authors of systematic reviews without meta-analysis should explain in their cover letter and in the body of their manuscript why meta-analysis was not considered appropriate (e.g., reviewed studies were not of sufficient quality).

Authors should avoid use of non-standard abbreviations.

Title page

Title Page

Please use this **template title page** for providing the following information.

The title page should include:

- The name(s) of the author(s)
- A concise and informative title
- The affiliation(s) of the author(s), i.e. institution, (department), city, (state), country
- A clear indication and an active e-mail address of the corresponding author
- If available, the 16-digit ORCID of the author(s)

If address information is provided with the affiliation(s) it will also be published.

For authors that are (temporarily) unaffiliated we will only capture their city and country of residence, not their e-mail address unless specifically requested.

Abstract

Please provide an abstract of 150 to 250 words. The abstract should not contain any undefined abbreviations or unspecified references.

Keywords

Please provide 4 to 6 keywords which can be used for indexing purposes.

Declarations

All manuscripts must contain the following sections under the heading 'Declarations'.

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

To be used for non-life science journals

Funding (information that explains whether and by whom the research was supported)

Conflicts of interest/Competing interests (include appropriate disclosures)

Availability of data and material (data transparency)

Code availability (software application or custom code)

Authors' contributions (optional: please review the submission guidelines from the journal whether statements are mandatory)

To be used for life science journals + articles with biological applications

Funding (information that explains whether and by whom the research was supported)

Conflicts of interest/Competing interests (include appropriate disclosures)

Ethics approval (include appropriate approvals or waivers)

Consent to participate (include appropriate statements)

Consent for publication (include appropriate statements)

Availability of data and material (data transparency)

Code availability (software application or custom code)

Authors' contributions (optional: please review the submission guidelines from the journal whether statements are mandatory)

Please see the relevant sections in the submission guidelines for further information as well as various examples of wording. Please revise/customize the sample statements according to your own needs.

References

Citation

Cite references in the text by name and year in parentheses. Some examples:

- Negotiation research spans many disciplines (Thompson 1990).
- This result was later contradicted by Becker and Seligman (1996).
- This effect has been widely studied (Abbott 1991; Barakat et al. 1995; Kelso and Smith 1998; Medvec et al. 1999).

Ideally, the names of six authors should be given before et al. (assuming there are six or more), but names will not be deleted if more than six have been provided.

Reference list

The list of references should only include works that are cited in the text and that have been published or accepted for publication. Personal communications and unpublished works should only be mentioned in the text. Do not use footnotes or endnotes as a substitute for a reference list.

Reference list entries should be alphabetized by the last names of the first author of each work.

Journal names and book titles should be *italicized*.

- Journal article Harris, M., Karper, E., Stacks, G., Hoffman, D., DeNiro, R., Cruz, P., et al. (2001). Writing labs and the Hollywood connection. *Journal of Film Writing*, 44(3), 213–245.
- Article by DOI Slifka, M. K., & Whitton, J. L. (2000) Clinical implications of dysregulated cytokine production. *Journal of Molecular Medicine*, https://doi.org/10.1007/s001090000086
- Book Calfee, R. C., & Valencia, R. R. (1991). *APA guide to preparing manuscripts for journal publication*. Washington, DC: American Psychological Association.
- Book chapter O'Neil, J. M., & Egan, J. (1992). Men's and women's gender role journeys: Metaphor for healing, transition, and transformation. In B. R. Wainrib (Ed.), *Gender issues across the life cycle* (pp. 107–123). New York: Springer.
- Online document Abou-Allaban, Y., Dell, M. L., Greenberg, W., Lomax, J., Peteet, J., Torres, M., & Cowell, V. (2006). Religious/spiritual commitments and psychiatric practice. Resource document. American Psychiatric Association. http://www.psych.org/edu/other_res/lib_archives/archives/200604.pdf. Accessed 25 June 2007.

For authors using EndNote, Springer provides an output style that supports the formatting of in-text citations and reference list.

Tables

- All tables are to be numbered using Arabic numerals.
- Tables should always be cited in text in consecutive numerical order.
- For each table, please supply a table caption (title) explaining the components of the table.
- Identify any previously published material by giving the original source in the form of a reference at the end of the table caption.
- Footnotes to tables should be indicated by superscript lower-case letters (or asterisks for significance values and other statistical data) and included beneath the table body.

Ethical Responsibilities of Authors

This journal is committed to upholding the integrity of the scientific record. As a member of the Committee on Publication Ethics (COPE) the journal will follow the COPE guidelines on how to deal with potential acts of misconduct.

Authors should refrain from misrepresenting research results which could damage the trust in the journal, the professionalism of scientific authorship, and ultimately the entire scientific endeavour. Maintaining integrity of the research and its presentation is helped by following the rules of good scientific practice, which include*:

- The manuscript should not be submitted to more than one journal for simultaneous consideration.
- The submitted work should be original and should not have been published elsewhere in any form or language (partially or in full), unless the new work concerns an expansion of previous work. (Please provide transparency on the re-use of material to avoid the concerns about text-recycling ('self-plagiarism').
- A single study should not be split up into several parts to increase the quantity of submissions and submitted to various journals or to one journal over time (i.e. 'salami-slicing/publishing').
- Concurrent or secondary publication is sometimes justifiable, provided certain conditions are met. Examples include: translations or a manuscript that is intended for a different group of readers.
- Results should be presented clearly, honestly, and without fabrication, falsification or inappropriate data manipulation (including image based manipulation). Authors should adhere to discipline-specific rules for acquiring, selecting and processing data.
- No data, text, or theories by others are presented as if they were the author's own ('plagiarism'). Proper acknowledgements to other works must be given (this includes material that is closely copied (near verbatim), summarized and/or paraphrased), quotation marks (to indicate words taken from another source) are used for verbatim copying of material, and permissions secured for material that is copyrighted.

Important note: the journal may use software to screen for plagiarism.

- Authors should make sure they have permissions for the use of software, questionnaires/(web) surveys and scales in their studies (if appropriate).
- Authors should avoid untrue statements about an entity (who can be an individual person or a company) or descriptions of their behavior or actions that could potentially be seen as personal attacks or allegations about that person.
- Research that may be misapplied to pose a threat to public health or national security should be clearly identified in the manuscript (e.g. dual use of research). Examples include creation of harmful consequences of biological agents or toxins, disruption of immunity of vaccines, unusual hazards in the use of chemicals, weaponization of research/technology (amongst others).
- Authors are strongly advised to ensure the author group, the Corresponding Author, and the order of authors are all correct at submission. Adding and/or deleting authors during the revision stages is generally not permitted, but in some cases may be warranted. Reasons for changes in authorship should be explained in detail. Please note that changes to authorship cannot be made after acceptance of a manuscript.

*All of the above are guidelines and authors need to make sure to respect third parties rights such as copyright and/or moral rights.

Upon request authors should be prepared to send relevant documentation or data in order to verify the validity of the results presented. This could be in the form of raw data, samples, records, etc. Sensitive information in the form of confidential or proprietary data is excluded.

If there is suspicion of misbehavior or alleged fraud the Journal and/or Publisher will carry out an investigation following COPE guidelines. If, after investigation, there are valid concerns, the author(s) concerned will be contacted under their given e-mail address and given an opportunity to address the issue. Depending on the situation, this may result in the Journal's and/or Publisher's implementation of the following measures, including, but not limited to:

- If the manuscript is still under consideration, it may be rejected and returned to the author.
- If the article has already been published online, depending on the nature and severity of the infraction:

  - an erratum/correction may be placed with the article

  - an expression of concern may be placed with the article

  - or in severe cases retraction of the article may occur.

