The Structure and Psychometric Properties of the BERRI, an Outcome Measure for Looked After Children in Residential Care

Thesis submitted in part fulfilment of the degree of

Doctorate in Clinical Psychology (DClinPsy)

University of Leicester

By

Abigail Harris

Supervised by: Dr Gareth Morgan

April 2019

Declaration

I confirm that the literature review and empirical research reported in this thesis are original pieces of work. They are submitted in part fulfilment of the Doctorate in Clinical Psychology (DClinPsy) and have not been submitted for any other academic award. This thesis was checked for completion prior to submission.

The Structure and Psychometric Properties of the BERRI, an Outcome Measure for Looked After Children in Residential Care

By Abigail Harris

Abstract

Mental health outcome measures are being increasingly used to monitor the efficacy of interventions put in place to support children. For Looked After Children (LAC), mental health outcome measures are of further importance for placement planning. It is important that the measures used for these purposes are psychometrically robust.

Literature Review

The psychometric properties of 25 informant rated mental health outcome measures for young people were systematically reviewed. A novel quality appraisal tool was developed to evaluate evidence pertaining to internal consistency, test-retest and inter-rater reliability (IRR), construct validity and responsiveness. No measure provided evidence of acceptable rigour in all assessed domains. Generally, this was a result of an absence of evidence. A need for further research pertaining to the psychometric properties of these measures, particularly with respect to their responsiveness to change, test-retest and inter-rater reliability was identified.

Research Report

This study aimed to explore the psychometric properties of the BERRI in its current form for use with LAC in residential care and to explore whether these properties might be enhanced through the extraction of factors. Evidence of good internal consistency and construct validity was found for all original scales. Inter-rater reliability was 'poor-moderate' for three of the five scales and 'moderate-good' for the remaining two and the BERRI total score. An exploration of the structure of the BERRI using principal components analysis revealed a five component structure. The psychometric properties of the BERRI were not improved through the empirical extraction of components. Suggestions were made with regards to the item content of the BEERI. Consideration was given to the clinical implications arising from the exploration of the measure's IRR and subsequently how IRR might be improved. Overall, the BERRI was felt to show promise as a targeted outcome measure for use with LAC in residential care.

Acknowledgements

This thesis is dedicated to my incredible husband, Tim. Thank you for your endless encouragement, support and patience. Your unwavering belief in me has made this challenge surmountable.

To my research supervisor, Dr Gareth Morgan, thank you for your sound and calm advice from the conception of this project to its end. Thank you for always being available at the end of the phone or email for many a statistical emergency and for going above and beyond to help me make this research the best it could be. Thank you also to Dr Alice Welham for always being willing to share your statistical wisdom!

To Dr Miriam Silver, thank you for the trust you placed in me to work with the measure you have put so much time and effort into creating. Your passion for improving the lives of Looked After Children has inspired me since the very beginning of my career in psychology. Thank you for nurturing my love for this area of work.

To the management team and care staff at Keys Childcare, thank you for your consistently positive approach towards this project. Thank you to the care staff who took time out of their busy working days to participate in the study, without you this research would not have been possible. A special thanks to Tracey Hopes for all of your help with the data collection; you are a star!

To my wonderful family, thank you for always encouraging me to keep going. Thank you for being so incredibly understanding and for accepting the presence of my laptop at the majority of our gatherings over the last three years! You have always been by my side, supporting me to be the best I can be. Without you I know that I wouldn't have had the opportunity to embark on this journey, let alone complete it.

And finally, to my fellow cohort members, 'The Sisterhood'. Thank you for the many (many) laughs and unfailing friendship, both of which have made the last three years ones I will always cherish.

Word Count

Thesis Abstract	298
Literature Review:	
Abstract	286
Full text (excluding figures and tabulated data)	7701
Empirical Report:	
Abstract	383
Full Text (excluding figures and tabulated data)	8428
Non-mandatory Appendices (excluding figures and tabulated data)	1121
Total word count (excluding thesis abstract, references and mandatory appendices)	17,919

Contents Page

List of tables

Tables marked with an asterisk (*) are tabulated data and thus not included in the word count.

Part one: Literature review

Part two: Research report

List of figures

Part one: Literature review

Part two: Research report

List of appendices

Mandatory appendices are marked with an asterisk (*)

Part one

A Systematic Review of the Psychometric Properties of Informant Rated Global Mental Health Outcome Measures for Young People

Abstract

Background: In the context of high levels of distress being experienced by children and young people, mental health outcome measures are being increasingly used to monitor the efficacy of the interventions put in place to support them and the services providing such interventions. It is important that the measures used for this purpose are psychometrically robust. This review aimed to complement the existing literature on the psychometric properties of self-report outcome measures for children and young people by systematically reviewing the evidence pertaining to the psychometric properties of informant rated mental health outcome measures for children and young people.

Method: PsycINFO, MEDLINE and Web of Science databases were systematically searched to identify articles reporting data pertaining to the psychometric properties of informant rated mental health outcome measures. A novel quality appraisal tool was developed to evaluate the evidence pertaining to the measures' internal consistency, test-retest and inter-rater reliability, construct validity and responsiveness.