The reason will be given in the published erratum/correction, expression of concern or retraction note. Please note that retraction means that the article is **maintained on the platform**, watermarked "retracted" and the explanation for the retraction is provided in a note linked to the watermarked article.

- The author's institution may be informed
- A notice of suspected transgression of ethical standards in the peer review system may be included as part of the author's and article's bibliographic record.

Fundamental errors

Authors have an obligation to correct mistakes once they discover a significant error or inaccuracy in their published article. The author(s) is/are requested to contact the journal and explain in what sense the error is impacting the article. A decision on how to correct the literature will depend on the nature of the error. This may be a correction or retraction. The retraction note should provide transparency which parts of the article are impacted by the error.

Suggesting / excluding reviewers

Authors are welcome to suggest suitable reviewers and/or request the exclusion of certain individuals when they submit their manuscripts. When suggesting reviewers, authors should make sure they are totally independent and not connected to the work in any way. It is strongly recommended to suggest a mix of reviewers from different countries and different institutions. When suggesting reviewers, the Corresponding Author must provide an institutional email address for each suggested reviewer, or, if this is not possible to include other means of verifying the identity such as a link to a personal homepage, a link to the publication record or a researcher or author ID in the submission letter. Please note that the Journal may not use the suggestions, but suggestions are appreciated and may help facilitate the peer review process.

Authorship principles

These guidelines describe authorship principles and good authorship practices to which prospective authors should adhere to.

Authorship clarified

The Journal and Publisher assume all authors agreed with the content and that all gave explicit consent to submit and that they obtained consent from the responsible authorities at the institute/organization where the work has been carried out, **before** the work is submitted.

The Publisher does not prescribe the kinds of contributions that warrant authorship. It is recommended that authors adhere to the guidelines for authorship that are applicable in their specific research field. In absence of specific guidelines it is recommended to adhere to the following guidelines*:

All authors whose names appear on the submission

1) made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work;

2) drafted the work or revised it critically for important intellectual content;

3) approved the version to be published; and

4) agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

* Based on/adapted from:

ICMJE, Defining the Role of Authors and Contributors,

Transparency in authors' contributions and responsibilities to promote integrity in scientific publication, McNutt at all, PNAS February 27, 2018

Disclosures and declarations

All authors are requested to include information regarding sources of funding, financial or non-financial interests, study-specific approval by the appropriate ethics committee for research involving humans and/or animals, informed consent if the research involved human participants, and a statement on welfare of animals if the research involved animals (as appropriate).

The decision whether such information should be included is not only dependent on the scope of the journal, but also the scope of the article. Work submitted for publication may have implications for public health or general welfare and in those cases it is the responsibility of all authors to include the appropriate disclosures and declarations.

Data transparency

All authors are requested to make sure that all data and materials as well as software application or custom code support their published claims and comply with field standards. Please note that journals may have individual policies on (sharing) research data in concordance with disciplinary norms and expectations. Please check the Instructions for Authors of the Journal that you are submitting to for specific instructions.

Role of the Corresponding Author

One author is assigned as Corresponding Author and acts on behalf of all co-authors and ensures that questions related to the accuracy or integrity of any part of the work are appropriately addressed.

The Corresponding Author is responsible for the following requirements:

- ensuring that all listed authors have approved the manuscript before submission, including the names and order of authors;
- managing all communication between the Journal and all co-authors, before and after publication;*
- providing transparency on re-use of material and mention any unpublished material (for example manuscripts in press) included in the manuscript in a cover letter to the Editor;
- making sure disclosures, declarations and transparency on data statements from all authors are included in the manuscript as appropriate (see above).

* The requirement of managing all communication between the journal and all co-authors during submission and proofing may be delegated to a Contact or Submitting Author. In this case please make sure the Corresponding Author is clearly indicated in the manuscript.

Author contributions

Please check the Instructions for Authors of the Journal that you are submitting to for specific instructions regarding contribution statements.

In absence of specific instructions and in research fields where it is possible to describe discrete efforts, the Publisher recommends authors to include contribution statements in the work that specifies the contribution of every author in order to promote transparency. These contributions should be listed at the separate title page.

Examples of such statement(s) are shown below:

• Free text:

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [full name], [full name] and [full name]. The first draft of the manuscript was written by [full name] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Example: CRediT taxonomy:

• Conceptualization: [full name], …; Methodology: [full name], …; Formal analysis and investigation: [full name], …; Writing - original draft preparation: [full name, …]; Writing - review and editing: [full name], …; Funding acquisition: [full name], …; Resources: [full name], …; Supervision: [full name],….

For **review articles** where discrete statements are less applicable a statement should be included who had the idea for the article, who performed the literature search and data analysis, and who drafted and/or critically revised the work.

For articles that are based primarily on the **student's dissertation or thesis**, it is recommended that the student is usually listed as principal author:

A Graduate Student's Guide to Determining Authorship Credit and Authorship Order, APA Science Student Council 2006

Affiliation

The primary affiliation for each author should be the institution where the majority of their work was done. If an author has subsequently moved, the current address may additionally be stated. Addresses will not be updated or changed after publication of the article.

Changes to authorship

Authors are strongly advised to ensure the correct author group, the Corresponding Author, and the order of authors at submission. Changes of authorship by adding or deleting authors, and/or changes in Corresponding Author, and/or changes in the sequence of authors are **not** accepted **after acceptance** of a manuscript.

- **Please note that author names will be published exactly as they appear on the accepted submission!**

Please make sure that the names of all authors are present and correctly spelled, and that addresses and affiliations are current.

Adding and/or deleting authors at revision stage are generally not permitted, but in some cases it may be warranted. Reasons for these changes in authorship should be explained. Approval of the change during revision is at the discretion of the Editor-in-Chief. Please note that journals may have individual policies on adding and/or deleting authors during revision stage.

Author identification

Authors are recommended to use their ORCID ID when submitting an article for consideration or acquire an ORCID ID via the submission process.

Deceased or incapacitated authors

For cases in which a co-author dies or is incapacitated during the writing, submission, or peer-review process, and the co-authors feel it is appropriate to include the author, co-authors should obtain approval from a (legal) representative which could be a direct relative.

Authorship issues or disputes

In the case of an authorship dispute during peer review or after acceptance and publication, the Journal will not be in a position to investigate or adjudicate. Authors will be asked to resolve the dispute themselves. If they are unable the Journal reserves the right to withdraw a manuscript from the editorial process or in case of a published paper raise the issue with the authors' institution(s) and abide by its guidelines.

Confidentiality

Authors should treat all communication with the Journal as confidential which includes correspondence with direct representatives from the Journal such as Editors-in-Chief and/or Handling Editors and reviewers' reports unless explicit consent has been received to share information.

Compliance with Ethical Standards

To ensure objectivity and transparency in research and to ensure that accepted principles of ethical and professional conduct have been followed, authors should include information regarding sources of funding, potential conflicts of interest (financial or non-financial), informed consent if the research involved human participants, and a statement on welfare of animals if the research involved animals.

Authors should include the following statements (if applicable) in a separate section entitled "Compliance with Ethical Standards" when submitting a paper:

- Disclosure of potential conflicts of interest
- Research involving Human Participants and/or Animals
- Informed consent

Please note that standards could vary slightly per journal dependent on their peer review policies (i.e. single or double blind peer review) as well as per journal subject discipline. Before submitting your article check the instructions following this section carefully.