Results: A total of 60 published articles were identified; these described 25 measures meeting the inclusion criteria. None of the included measures provided evidence of acceptable rigour in all five of the domains, per the numerical and methodological standards indicated by the quality appraisal tool. Generally, this was a result of an absence of evidence concerning the psychometric properties of the measures; however, there were instances where the available data was suggestive of poor psychometric rigour.

Conclusions: The findings of this review indicated that caution should be employed when using any of the reviewed informant rated measures to assess mental health outcomes for children and young people. A need for further research pertaining to the psychometric properties of these measures, particularly with respect to their responsiveness to change, test-retest and inter-rater reliability was identified.

1. Introduction

It is difficult at the time of writing for any length of time to pass without the publication of a news article describing the high levels of emotional distress being experienced by young people. For example, in the UK, a recent report published by The Children's Society identified that out of 11,000 fourteen year olds surveyed, one in six had self-harmed in the past year (The Children's Society, 2018). It follows that reports of high levels of distress are frequently accompanied by references to the equally high demand being placed on health services to respond (e.g. Siddique, 2018). This level of demand, coupled with limited resources and the known impact of mental health difficulties on the lives of children and young people (e.g. Das *et al*. 2016), points to the importance of ensuring that mental health services are effective. As such, in the U.K. the use of outcome measures to monitor and evaluate the efficacy of mental health interventions for children and young people is central to the national framework, 'Every Child Matters'. This framework goes further to state that such outcome data should be used to aid the development of services and enhance the level of care they provide (Department of Health, 2007).

According to Kwan and Rickwood (2015; p1) 'an outcome measure in mental health care can be defined as a tool used to measure the effect on a person's mental health as a result of health care intervention, plus any additional extra-therapeutic influences'. To serve this function, outcome measures are administered at two or more time points over the course of a therapeutic intervention. Historically, much of the value of mental health outcome measures has been perceived to be their use in conducting research, which further aims to facilitate evidence based practice. In the U.K., the primary example of this approach are the guidelines provided by the National Institute of Health and Clinical Excellence (NICE) on the best intervention approaches to use for

various presentations of distress. As Green and Latchford (2012) point out, although research such as that contributing to NICE guidelines can be helpful in demonstrating what interventions are not helpful, reliance on it to ascertain which interventions might be more helpful than others is problematic. First, the research available at present does not support the notion of a superior therapeutic intervention (e.g. Luborsky *et al*., 2002). Second, the reliance on randomised control trials (RCTs) among the research designed to inform clinical practice results in a failure to take into account the extra-therapeutic factors which have been found to have the most impact on outcomes (Wampold, 2001; Beutler, 2009). Third, Luborsky *et al*. (1999) identified that of the variance reported in such RCTs, 69% was accounted for by the researchers' theoretical affiliations.

Perhaps more helpfully, outcome measures can be used to routinely collect 'real time' evidence concerning the effectiveness of and outcomes associated with interventions in clinical practice – i.e. to gather practice based evidence. Use of outcome measures in this way has been found to have benefits at both a client and service level. With regards to the former, in order to adapt and manage interventions over their course to meet clients' needs, clinicians need to make accurate assessments concerning the process of change, something which research has demonstrated clinical judgement alone cannot be relied upon to do (Hatfield *et al*., 2010). Research has demonstrated that the use of outcome measures to support assessments concerning the process of change can reduce the level of client deterioration over the course of therapy (e.g. Lambert *et al*., 2003). Bickman *et al*. (2011) further identified that using outcome measures in this manner increased the speed of improvement for young people engaging in psychotherapy. In addition to speed of change, the use of outcome measurement systems was found by Harmon *et al*., (2007) to increase the magnitude of progress made in a sample of clients who engaged in one of three feedback systems

compared to a treatment as usual control group. On a service level, the routine collection of outcome data can be used to inform decisions concerning how best to allocate limited funding resources to achieve maximum impact (Hall et al., 2014), a function which has increased significance given the current climate of austerity in the U.K.

The potential for outcome measures to contribute positively to the care of children and young people experiencing distress is evident. It is welcomed therefore that there is a growing number of children's mental health outcome measures being used in clinical practice and research (Deighton *et al*., 2014). However, for these measures to be useful they must be valid, reliable, responsive to change and meaningful to clinicians and service users (Happell., 2008). The Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) is a brief emotional screening questionnaire reputed to be the most widely used measure of its type in the UK (e.g. Vaz *et al*., 2016). The ASEBA (Achenbach System of Empirically Based Assessment, formally known as the Child Behaviour Checklist, CBCL) is a multi-informant tool designed to assess adaptive functioning and behavioural, emotional and social problems (Achenbach & Rescorla, 2001). A systematic review of the psychometric properties of the parent and teacher versions of the SDQ was conducted by Stone *et al.* in 2010. This review reported the SDQ's internal consistency to be acceptable, test-retest reliability to be good and inter-rater reliability to be higher than that reported for other measures of child psychopathology. In addition, the SDQ was said to have moderate to good construct validity and to be able to differentiate between clinical and non clinical samples. Similarly, Achenbach and Rescorla (2001) provide a comprehensive review of the ASEBA and identified it to have moderate to high internal consistency, high inter-rater reliability and test-retest reliability, good construct validity and to be consistently able to differentiate between referred and non-referred samples.

Unfortunately, the psychometric properties of children's mental health outcome measures are not consistently assessed during their development or prior to their implementation in research and clinical practice. For example, the Child and Adolescent Functional Assessment Scale (CAFAS; Hodges, 1989) was designed as a measure of functional impairment and was adopted by more than 20 states in the USA as a tool for documenting outcomes and making decisions regarding treatment eligibility, in addition to being used in numerous pieces of research. A review of the psychometric properties of the CAFAS by Bates (2001) revealed a lack of empirical evidence supporting the use of the measure for the purposes outlined above. Further, several psychometric limitations were identified in the scale. It is important that clinicians and researchers are aware of the psychometric strengths and limitations of the measures they employ so that informed decisions can be made about the appropriate use of such measures and what conclusions can fairly be drawn from them. To this end, Deighton *et al*. (2014) conducted a review of the psychometric properties and suitability for use of child self-report outcome measures. They reviewed 11 measures identified as having potential for use in clinical practice and concluded that each of these measures had strengths and limitations but that no individual measure had sufficient psychometric rigour to suggest that it was able to reliably measure both symptom severity and responsiveness.

The utility and implications of Deighton *et al.'s* (2014) review are apparent, however their focus on self-report measures leaves a gap in the literature regarding the psychometric properties of informant rated measures concerning children's distress and wellbeing. More recently, there has been an increasing emphasis on the importance of the voice of children in their own journey of recovery (Department of Health, 2012).

With this has come an understandable and welcomed rise in the use of self-report outcome measures in the field of children's mental health and evidence in support of children as accurate reporters of their own strengths and difficulties. Despite this there is evidence to suggest that children's reports of their mental health via outcome measures may lack reliability as a result of children struggling to generalise their mental state over a period of time (instead giving an 'in moment' report). Furthermore, young people's representations of the extent of their behavioural difficulties have been found to differ from objective reports (Edelbrock *et al*., 1985; Marsh *et al*., 2005). It has long been reported that concordance between parent and child ratings is lacking: For example, in a meta-analytic review, Achenbach *et al*. (1987) found very low correlations (r = .25) between parent and child ratings. This discrepancy has been widely reported to be indicative of children's ratings of themselves and parents' ratings of their children reflecting uniquely differently information (e.g. Grietens *et al*., 2004). Consequentially, it is a generally held belief that best practice, both clinically and in research, is to gather reports from multiple informants (Hunsley & Mash, 2007). Systematic reviews of mental health outcome measures for young people conducted by Kwan and Rickwood (2015) and Hunter *et al*. (1996) considered both self and informant rated measures, however provided a broad overview of the measures available rather than focusing on their psychometric properties.

1.1 Aims & Objectives

In line with the recognised need to use both self and informant rated measures of young people's mental health and the importance of these measures being psychometrically robust, this review aims to complement the existing literature by reviewing the psychometric properties of informant rated measures of children and young people's mental health. In line with the important role of outcome measurement

in service development and planning, this review focused on global, as opposed to presentation specific, measures of mental health designed to cover broad age ranges in order to facilitate comparisons between services and over time.

For the purposes of this review, mental health outcome measures are defined as questionnaires which seek to provide a measurement of mental health, encompassing 'negative' (i.e. distress) and/or 'positive' (i.e. wellbeing) components. Informants are defined as any person completing a measure concerning a young person apart from the young person themselves. This term includes parents, carers (professional and non-professional), teachers, mental health practitioners and other professionals. 'Young people' are defined as those aged 18 and under.

2. Method

The aim of this review was to assess the psychometric properties of informant rated mental health outcome measures that could be used in routine clinical practice to measure children and young people's mental health in primary and specialist services, in addition to research. It was therefore decided to focus on measures with a global rather than specific orientation across a wide age range

2.1 Search strategy

A set of initial search terms were developed in an attempt to identify all papers exploring the psychometric properties for indirect measures for children, prior to a second search for each of the identified measures in order to ensure the review was comprehensive. The initial search terms were developed through amalgamating terms utilised in previous reviews with similar aims (Deighton *et al.,* 2014: psychometric properties of self-report measures for children; Kwan & Rickwood, 2015: general review of measures for 12 to 25year-olds; Hunter *et al.,* 1996: general review of mental health outcome measures for young people). All three papers employed search terms related to the domains of: 'measurement'; 'mental health'; and 'children'. The search terms from these papers were amalgamated, resulting in the search terms displayed in Table 1.