The corresponding author should be prepared to collect documentation of compliance with ethical standards and send if requested during peer review or after publication.

The Editors reserve the right to reject manuscripts that do not comply with the above-mentioned guidelines. The author will be held responsible for false statements or failure to fulfill the above-mentioned guidelines.

Disclosure of potential conflicts of interest

Authors must disclose all relationships or interests that could have direct or potential influence or impart bias on the work. Although an author may not feel there is any conflict, disclosure of relationships and interests provides a more complete and transparent process, leading to an accurate and objective assessment of the work. Awareness of a real or perceived conflicts of interest is a perspective to which the readers are entitled. This is not meant to imply that a financial relationship with an organization that sponsored the research or compensation received for consultancy work is inappropriate. Examples of potential conflicts of interests **that are directly or indirectly related to the research** may include but are not limited to the following:

- Research grants from funding agencies (please give the research funder and the grant number)
- Honoraria for speaking at symposia

- Financial support for attending symposia
- Financial support for educational programs
- Employment or consultation
- Support from a project sponsor
- Position on advisory board or board of directors or other type of management relationships
- Multiple affiliations
- Financial relationships, for example equity ownership or investment interest
- Intellectual property rights (e.g. patents, copyrights and royalties from such rights)
- Holdings of spouse and/or children that may have financial interest in the work

In addition, interests that go beyond financial interests and compensation (non-financial interests) that may be important to readers should be disclosed. These may include but are not limited to personal relationships or competing interests directly or indirectly tied to this research, or professional interests or personal beliefs that may influence your research.

The corresponding author collects the conflict of interest disclosure forms from all authors. In author collaborations where formal agreements for representation allow it, it is sufficient for the corresponding author to sign the disclosure form on behalf of all authors. Examples of forms can be found

here:

The corresponding author will include a summary statement in the text of the manuscript in a separate section before the reference list, that reflects what is recorded in the potential conflict of interest disclosure form(s).

See below examples of disclosures:

Funding: This study was funded by X (grant number X).

Conflict of Interest: Author A has received research grants from Company A. Author B has received a speaker honorarium from Company X and owns stock in Company Y. Author C is a member of committee Z.

If no conflict exists, the authors should state:

Conflict of Interest: The authors declare that they have no conflict of interest.

ETHICS; INFORMED CONSENT……

B. Full Electronic Database Searches

1. MEDLINE

Search for: 3 and 4

Results: 878

Database: Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R) <1946
to March 26, 2021>
Search Strategy:
----
1 adolescent/ or child/ or child, preschool/ (3070569)
2 (child* or p?ediatric* or adolescen* or youngster* or teen* or young person* or young people* or school-age*).ti.
(1088103)
3 1 or 2 (3295903)
4 (cognit* adj3 (test* or assess* or measure* or screen*)).ti. (5432)
5 3 and 4 (878)

****

2. EMBASE

Search for: 6 and 7

Results: 907

Database: EMBASE <1947-Present>
Search Strategy:
----
1 pediatrics/ (88703)

2 child/ (2049274)
3 child/ or preschool child/ or school child/ (2338491)
4 adolescent/ (1707589)
5 (child* or p?ediatric* or adolescen* or youngster* or teen* or young person* or young people* or school-age*).ti.
(1415254)
6 1 or 2 or 3 or 4 or 5 (3483656)
7 (cognit* adj3 (test* or assess* or measure* or screen*)).ti. (7816)
8 6 and 7 (907)

****

3. CINAHL

S11	S9 AND S10	Expanders - Apply equivalent subjects modes - Boolean/Phrase Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL Plus with Full Text	(671)
S10	TI screen* OR TI test*	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase Interface – EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL Plus with Full Text	(159,793)
S9	S5 AND S8	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL Plus with Full Text	(21,147)

S8	S6 OR S7	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL Plus with Full Text	(99,290)
S7	TI cognit* OR TI intellect*	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL Plus with Full Text	(63,228)
S6	(MH "Cognition")	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL Plus with Full Text	(56,255)
S5	S1 OR S2 OR S3 OR S4	Expanders - Apply equivalent subjects Search modes - Boolean/Phrase Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL Plus with Full Text	(960,458)

S4 TI child* OR TI p?ediatric* OR TI adolescen* OR TI youngster* OR TI teen* OR TI young person OR TI young people* OR TI school-age*

Expanders - Apply equivalent subjects

Search modes - Boolean/Phrase

Interface - EBSCOhost Research Databases

Search Screen - Advanced Search

Database - CINAHL Plus with Full Text (390,934)

S3 (MH "Adolescence") Expanders - Apply equivalent subjects

Search modes - Boolean/Phrase

Interface - EBSCOhost Research Databases

Search Screen - Advanced Search

Database - CINAHL Plus with Full Text (543,089)

S2 (MH "Child") OR (MH "Child, Preschool")

Expanders - Apply equivalent subjects

Search modes - Boolean/Phrase

Interface - EBSCOhost Research Databases

Search Screen - Advanced Search

Database - CINAHL Plus with Full Text (534,173)

S1 **(**MH "Pediatrics") Expanders - Apply equivalent subjects

Search modes - Boolean/Phrase

Interface - EBSCOhost Research Databases

Search Screen - Advanced Search

Database - CINAHL Plus with Full Text (20,417)

4. APA PsycInfo

Search for: 5 and 8 and 11

Results: 962

Database: APA PsycInfo <1806 to March Week 3 2021>
Search Strategy:
----

1 pediatrics/ (26144)
2 child psychology/ or developmental psychology/ (11626)
3 adolescent psychology/ (4353)
4 (child* or p?ediatric* or adolescen* or youngster* or teen* or young person* or young people* or school-age*).ti.
(476107)
5 1 or 2 or 3 or 4 (489204)
6 cognition/ (34466)
7 cognit*.ti. (137874)
8 6 or 7 (156464)
9 screening/ or screening tests/ (16254)
10 (test* or assess* or measure* or screen*).ti. (283526)
11 9 or 10 (288396)
12 5 and 8 and 11 (962)
****

5. APA PsycTests

Search for: 5 and 9 and 12

Results: 23

Database: APA PsycTests <1910 to March 2021>
Search Strategy:
----
1 pediatrics/ (401)
2 child psychology/ or developmental psychology/ (351)
3 adolescent psychology/ (407)
4 (child* or p?ediatric* or adolescen* or youngster* or teen* or young person* or young people* or school-age*).ti.
(3200)
5 1 or 2 or 3 or 4 (3827)
6 cognition/ (174)
7 intellectual development/ or cognitive development/ (138)
8 cognit*.ti. (690)
9 6 or 7 or 8 (947)
10 screening tests/ or screening/ (1436)

11 (test* or assess* or measure* or screen*).ti. (27361)
12 10 or 11 (28031)
13 5 and 9 and 12 (23)

****

C. COSMIN Modified GRADE Approach for grading the quality of evidence

Quality of evidence	**Lower if**
High	Risk of bias 1 Serious 2 Very serious 3 Extremely serious Inconsistency 1 Serious 2 Very serious Imprecision 1 total n=50-100 2 total n<50 Indirectness 1 Serious 2 Very serious
Moderate	
Low	
Very Low	

n=sample size

D. Results of studies on measurement properties for each cognitive screen

Results of studies on measurement properties for BI Alert

BI Alert (ref)	**Country (language) in which the OM was evaluated**	**Structural validity**			**Internal consistency**			**Cross-cultural validity\ measurement invariance**			**Reliability**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
Rasquin et al. (2011)	Netherlands (English)	133	Adequate	Exploratory Factor Analysis: Two factor solution (?)	133	Very Good	Cronb. alpha; parents 0.68; teachers 0.82 (?)	71	Inadequate	MGCFA: No important differences found (+)	77	Doubtful	Pearson's CC reasonable (r = 0.46 - 0.82) (?)
Pooled or summary result (overall rating)		**133**		**2 factors (1?)**	**133**		**1?**	**71**		**1+**	**77**		**r 0.46 – 0.82 (1?)**