Table 1: Search Terms

Search category	Associated search terms: The search terms within each category were combined with the Boolean operator, 'OR'. The categories were combined using the Boolean operator, 'AND'.
Measurement	Assessment; checklist; "assessment schedule"; measure*; questionnaire; rating; repository; "rating scale"; scale; screen*; "screening tool"; survey; tool; inventory; instrument
Mental Health	Behav*; conduct; emotion*; feeling*; "mental disorder"; "Mental health"; "mental illness"; mood; prosocial; psych*; adjustment; distress; quality of life; wellbeing; difficult*; resilien*
Children	Child*; paed*; young; youth*; adolescen*

Searches were conducted on 17th July 2018, utilising three databases chosen for their combined breadth of applied disciplines: PsycInfo; Medline and Web of Science (see Appendix F for further justification). Filters restricted returns to journal articles, published in English. The search was restricted to papers published between 1990 and 2018 in order to restrict search results to papers pertaining to measures relevant to current clinical practice.

In line with the aims of this review, articles were included if they pertained to *measures* which:

i. were informant rated global mental health questionnaire based (as opposed to diagnostic interviews or observational rating scales) measures designed for evaluating outcomes in therapy;
ii. were designed for use with a broad age group young people aged 18 and under;

iii. were focused on measuring mental health as opposed to quality of life or physical health
iv. were the most up to date version of a measure (in instances where more than one version had been developed and tested)
v. were **not** designed solely for use in school settings or to be rated by teachers only

Articles relating to measures meeting the inclusion criteria were further included if:

i. they pertained to the psychometric properties of the measure in question
ii. they separated informant and self-report data when reporting psychometric properties of measures with both self-report and informant rated versions
iii. they were published in English
iv. they were published in a journal

Due to the psychometric properties of the SDQ and ASEBA being widely reported and reviewed, these measures were excluded from this review.

2.2 Search results

The searches yielded 6,068, 6,125 and 5,301 results in PsycINFO, Web of Science and Medline respectively. Results were exported into the reference managing programme Mendeley, which identified and removed 12,119 duplicates. The remaining 5,375 articles were then checked against a catalogue of mental health outcomes suitable for children and young people (Wolpert, *et al*. 2009). This check was carried out in

order to ensure the adequacy of the search terms employed. All carer-rated measures within the reference document were represented by at least one returned article. Article titles were screened against the inclusion/exclusion criteria. Papers were retained if it was unclear from the title whether they met the inclusion criteria. The abstracts of the remaining 964 articles were read against inclusion criteria, resulting in a further 859 being excluded. The remaining 105 articles were read in full and a further 47 were excluded. Fifty-eight articles describing 25 measures were retained. A further search was then conducted in PsycInfo for each measure in order to capture any missing articles, including those published prior to the year 1990. This yielded two further articles. In total, a search of three databases yielded 60 articles that met the inclusion criteria; these articles described 25 measures. A Google search was completed to identify any grey literature of relevance, however this yielded no new results. Please see Figure 1 for a flow chart depicting the systematic literature search outlined above.

The measures and the articles relating to them (along with the psychometric properties they consider) are outlined in Table 2. The retained measures were broad with regards to their focus. All included measures considered emotional wellbeing, while some considered behaviour or conduct and interpersonal relationships in addition. A number of measures used diagnostic categories to organise their items into scales. Items were also frequently clustered into 'internalising' and 'externalising' scales. The majority of the included measures were deficit focused, however some were strength based and others considered strengths in addition to difficulties. A small number of the included articles considered the psychometric properties of translated versions of an original measure. These articles were published in English and therefore met the inclusion criteria. Translations included Norwegian, Spanish and German. The majority of included measures were designed for completion by parents or carers, while a minority

were designed only for completion by clinicians. Several measures could be completed by both parental caregivers and clinicians, with teacher ratings being considered in addition in some instances. The included articles provided examples of measures being used in a range of populations and settings, including non-clinical populations (for example, primary care settings), children accessing community mental health services and children in inpatient mental health settings. The included measures were designed for use with children up to the age of 18. Some measures were specifically designed for primary school aged children, while others focused on adolescents. Some measured covered the entire age range.

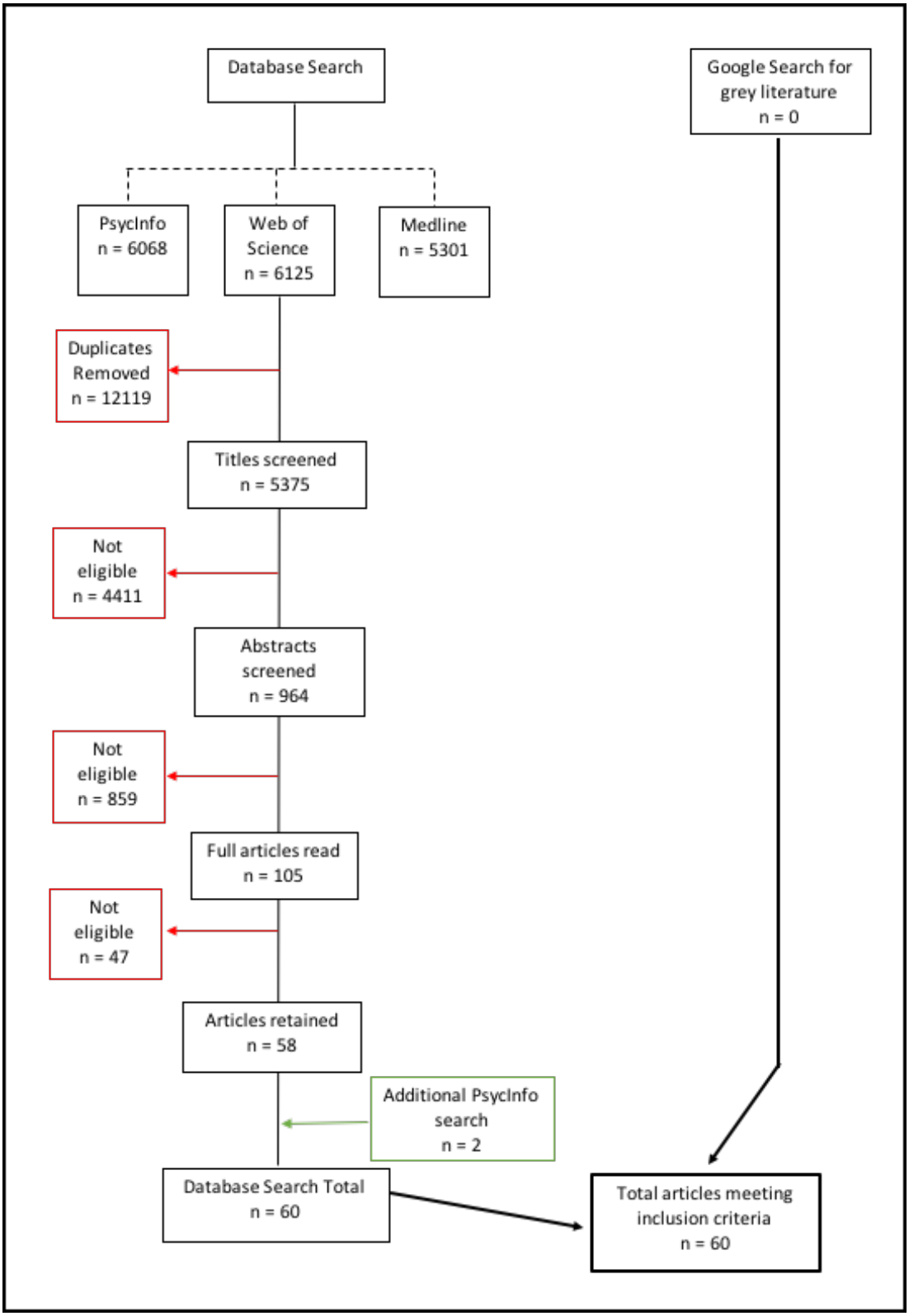


Figure 1: A flow chart depicting the systematic literature search

Table 2: A summary of the included measures and articles

Measure	Brief description	Article(s)	Sample	Internal consistency (Yes/No)	Inter-rater reliability (Yes/No)	Test-retest reliability (Yes/No)	Construct Validity (Yes/No)	Responsive-ness (Yes/No)
The Assessment Checklist for Adolescents (ACA)	ACA is a 105-item carer-report mental health rating scale Age range: 11 - 18 9 scales: non-reciprocal interpersonal behaviour; sexual behaviour problems; food maintenance behaviour; suicide discourse; social instability/behavioural dysregulation; emotional dysregulation/distorted social cognition; dissociation/trauma symptoms; negative self-image and low-confidence.	Tarren-Sweeny (2013)	n = 372 children in long term care rated by carers Age: 11 - 18	Yes	No	No	Yes	No
The Assessment Checklist for Children (ACC)	ACC is a 120-item carer-report mental health rating scale Age range: 4 -11 12 scales: Sexual behaviour; pseudomature interpersonal behaviour; non-reciprocal interpersonal behaviour;	Tarren-Sweeny (2007)	n = 412 children in long term care rated by carers Age: 4 -11	Yes	No	No	Yes	No

	indiscriminate interpersonal behaviour; insecure interpersonal behaviour; anxious–distrustful; abnormal pain response; food maintenance; self-injury; suicide discourse; negative self-image; and low confidence.							
Brief Assessment Checklist for Adolescents (BAC-A)	20 item unidimensional version of the ACC	Denton (2016)	n = 111 children in long term care rated by carers Age: 11-18	Yes	No	No	Yes	No
		Goemans *et al.* (2018)	n = 101 children in foster care rated by carers Age: 12 - 17	Yes			Yes	
		Tarren-Sweeny (2013)	n = 230 children in long term care rated by carers Age: 11-18	Yes	No	No	Yes	No
Brief Assessment Checklist for Children (BAC-C)	20 item unidimensional version of the BCC	Frogley (2016)	n = 178(185 for internal consistency) children in long term alternate care rated by carers	Yes	No	No	Yes	No