BI Alert	**Country (language) in which the OM was evaluated**	**Measurement error**			**Criterion validity**			**Hypotheses testing**			**Responsiveness**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)

Rasquin et al. (2011)	Netherlands (English)							133	Adequate	Result is in accordance with hypothesis for convergent validity (1+)			
Pooled or summary result (overall rating)								**133**		**1+**			

Results of studies on measurement properties for CNS Vital Signs

CNS Vital Signs (ref)	**Country (language) in which the OM was evaluated**	**Structural validity**			**Internal consistency**			**Cross-cultural validity\ measurement invariance**			**Reliability**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
Gualtieri & Johnson (2006)	USA (English)							84	Inadequate	No multiple group factor analysis OR DIF analysis	84	Doubtful	ICC/KAPPA not reported; Pearson's/Spearman's Correlation

										performed (?)			Coefficients reported (r = 0.314-0.874) (?)
Brooks & Sherman (2012)	Canada (English)							44	Adequate	MGCFA: No important differences found (+)			
Brooks et al. (2014)	Canada (English)							105	Very Good	MGCFA: No important differences found (+)			
Gualtieri & Hervey (2015)	USA (English)	3420	Very Good	Normal sample: CFI=.961; meets criteria of CFA (3 factors; memory, processing speed and attention) (+)									
Plourde &	Canada (English)							66	Inadequate	No MGCFA or DIF analysis			

Brooks (2017)										performed (?)			
Brooks et al. (2019)	Canada (English)	280	Adequate	EFA used: 3 factor solution (speed, memory and inhibition) (?)									
Pooled or summary result (overall rating)		**3,700**		**3 factors (1+)**				**299**		**No important differences found in studies where a robust design and analysis was used (2+)**	**84**		**A robust test was not used (ICC/Kappa); Pearsons/Spearmans tell order of scores remained the same (?)**

CNS Vital Signs	**Country (language) in which the OM**	**Measurement error**			**Criterion validity**			**Hypotheses testing**			**Responsiveness**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)

	was evaluated												
Gualtieri & Johnson (2006)	USA (English)				84	Very Good	Not all information for + provided (?)	84	Very Good	75% of results across all subtests were not in accordance with the hypotheses for convergent and discriminant validity (-)			
Brooks & Sherman (2012)	Canada (English)							44	Very Good	Above 75% of results were in accordance with hypotheses with significant differences between subgroups (+)			
Brooks et al. (2014)	Canada (English)							105	Very Good	75% of result is not in accordance with hypotheses testing for known-groups validity (-)			

Brooks et al. (2016)	Canada (English)				105	Very Good	60.8% overall correct classification rate (sensitivity = 0.929) (+)						
Plourde & Brooks (2017)	Canada (English)				66	Very Good	DFA analysis (sensitivity = 60.0%) (-)	66	Very Good	Above 75% of results were in accordance with hypotheses with significant differences between subgroups with large effect sizes (+)			
Pooled or summary result (overall rating)					**255**		**Inconsistent findings across different TBI subgroups (?)**	**299**		**Inconsistent results across studies (?)**			

Results of studies on measurement properties for CALS

CALS (ref)	Country (language) in which the OM was evaluated	Structural validity			Internal consistency			Cross-cultural validity\ measurement invariance			Reliability		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
Slomine et al. (2008)	USA (English)	100	Adequate	Not all information for '+' reported; EFA analysis conducted (?)	100	Very Good	At least low evidence for sufficient structural validity not met; Cronb. alpha ≥ 0.70; α = 0.96 (?)				9	Doubtful	ICC ≥ 0.70 (+)
Pooled or summary result (overall rating)		**100**		**2 factors (?)**	**100**		**α = 0.96 (?)**				**9**		**ICC = 0.99 (+)**

CALS	Country (language) in which the OM was evaluated	Measurement error			Criterion validity			Hypotheses testing			Responsiveness		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
Slomine et al. (2008)	USA (English)							100	Very Good	Results are in accordance with the	100	Very Good	Results are in accordance with the hypothesis -

										hypothesis - strong correlations between CALS and WeeFIM (+)			strong correlations between CALS and WeeFIM between admission and discharge (+)
Pooled or summary result (overall rating)								**100**		**(1+)**	**100**		**(1+)**

Results of studies on measurement properties for LANSE-C/A

LANSE-C/A	**Country (language) in which the OM was evaluated**	**Measurement error**			**Criterion validity**			**Hypotheses testing**			**Responsiveness**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
Lebby et al. (2015)	USA (English)				249	Very Good	Sensitivity of 94.62% using a cut off of 2 failed subtests (+)	249	Very Good	13 of the 14 subtests reached statistical significance at the			

										P < .001 level; when comparing sub-groups (+)			
Kahn, Asbell & Donders (2015)	USA (English)						Not included as no reference to gold standard for criterion				56	Adequ ate	Did not meet 75% criteria for convergent validity – possibly poor instrument comparison (-)
Pooled or summary result (overall rating)					**249**		**(1+)**	**249**		**>75% in accordance with hypothesis for known-groups (1+)**	**56**		**Insufficient evidence to meet criteria (-)**

Results of studies on measurement properties for S-FAVRES

S-FAVRES (ref)	**Country (language) in**	**Structural validity**	**Internal consistency**	**Cross-cultural validity\ measurement invariance**	**Reliability**

	which the OM was evaluated	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
MacDonald (2015)	Canada and USA (English)				241	Very Good	Did not meet criteria of Cronb. Aplha >0.70 for all subscales; criteria for at least low evidence of structural validity not met (?)	241	Inadequate	No important differences found between group factors (+)	10	Doubtful	ICC or weighted Kappa < 0.70 (-)
Pooled or summary result (overall rating)					**241**		**(?)**	**241**		**(1+)**	**10**		**ICC = 0.28-0.80 (-)**

S-FAVRES	**Country (language) in which the OM was evaluated**	**Measurement error**			**Criterion validity**			**Hypotheses testing**			**Responsiveness**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
MacDonald (2015)	Canada and USA (English)				241	Very Good	AUC = 0.85 (+)	241	Convergent (Doubtful); known-	Did not meet criteria for convergent			

									groups (Very Good)	validity (-); 75% of results were in accordance with hypothesis for discriminative validity (+)			
Pooled or summary result (overall rating)					**241**		**AUC = 0.85 (1+)**	**241**		**Evidence for construct validity using known groups (1 +)**			

Sufficient (+), Insufficient (–), or Indeterminate (?)

E. The British Journal of Developmental Psychology: Publication Guidelines

https://onlinelibrary.wiley.com/page/journal/2044835x/homepage/forauthors.html

Author Guidelines

Author Guidelines

The British Journal of Developmental Psychology publishes full-length (5000 words), empirical, conceptual, review and discussion papers, as well as brief reports (2000 words), in the areas described in the journal **overview**. Only papers which report methodologically sound and rigorous research and which make a substantive contribution to the theory and understanding in developmental psychology will be accepted for publication.
All papers published in The British Journal of Developmental Psychology are eligible for Panel A: Psychology, Psychiatry and Neuroscience in the Research Excellence Framework (REF).

1. Circulation
The circulation of the Journal is worldwide. Papers are invited and encouraged from authors throughout the world.

2. Length
Papers should be no more than 5000 words (excluding the abstract, reference list, tables and figures). In exceptional cases the Editor retains discretion to publish papers beyond this length where the clear and concise expression of the scientific content requires greater length (e.g., explanation of a new theory or a substantially new method). Authors must contact the Editor prior to submission in such a case.
Brief reports are limited to a maximum 2000 words (including the abstract, reference list, tables and figures) and have no more than 15 references. Brief reports will be treated as a priority during the review process and published in the next available issue once they are accepted.

3. Submission and reviewing
All manuscripts must be submitted via **Editorial Manager**. The Journal operates a policy of anonymous (double blind) peer review. We also operate a triage process in which submissions that are out of scope or otherwise inappropriate will be rejected by the editors without external peer review. Before submitting, please read **the terms and conditions of submission** and the **declaration of competing interests**. You may also like to use the **Submission Checklist** to help you prepare your paper. Papers will be evaluated by the Editor and referees in terms of their fit to the journal's aims and scope, theoretical interest, practical interest, timeliness, topicality and readability.

4. Manuscript requirements
• Contributions must be typed in double spacing with wide margins. All sheets must be numbered.
• Manuscripts should be preceded by a title page which includes a full list of authors and their affiliations, as well as the corresponding author's contact details. You may like to use **this** template. When entering the author names into Editorial Manager, the corresponding author will be asked to provide a CRediT contributor role to classify the role that each author played in creating the manuscript. Please see the **Project CRediT** website for a list of roles.
• The manuscript title must indicate the subject matter accurately but succinctly. Titles should be no longer than 120 characters (including spaces).
• All articles should be preceded by an abstract of between 100 and 150 words, giving a concise

statement of the intention, results or conclusions of the article and brief information regarding the ages and background and distinctive characteristics of any sample. The abstract should not include any sub-headings.
• All authors are required to provide a Statement of Contribution that identifies existing knowledge in the area and summarises the new knowledge added by the submitted paper. It should include two subheadings with 2 or 3 bullet points of no more than 100 characters under each, outlining (i) what is already known on this subject, and (ii) what the present study adds. The Statement of Contribution is submitted as a separate file.
• Conflict of Interest Statement: Where necessary authors must indicate any conflicts of interest. If such conflicts exist, a statement will be included at the end of each published manuscript. You will be asked to provide information to generate this statement during the submission process.
• The main document must be anonymous. Please do not mention the authors' names or affiliations and always refer to any previous work in the third person.
• SI units must be used for all measurements, rounded off if appropriate.
• Empirical reports must give details of the ages and other key characteristics (e.g., gender, ethnicity, socioeconomic status) of any sample.
• All manuscripts must explicitly indicate from where ethical approval was obtained for empirical research in the appropriate section of the Methods.
• Whenever possible, effect sizes should be reported.
• Tables should be typed in double spacing, each on a separate page with a self-explanatory title. Tables should be comprehensible without reference to the text. They should be placed at the end of the manuscript but they must be mentioned in the text.
• Figures can be included at the end of the document or attached as separate files, carefully labelled with symbols in a form consistent with text use. Unnecessary background patterns, lines and shading should be avoided. Captions should be listed on a separate sheet. The resolution of digital images must be at least 300 dpi. All figures must be mentioned in the text.
• For reference citations, please use APA style. Particular care should be taken to ensure that references are accurate and complete. Give all journal titles in full and provide doi numbers where possible for journal articles. For example:
Author, A., Author, B., & Author, C. (1995). *Title of book*. City, Country: Publisher.
Author, A. (2013). Title of journal article. *Name of journal*, 1, 1-16. doi: 10.1111/bjep.12031
• Authors must avoid the use of sexist or any other discriminatory language.
• Authors are responsible for acquiring written permission to publish lengthy quotations, illustrations, etc. for which they do not own copyright.
• For guidelines on editorial style, please consult the **APA Publication Manual** published by the American Psychological Association.
• Manuscripts describing clinical trials are encouraged to submit in accordance with the **CONSORT statement** on reporting randomised controlled trials.
• Manuscripts reporting systematic reviews and meta-analyses are encouraged to submit in accordance with the **PRISMA statement**.
• Manuscripts reporting interventions are encouraged to describe them in accordance with the **TIDieR checklist**.

If you need more information about submitting your manuscript for publication, please email Hannah Wakley, Managing Editor **bjdp@wiley.com** or phone +44 (0) 116 252 9504.

5. Supporting Information
Supporting Information can be a useful way for an author to include important but ancillary information with the online version of an article. Examples of Supporting Information include appendices, additional tables, data sets, figures, movie files, audio clips, and other related nonessential multimedia files. Supporting Information should be cited within the article text, and a

descriptive legend should be included. Please indicate clearly on submission which material is for online only publication. It is published as supplied by the author, and a proof is not made available prior to publication; for these reasons, authors should provide any Supporting Information in the desired final format.
For further information on recommended file types and requirements for submission, please visit the **Supporting Information page** on Author Services.

6. OnlineOpen
OnlineOpen is available to authors of primary research articles who wish to make their article available to non-subscribers on publication, or whose funding agency requires grantees to archive the final version of their article. With OnlineOpen, the author, the author's funding agency, or the author's institution pays a fee to ensure that the article is made available to non-subscribers upon publication via Wiley Online Library, as well as deposited in the funding agency's preferred archive. A full list of **terms and conditions** is available on Wiley Online Library.
Any authors wishing to send their paper OnlineOpen will be required to complete the **payment form**.
Prior to acceptance there is no requirement to inform an Editorial Office that you intend to publish your paper OnlineOpen if you do not wish to. All OnlineOpen articles are treated in the same way as any other article. They go through the journal's standard peer-review process and will be accepted or rejected based on their own merit.

7. Author Services
Author Services enables authors to track their article – once it has been accepted – through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The author will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. You can then access **Kudos** through Author Services, which will help you to increase the impact of your research. Visit **Author Services** for more details on online production tracking and for a wealth of resources including FAQs and tips on article preparation, submission and more.

8. Copyright and licences
If your paper is accepted, the author identified as the formal corresponding author for the paper will receive an email prompting them to login into Author Services, where via the Wiley Author Licensing Service (WALS) they will be able to complete the license agreement on behalf of all authors on the paper.
For authors signing the copyright transfer agreement
If the OnlineOpen option is not selected the corresponding author will be presented with the copyright transfer agreement (CTA) to sign. The terms and conditions of the CTA can be previewed in the samples associated with the **Copyright FAQs**.
For authors choosing OnlineOpen
If the OnlineOpen option is selected the corresponding author will have a choice of the following Creative Commons Licence Open Access Agreements (OAA):

- Creative Commons Attribution Non-Commercial Licence (CC-BY-NC)
- Creative Commons Attribution Non-Commercial -NoDerivs Licence (CC-BY-NC-ND)

To preview the terms and conditions of these open access agreements please visit the **Copyright FAQs** and you may also like to visit the **Wiley Open Access Copyright and Licence** page.
If you select the OnlineOpen option and your research is funded by The Wellcome Trust and members of the Research Councils UK (RCUK) or the Austrian Science Fund (FWF) you will be given the opportunity to publish your article under a CC-BY licence supporting you in complying with your Funder requirements. For more information on this policy and the Journal's compliant self-archiving policy please visit our **Funder Policy** page.