			Age: 4 -11					
		Goemans *et al.* (2018)	n = 117 children in foster care rated by carers Age: 4-11	Yes			Yes	
		Tarren-Sweeny (2013)	n = 347 children in long term care rated by carers Age: 4 -11	Yes	No	No	Yes	No
Behaviour Assessment for Children – 2 (Parent Rating Scale) (BASC-2 – PRS)	A measure of adaptive and problem behaviours 134-160 items 14 Subscales: adaptability; aggression; anxiety; attention problems; atypicality; conduct problems; depression; functional communication; hyperactivity; leadership; social skills; somatization; withdrawal and activities of daily living.	Gabrielli *et al.* (2015)	n = 479 rated by carers Age: 8 - 18	Yes	No	No	No	No
Behavioural and Emotional Rating Scale 2 (BERS-2)	52 items strength based assessment of strengths and competencies Age rage: 5-18	Buckley, J. *et al.* (2006)	n = 927 children rated by carers Age: 5 – 18	Yes	No	No	No	No

	5 subscales: interpersonal strengths; intrapersonal strengths; affective strengths; family involvement; school functioning	Gonzalez, J. *et al.* (2006)	n = 927 children rated by parents Age: 0 – 18	No	Yes	No	No	No
		Lambert, M. *et al.* (2015)	n = 7487 children rated by carers Age: 6 – 18	No	No	No	Yes	No
		Mooney, P. *et al.* (2005)	n = 78 for test retest; 85 for construct validity children rated by carers Age: 5 -12	No	No	Yes	Yes	No
		Sointu, E. *et al.* (2015)	n = 334 children rated by 77 teachers Age: 11-17	Yes	No	No	No	No
Brief Problem Monitor (BPM)	Shortened version of the Child Behaviour Check list 19 items 3 Scales: internalising; externalising; attention	Piper *et al.* (2014)	n = 567 children rated by carers Age: 6 - 18	Yes	No	No	Yes	No
		Richter (2014) *Norwegian version*	n = 2582 children rated by mothers Age: 6 - 16	Yes	No	No	Yes	No
Brief Psychiatric	21 items	Gale *et al.* (1986)	n = 28 children (in outpatient clinic) and 20 (in inpatient unit) rated by 3 clinicians	No	Yes	No	No	No

Rating Scale for Children (BPRS-C)	clinician Rated outcome measure 7 scales: behavioural problems; depression; thinking disturbance; psychomotor excitation; withdrawal-retardation; anxiety and organicity		Age: 5 – 18					
		Lachar *et al.* (2001)	n = 547 children (n = 90 for inter-rater reliability) rated by clinicians Age: 3 - 18	Yes	Yes	No	No	No
		McLlhaney *et al.* (2008)	n = 522 children in residential treatment and 396 in foster care rated by 49 clinicians Age: 4 – 18	No	Yes	Yes	Yes	Yes
		Mullins *et al.* (1985)	n = 40 children in psychiatric inpatient unit rated by 3 clinicians Age: 5 - 18	No	Yes	No	No	No
		Shafer (2013)	n = 6712 children (inpatient sample) and 21,459 (community mental health sample) rated by clinicians Age: 3 – 17	Yes	No	No	Yes	No
Brief Screening Measure of Emotional Distress in children (BSMED-C)	A parent rated measure of emotional distress 8 items 2 scales: distress and depression	Parker *et al.* (2001)	n = 20 for inter-rater reliability n = 2071 for factor analysis School children in Singapore Rated by parents	Yes	Yes	No	No	No

			Age: 10 – 12					
Child and Adolescent Behaviour Assessment (CABA)	A brief structured scale to assess 'problem behaviour' 32 items 3 Scales: externalising; internalising and risk behaviour Age: 5 - 18	Morin *et al.* (2016)	n = 32,689 children admitted for psychiatric treatment rated by 'an informant known to patient' Age: 5 - 18	Yes	No	No	Yes	No
Child Adjustment and Parent Efficacy Scale (CAPES)	A measure of child emotional and behavioural problems and parental self-efficacy. 27 items 4 scales: intensity; behaviour; self-efficacy and emotional maladjustment	Mejia *et al.* (2016) *Spanish version*	n = 174 children (n = 50 for test-retest) rated by parents Age: 2 - 12	Yes	No	Yes	Yes	No
		Morawska *et al.* (2014)	n = 374 children rated by parents Age: 2 - 12	Yes	No	No	No	No
Children's Emotional Adjustment Scale (CEAS)	A measure designed to capture emotional adjustment in children. 47 items 4 subscales: temper control; anxiety control; mood repair and social assertiveness	Thorlacius & Gudmund-sson (2014)	N = 606 children rated by mothers Age: 6 - 13	Yes	No	No	Yes	No