9. Colour illustrations
Colour illustrations can be accepted for publication online. These would be reproduced in greyscale in the print version. If authors would like these figures to be reproduced in colour in print at their expense they should request this by completing a **Colour Work Agreement** form upon acceptance of the paper.

10. Pre-submission English-language editing
Authors for whom English is a second language may choose to have their manuscript professionally edited before submission to improve the English. A list of independent suppliers of editing services can be found in **Author Services**. All services are paid for and arranged by the author, and use of one of these services does not guarantee acceptance or preference for publication.

11. The Later Stages
The corresponding author will receive an email alert containing a link to a web site. The proof can be downloaded as a PDF (portable document format) file from this site. Acrobat Reader will be required in order to read this file. This software can be downloaded (free of charge) from **Adobe's web site**. This will enable the file to be opened, read on screen and annotated direct in the PDF. Corrections can also be supplied by hard copy if preferred. Further instructions will be sent with the proof. Excessive changes made by the author in the proofs, excluding typesetting errors, will be charged separately.

12. Early View
British Journal of Developmental Psychology is covered by the Early View service on Wiley Online Library. Early View articles are complete full-text articles published online in advance of their publication in a printed issue. Articles are therefore available as soon as they are ready, rather than having to wait for the next scheduled print issue. Early View articles are complete and final. They have been fully reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after online publication. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so they cannot be cited in the traditional way. They are cited using their Digital Object Identifier (DOI) with no volume and issue or pagination information: e.g. Jones, A.B. (2010). Human rights Issues. *Journal of Human Rights*. Advance online publication. doi:10.1111/j.1467-9299.2010.00300.x Further information about the process of peer review and production can be found in this document. **What happens to my paper?** Appeals are handled according to the **procedure recommended by COPE**.

F. NDAU Referral Pack and Consent Forms

CONSENT AND REFERRAL FORMS

Neurodevelopment Assessment Unit
Cardiff University Centre for Human Developmental Science
School of Psychology
Cardiff, CF10 3AT

NDAU

Neurodevelopment Assessment Unit
Cardiff University Centre for Human Developmental Science
School of Psychology
Cardiff, CF10 3AT
02920 870354

Expression of Interest Form

Name of parent:	
Name of child:	
Child's date of birth:	
Child's gender	**Female** ☐ **Male** ☐
Family contact address:	
Family contact email:	
Family contact telephone:	

Name of Referrer	
Contact address of referrer:	
Contact email of referrer:	
Contact telephone of referrer:	

Once complete, please return this form to:

The Neurodevelopment Assessment Unit
Cardiff University Centre for Human Developmental Science
School of Psychology
Cardiff, CF10 3AT

Or email this form to: NDAU@cardiff.ac.uk

Neurodevelopment Assessment Unit
Cardiff University Centre for Human Developmental Science
School of Psychology
Cardiff, CF10 3AT
02920 870354

Overview of Pupil's Needs

This sheet is designed for the class teacher and/or SENCo to complete in order to provide the NDAU staff with a brief summary of the pupil's needs, the current interventions used and any current assessment information.

Overview of Presenting Needs	*Please give a brief summary of the pupil's presenting needs or areas for development. For example, social, emotional, motor, learning, etc.*
Overview of Interventions	*E.g. current approaches with pupil*
Overview of School-Based Assessment Data	*E.g. baseline assessments, foundation phase profile level, national curriculum levels, literacy levels, Language Link scores etc.*

Agency involvement: Please check school files and record external agency involvement.

	Involved? (Y/N)	***Brief Details***
Behaviour Support Services		
Learning Support Services		
Child and Family Service / CAMHS		
Children's / Social Services		
Speech & Language Therapy Service		
Occupational Therapy		
Other Health Services		
Other		

PLEASE ALSO ENSURE THAT YOU OR A RELEVANT MEMBER OF STAFF COMPLETES THE STRENGTHS AND DIFFICULTIES QUESTIONNAIRE (SDQ- Teacher version).

Please indicate here whether you are happy with the school's SDQ ratings of the child to be included in the report: YES / NO (please circle)

Neurodevelopment Assessment Unit
Cardiff University Centre for Human Developmental Science
School of Psychology
Cardiff, CF10 3AT
02920 870354

Referral Pathway Flow Chart

School SENCo identifies appropriate pupil to refer to the Neurodevelopment Assessment Unit (NDAU)

↓

School SENCo provides parents or guardians with the Parent Information Form

↓

School SENCo and Parent complete the Expression of Interest Form, Parental Consent Form and Overview of Pupil Needs Form. The child's teacher also completes the Strengths and Difficulties Questionnaire. These should be sent to the NDAU.

↓

NDAU will send parents an appointment date and time

↓

Pupil and parent attend appointment at NDAU in Cardiff University

(Travel expenses will be paid)

↓

NDAU creates a summary report, which is sent to school

↓

Parents and SENCo meet to discuss summary report and plan future interventions

Neurodevelopment Assessment Unit
Cardiff University Centre for Human Developmental Science
School of Psychology
Cardiff, CF10 3AT
02920 870354

STUDY CONSENT FORM

(for parents of children aged 4-7 years)

This is to be completed by parents/care-givers on behalf of their child and themselves. Please initial box

1. I confirm that I have read and understood the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐
2. I understand that my participation and that of my child is voluntary and that I am free to withdraw at any time without giving any reason, without my legal rights being affected. ☐
3. I am happy for the research team to make contact with me if there are any future research studies that might be of interest to me. ☐
4. I agree for my child to perform the developmental assessments as part of the study named above, including measuring my child's heart-rate. ☐
5. I agree to complete the parental interview and questionnaires as part of the study named above. ☐
6. I understand that relevant sections of my child's data collected during the study (including my ratings about my child on the Strengths and Difficulties Questionnaire) may be looked at by individuals from the NDAU study team, from regulatory authorities or by my child's referring agent, where it is relevant to their taking part in this research. I give permission for these individuals to have access to my child's data. ☐
7. I understand that an assessment report of my child's strengths and difficulties will be sent to the referring agent to guide their intervention with my child within the school environment. I understand that I do not receive a copy of this report. ☐
8. I understand that a video recording will be made of my child's assessments for research, safety and training purposes. I understand that brief clips from the video may be used to illustrate important aspects of child development, and to train new researchers, and so such clips may be shown to students or at professional meetings. I give consent for such clips to be taken from this video record, with the understanding that my name or my child's name will never be associated with the video clip. I understand that the video will remain in the possession of Prof. Van Goozen and the NDAU research team, and will never be given to other unauthorised individuals. ☐
9. I agree that assessment can be linked to routinely collected, anonymised datasets (such as those held in the Secure Anonymised Information Linkage [SAIL] databank), in order to answer future questions related to mental health. I understand that the data within any such dataset will be fully anonymised and my child would not be identifiable in any way. ☐

______________________	___________	________________________
Name of parent	Date	Signature

______________________ __________ __________________________

Name of person taking consent Date Signature

The information provided will be held in compliance with GDPR regulations. Cardiff University is the data controller and Matt Cooper is the data protection officer (inforequest@cardiff.ac.uk). The lawful basis for processing this information is public interest. This information is being collected by Professor Stephanie van Goozen.

The information on the consent form will be held securely and separately from the research information. Only the researcher will have access to this form and it will be destroyed after 7 years.

The research information you provide will be used for the purposes of research only and will be stored securely. Only members of the NDAU research team will have access to this information. After 7 years the data will be anonymised (any identifying elements removed) and this anonymous information may be kept indefinitely or published.

G. Ethical Approval of Empirical Project

From: psychethics <psychethics@cardiff.ac.uk>

Subject: Ethics Feedback - EC.16.10.11.4592GRA5

Date: 5 July 2018 at 10:34:22 BST

To: Stephanie Van Goozen <VangoozenS@cardiff.ac.uk>

Dear Steph,

The Ethics Committee has considered the amendment to your Staff project proposal: A Feasibility Study of a Neurodevelopmental Disorders Assessment Unit (EC.16.10.11.4592GRA5).

The amendment has been approved on the condition that a comment is added to the information, stating that if a child shows distress the monitor can be removed immediately.

Please note that if any changes are made to the above project then you must notify the Ethics Committee.

Best wishes,

Mark Jones

School of Psychology Research Ethics Committee

Cardiff University	Prifysgol Caerdydd
Tower Building	Adeilad y Tŵr
70 Park Place	70 Plas y Parc
Cardiff	Caerdydd
CF10 3AT	CF10 3AT
Tel: +44(0)29 208 70360	Ffôn: +44(0)29 208 70360
Email: psychethics@cardiff.ac.uk	E-bost: psychethics@caerdydd.ac.uk
http://psych.cf.ac.uk/aboutus/ethics.html	

H. TRIANGLES SCORING CRITERIA

1. Appropriateness/accuracy (i.e., whether participants correctly identify what is going on in the animation). You can score appropriateness/accuracy by following instructions by Abell et al. (2000) (see below). You will be scoring accuracy for both the goal directed videos and the theory of mind videos. As each clip scores a total of 2, the max score for goal directed would be 4 (2 x 2) and the max for ToM would be 6 (3 x 2, as 3 clips only, 1 as practice).

2. Intentionality or mental state language and Length (see Castelli et al 2000, below).

Abell et al. (2000) Scoring Criteria for Appropriateness

A.1. Accuracy of description: general rules

Each description is scored 2, 1, or 0 according to how accurately it reflects the sequence.
2 spot-on description of the story or the actions represented; can be concise just capturing gist, or can be discursive
1 partial description of the sequence; description is related to the sequence, but imprecise

or incomplete
0 bizarre descriptions, plainly wrong descriptions, and responses that focus solely on a minor unimportant aspect of the sequence

G-D movement sequences
Fighting: Character roles: two deer. No enclosure.
2 action implying physical fight, e.g. bashing each other
1 action that conveys the idea of a conflict, but is either too specific or too vague, e.g. biting; pushing
0 action that does not relate to conflict, e.g. following each other

Chasing: Character roles: two cats. Enclosure
2 description that conveys the idea of a chase
1 description that is related to but somewhat remote from chasing
0 action that does not relate to chasing, e.g. going up and down

ToM movement sequences
Surprising: Character roles: grandma and grandson. Enclosure.
2 any mention of boy tricking, surprising his grandma; hiding, hide and seek
1 description which gives part of the story but misses the critical point (see above)
0 description which gives only minor part of action e.g. knocking on the door, or does not relate to any of the events in the sequence.

Coaxing: Character roles: mother and child. Enclosure.
2 descriptions that conveys child's reluctance to go out and mother's attempts to get child out, e.g. persuading
1 partially correct description focussing on one aspect of the story or one character only, e.g. child does not want to go out; or, mother is pushing child to go out
0 actions that do not relate to the events or relate to a minor aspect of the sequence only, e.g. dancing together, or unrelated description.

Mocking: Character roles: teacher and boy. No enclosure.

2 description that conveys that boy is copying teacher without the teacher noticing, including pretending, hiding, being naughty

1 partially correct, e.g. following, copying

0 focus on a single unimportant event, e.g. boy ran away, or unrelated description

Seducing: Character roles: girl prisoner and guard. Enclosure.

2 description that conveys the girl prisoner luring, persuading or tricking the guard

1 partial story with minimal action for each character, e.g. girl trying to escape; guard blocking

0 description which focusses on unimportant event or is extremely minimal, e.g. she got out, or unrelated description

A.2. Type of description: general rules

Interaction: Specific reference to purposeful movement, without reference to mental states, e.g. following; fighting; copying; having a race. More than one action may be described, e.g. leading and following. May involve use of direct speech without mental state verb; may include qualification of verb by `trying to', e.g. boy trying to ask her something, but teacher kept walking away.

NOT: purposeless action. NOT: implied mental state attribution.

Mentalising attribution: Use of mental state verbs to describe reciprocal interactions, e.g. wanting; hiding; tricking; pretending; being naughty. NOT: complex interaction, e.g. chasing each other round the house; x pushing y out of the way. NOT: solely direct speech. NOT: solely `trying to.'

Castelli et al. (2000)

APPENDIX 2

Scoring Criteria and Examples for Verbal Descriptions of Animations

Score (0–5) for Intentionality:

0 = action, non-deliberate (e.g., “Bouncing,” “Moving around,” “Rotating”)

1 = deliberate action with no other (e.g., “Ice-skating”)

2 = deliberate action with another (e.g., “Blue and red are fighting,” “Parent is followed by child”)

3 = deliberate action in response to other's action (e.g., "Big is chasing little," "Red is allowing the Blue to get close to him," "Big is guarding little who was trying to escape")

4 = deliberate action in response to other's mental state (e.g., "The little one is mocking the big one," "Two people are arguing," "A parent is encouraging a child to go outside")

5 = deliberate action with goal of affecting other's mental state (e.g., "The blue triangle wanted to surprise the red one," "Child pretending not to be doing anything")

Score (0–4) for Length:

0 = no response

1 = one clause

2 = two clauses

3 = four clauses

4 = more than four clauses

I. SDQ Scoring

Scoring the Strengths & Difficulties Questionnaire for age 4-17 or 18+

The 25 items in the SDQ comprise 5 scales of 5 items each. It is usually easiest to score all 5 scales first before working out the total difficulties score. 'Somewhat True' is always scored as 1, but the scoring of 'Not True' and 'Certainly True' varies with the item, as shown below scale by scale. For each of the 5 scales the score can range from 0 to 10 if all items were completed. These scores can be scaled up pro-rata if at least 3 items were completed, e.g. a score of 4 based on 3 completed items can be scaled up to a score of 7 (6.67 rounded up) for 5 items.