Child Symptom Inventory-4 (CSI-4)	The CSI–4 contains symptom categories for the following DSM–IV 'disorders': ADHD, Inattentive; ADHD, Hyperactive–Impulsive; ADHD, Combined; ODD; CD; GAD; social phobia; SAD; MDD; dysthymic disorder; schizophrenia; autistic disorder and Asperger's disorder The CSI–4 also contains single items to screen for simple phobias, obsessions, compulsions, motor tics, vocal tics, enuresis, and encopresis.	Sprafkin *et al.* (2002)	N = 247 boys rated by informants: type unclear Age: 6 - 10	Yes	No	Yes	Yes	No
The Devereux Student Strengths Assessment (DESSA)	A measure of social-emotional competencies related to resilience Age: 5 – 14 8 subscales: self-awareness; social-awareness; self-management; goal-directed behaviour; relationship skills; personal responsibility decision making; optimistic thinking.	LeBuffe *et al.* (2018)	n = 778 children rated by teachers n = 472 children rated by after school staff n = 1244 rated by parents Age: 5 -14	Yes	No	Yes	Yes	No
		Nickerson & Fishman (2009)	n = 94 children rated by teachers n = 133 children rated by parents Age: 6 - 12	No	No	No	Yes	No

Devereux Scales of Mental Disorders (DSMD)	A measure designed to measure broad indicators of psychopathology. Age: 5 – 12 (111 items) Age: 13-18 (112 items) 8 subscales: conduct; attention; delinquency; anxiety; depression; autism; acute problems; externalizing composite; internalizing composite and critical pathology composite.	Curry & Ilardi (2000)	n = 108 young people admitted to inpatient 'psychiatry programme' rated by parents Age: 13-18	No	No	No	Yes	No
		Gimpel & Nagle (1999)	Children divided according into severely emotionally disturbed (SED) and non SED groups. n = 194 (87 SED and 107 non SED) Age: 5-12 n = 190 (84 SED and 106 non SED) Age: 13-18 Rated by teachers and parents	Yes	No	No	No	No
		Reddy, L. *et al.* (2007)	n = 74 students classified as 'regular education' and 74 children classified as having 'emotional disturbance' rated by parents Ages: 5 -18	No	No	No	No	No

		Smith & Reddy (2002)	n = 138 children in an inpatient psychiatric unit rated by parents and teachers. Age: 5-18	No	No	No	Yes	No
Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA)	A measure of 'problems and impairment' Age: 5 - 18 13 items 4 scales: behaviour; impairment; symptoms; social	Ballesteros *et al.* (2018) (Spanish transla-tion)	N = 64 children from psychiatric hospitals rated by 2 clinicians and a parent Age: 7 - 17	Yes	Yes	Yes	Yes	No
		Brann *et al.* (2001)	n = 24 (for inter-rater) and n = 145 (paired ratings for responsiveness) children rated by clinicians Age: 5 - 18	No	Yes	No	No	Yes
		Brann & Coleman (2010)	n = 911children rated by clinicians Mean age: 11.5	No	No	No	No	Yes

Gerralda *et al.* (2000)	n = 248 total rated by clinicians Age: 3 - 18	No	Yes	Yes	No	Yes
Gowers *et al.* (1999)	n = 1276 rated by clinicians Age range unclear	No	Yes	No	No	Yes
Harnett *et al.* (2015)	n = 51 adolescents admitted to an inpatient unit rated by clinicians Age: 12 – 17	Yes	No	Yes	Yes	Yes
Tiffin & Rolling (2012)	n = 1335 children rated by 164 clinicians Age: 6 - 15	Yes	No	No	No	No

		Von Wyl *et al.* (2017) (German transla-tion)	n = 1,553 children rated by clinicians (n = 1408 also had parent completed SDQ) Age: 4 - 17	Yes	No	No	Yes	No
		Yates *et al.* (1999)	n = 248 children rated by clinicians Age: 3 - 18	No	No	No	Yes	No
		Yuan (2015)	n = 32 children from inpatient unit rated by clinicians Age range: 12 – 17	No	No	No	No	Yes
The Nisonger Child Behaviour Rating Form: typical IQ version (NCBRF-TIQ)	An outcome measure for children with average IQ 66 items 2 Scales: positive social and problem behaviour.	Aman *et al.* (2008)	n = 523 rated by parents Age: 5 - 15	Yes	No	No	No	No

Primary Care Mental Health Screening Tool (PCMHS)	Designed to be a broad, low-burden screening tool for mental health difficulties 40 items 8 subscales: inattention; hyperactivity; oppositionality; conduct problems; learning problems; anxiety; depression; autism spectrum problems	Hartung & Lefler (2010)	n = 303 children rated by carers Age: 3 - 12	Yes	No	No	No	No
		Lefler *et al.* (2012)	n = 58 children rated by carers Age: 3 - 8	No	No	No	Yes	No
Paediatric symptom checklist (PSC-17)	A 17 item parent rated measure about young people's psychological functioning	Gardner *et al.* (1999)	n = 18,451 children rated by carers Age: 4 -15	Yes	No	No	No	No
		Gardner *et al.* (2007)	n = 269 children rated by carers Age: 8 - 15	No	No	No	No	No
		Jacobson *et al.* (2018)	n = 6492 children rated by carers Age 5 - 17	No	No	Yes	Yes	No
		Murphy *et al.* (2016)	n = 80,680 children rated by carers Age: 4 - 15 years	Yes	Yes	No	No	No