Note that the items listed below are for 4-17-year-olds, but the scoring instructions are identical for the similarly-worded '18+' SDQ

Table 1: Scoring symptom scores on the SDQ for 4-17 year olds

	Not True	Somewhat True	Certainly True
Emotional problems scale			
ITEM 3: Often complains of headaches… *(I get a lot of headaches…)*	**0**	**1**	**2**
ITEM 8: Many worries… (*I worry a lot)*	**0**	**1**	**2**
ITEM 13: Often unhappy, downhearted… (*I am often unhappy….)*	**0**	**1**	**2**

ITEM 16: Nervous or clingy in new situations… (*I am nervous in new situations…)*	**0**	**1**	**2**
ITEM 24: Many fears, easily scared *(I have many fears…)*	**0**	**1**	**2**
Conduct problems Scale			
ITEM 5: Often has temper tantrums or hot tempers (*I get very angry)*	**0**	**1**	**2**
ITEM 7: Generally obedient… *(I usually do as I am told)*	**2**	**1**	**0**
ITEM 12: Often fights with other children… *(I fight a lot)*	**0**	**1**	**2**
ITEM 18: Often lies or cheats *(I am often accused of lying or cheating)*	**0**	**1**	**2**
ITEM 22: Steals from home, school or elsewhere *(I take things that are not mine)*	**0**	**1**	**2**
Hyperactivity scale			
ITEM 2: Restless, overactive… *(I am restless…)*	**0**	**1**	**2**
ITEM 10: Constantly fidgeting or squirming *(I am constantly fidgeting….)*	**0**	**1**	**2**
ITEM 15: Easily distracted, concentration wanders *(I am easily distracted)*	**0**	**1**	**2**
ITEM 21: Thinks things out before acting *(I think before I do things)*	**2**	**1**	**0**
ITEM 25: Sees tasks through to the end… *(I finish the work I am doing)*	**2**	**1**	**0**
Peer problems scale			
ITEM 6: Rather solitary, tends to play alone *(I am usually on my own)*	**0**	**1**	**2**
ITEM 11: Has at least one good friend *(I have one goof friend or more)*	**2**	**1**	**0**
ITEM 14: Generally liked by other children *(Other people my age generally like me)*	**2**	**1**	**0**
ITEM 19: Picked on or bullied by other children… *(Other children or young people pick on me)*	**0**	**1**	**2**
ITEM 23: Gets on better with adults than with other children *(I get on better with adults than with people my age)*	**0**	**1**	**2**
Prosocial scale			
ITEM 1: Considerate of other people's feelings *(I try to be nice to other people)*	**0**	**1**	**2**
ITEM 4: Shares readily with other children… *(I usually share with others)*	**0**	**1**	**2**
ITEM 9: Helpful if someone is hurt… *(I am helpful is someone is hurt…)*	**0**	**1**	**2**
ITEM 17: Kind to younger children *(I am kind to younger children)*	**0**	**1**	**2**
ITEM 20: Often volunteers to help others… *(I often volunteer to help others)*	**0**	**1**	**2**

Total difficulties score: This is generated by summing scores from all the scales except the prosocial scale. The resultant score ranges from 0 to 40, and is counted as missing of one of the 4 component scores is missing.

‘Externalising’ and ‘internalising’ scores: The externalising score ranges from 0 to 20 and is the sum of the conduct and hyperactivity scales. The internalising score ranges from 0 to 20 and is the sum of the emotional and peer problems scales. Using these two amalgamated scales may be preferable to using the four separate scales in community samples, whereas using the four separate scales may add more value in high-risk samples *(see Goodman & Goodman. 2009 Strengths and difficulties questionnaire as a dimensional measure of child mental health. J Am Acad Child Adolesc Psychiatry 48(4), 400-403).*

Generating impact scores

When using a version of the SDQ that includes an 'impact supplement', the items on overall distress and impairment can be summed to generate an impact score that ranges from 0 to 10 for parent- and self-report, and from 0 to 6 for teacher-report.

Table 2: Scoring the SDQ impact supplement

	Not at all	Only a little	A medium amount	A great deal
Parent report:				
Difficulties upset or distress child	**0**	**0**	**1**	**2**
Interfere with HOME LIFE	**0**	**0**	**1**	**2**
Interfere with FRIENDSHIPS	**0**	**0**	**1**	**2**
Interfere with CLASSROOM LEARNING	**0**	**0**	**1**	**2**
Interfere with LEISURE ACTIVITIES	**0**	**0**	**1**	**2**
Teacher report:				
Difficulties upset or distress child	**0**	**0**	**1**	**2**
Interfere with PEER RELATIONS	**0**	**0**	**1**	**2**
Interfere with CLASSROOM LEARNING	**0**	**0**	**1**	**2**
Self-report report:				
Difficulties upset or distress child	**0**	**0**	**1**	**2**
Interfere with HOME LIFE	**0**	**0**	**1**	**2**
Interfere with FRIENDSHIPS	**0**	**0**	**1**	**2**
Interfere with CLASSROOM LEARNING	**0**	**0**	**1**	**2**
Interfere with LEISURE ACTIVITIES	**0**	**0**	**1**	**2**

Responses to the questions on chronicity and burden to others are not included in the impact score. When respondents have answered 'no' to the first question on the impact supplement (i.e. when they do not perceive themselves as having any emotional or behavioural difficulties), they are not asked to complete the questions on resultant distress or impairment; the impact score is automatically scored zero in these circumstances.

Cut-points for SDQ scores for age 4-17: original 3-band solution & newer 4-band solution

Although SDQ scores can be used as continuous variables, it is sometimes convenient to categorise scores. The initial bandings presented for the SDQ scores were 'normal', 'borderline' and 'abnormal'. These bandings were defined based on a population-based UK survey, attempting to choose cut points such that 80% of children scored 'normal', 10% 'borderline' and 10% 'abnormal'.

More recently a four-fold classification has been created based on an even larger UK community sample. This four-fold classification differs from the original in that it (1) divided the top 'abnormal' category into two groups, each containing around 5% of the population, (2) renamed the four categories (80% 'close to average', 10% 'slightly raised, 5% 'high' and 5% 'very high' for all scales except prosocial, which is 80% 'close to average', 10% 'slightly lowered', 5% 'low' and 5% 'very low'), and (3) changed the cut-points for some scales, to better reflect the proportion of children in each category in the larger dataset.

Note that these cut points have not been validated for use with the 18+ SDQ, so we suggest that it is safest to use continuous scores rather than categories for this measure

Table 3: Categorising SDQ scores for 4-17 year olds (not validated for 18+)

	Original 3-band categorisation			**Newer 4-band categorisation**			
	Normal	Borderline	Abnormal	Close to average	Slightly raised (/slightly lowered)	High (/Low)	Very high (very low)
<u>Parent completed SDQ</u>							
Total difficulties score	0-13	14-16	17-40	0-13	14-16	17-19	20-40
Emotional problems score	0-3	4	5-10	0-3	4	5-6	7-10
Conduct problems score	2	3	4-10	0-2	3	4-5	6-10
Hyperactivity score	5	6	7-10	0-5	6-7	8	9-10
Peer problems score	6	3	4-10	0-2	3	4	5-10
Prosocial score	6-10	5	0-4	8-10	7	6	0-5
Impact score	0	1	2-10	0	1	2	3-10
<u>Teacher completed SDQ</u>							
Total difficulties score	0-11	12-15	16-40	0-11	12-15	16-18	19-40
Emotional problems score	0-4	5	6-10	0-3	4	5	6-10
Conduct problems score	0-2	3	4-10	0-2	3	4	5-10
Hyperactivity score	0-5	6	7-10	0-5	6-7	8	9-10
Peer problems score	0-3	4	5-10	0-2	3-4	5	6-10
Prosocial score	6-10	5	0-4	6-10	5	4	0-3
Impact score	0	1	2-6	0	1	2	3-6
<u>Self-completed SDQ</u>							
Total difficulties score	0-15	16-19	20-40	0-14	15-17	18-19	20-40
Emotional problems score	0-5	6	7-10	0-4	5	6	7-10
Conduct problems score	0-3	4	5-10	0-3	4	5	6-10
Hyperactivity score	0-5	6	7-10	0-5	6	7	8-10
Peer problems score	0-3	4-5	6-10	0-2	3	4	5-10
Prosocial score	6-10	5	0-4	7-10	6	5	0-4
Impact score	0	1	2-10	0	1	2	3-10

Note that both these systems only provide a rough-and-ready way of screening for disorders; combining information from SDQ symptom and impact scores from multiple informants is better, but still far from perfect.