Symptoms and Functioning Severity Scale (SFSS)	'A global measure of severity' Age: 11-18 33 items Subscales/domains: ADHD; conduct/oppositional disorder; depression and anxiety. Scores reported as total score, internalising and externalising	Athay *et al.* (2012)	n = 760 rated by 686 carers and 710 clinicians. Age: unspecified	Yes	No	No	No	No
SFSS short forms A and B (SFSS-SF)	14 items each (two abridged, non-redundant forms designed for alternate administration every week) No subscales. Total score only	Duppong Hurley *et al.* (2015)	n = 143 children in a psychiatric hospital rated by 53 clinicians Age: 11 - 17	Yes	No	No	No	No
		Gross *et al.* (2015)	n = 143 children in a psychiatric hospital rated by 52 care staff. Age: 11 - 17	Yes	No	No	Yes	No
		Lambert *et al.* (2015)	n = 143 children in a psychiatric hospital rated by 53 clinicians Age: 11 - 17	No	No	No	Yes	No

The Revised Child Anxiety and Depression Scale - Parent Version (RCADS-P)	Questionnaire measuring the reported frequency of various symptoms of anxiety and low mood. 6 Subscales: separation anxiety; social phobia; generalised anxiety; panic; obsessive compulsive Total anxiety and Total low mood scores also calculated Age: Up to 18	Ebesutani *et al.* (2009)	n = 490 children rated by caregivers Age: 6-18	Yes	No	No	No	No
		Ebesutani *et al.* (2010)	n = 1,288 children (n = 90 for test-retest) rated by caregivers Age 8 – 18	Yes	No	Yes	No	No
		Ebesutani, et al. (2015)	n = 307 children rated by caregivers Age: 3 – 17	Yes	No	No	Yes	No
		Park *et al.* (2016) (Spanish version)	n = 85 children rated by caregivers Age: unspecified	Yes	No	No	Yes	No

Revised Ontario Child Health Study scales (OCHS)	Aims to identify externalising and internalising 'disorders' 2 scales containing subscales: internalising (anxiety and depression) and externalising ('ODD', 'ADHD' and 'conduct disorder')	Boyle *et al.* (1993)	n = 1751 ('general population'), n = 1017 children accessing mental health services rated by caregivers Age: 6 – 16	Yes	No	Yes	No	No
The Target Symptom Rating (TSR)	A 13 item multi-informant measure of child and adolescent mental health 2 subscales: emotional problems and behavioural problems	Barber *et al.* (2002)	n = 1723 children admitted to psychiatric hospital (n = 30 for Interrater reliability) rated by 93 clinicians Age: 4 – 21	No	Yes	No	Yes	Yes
Youth Outcome Questionnaire (Y-OQ)	Constructed to track treatment progress Age range: 4-18 64 items 6 subscales: Intrapersonal distress; somatic; interpersonal relations; critical items; social problems & behavioural dysfunction	Dunn *et al.* (2005)	n = 217 (community sample), n = 5132 (clinical sample) rated by caregivers Age: 8 - 16	Yes	Yes	Yes	No	No
		McClendon *et al.* (2011)	n = 136 children rated by caregivers Age: 7 - 14	No	No	No	No	No

2.3 Quality Appraisal

For the purpose of this review, a novel quality appraisal tool was employed. This was developed following a review of available tools for appraising the quality of psychometric studies. The novel tool used here was an amalgamation of the 'Terwee tool' (Terwee *et al*., 2007) and the 'Andresen tool' (Andresen, 2000). Further information on these quality appraisal tools and rationale for the development of a new tool are discussed in Appendix G. The qualities considered by the novel tool and numerical cut-offs for these are described below.

2.3.1 Internal Consistency

Internal consistency is a measure of the degree of homogeneity among items in the scales comprising a questionnaire. This is important for questionnaires that are described as measuring a single concept (e.g. 'wellbeing') using multiple items. Cronbach's alpha is considered an appropriate measure of internal consistency (Terwee *et al*. 2007). Nunnally and Bernstein (1994) outlined that a low Cronbach's alpha can be indicative of a lack of homogeneity and thus poor internal consistency, while a very high Cronbch's alpha can denote a measure with items that are too highly correlated and thus redundant. They proposed Cronbach's alphas ranging from .70 to .90 as being indicative of good internal consistency. Others have suggested that the upper limit for acceptable internal consistency be raised to $\alpha = 0.95$ as a result of some measures they consider subjectively 'good' having scales with high Cronbach's alphas (e.g. Terwee *et al*. 2007). Despite this, at present the upper limit recommended remains at $\alpha = 0.90$ (Streiner, 2003) and as such, for the purpose of this review, a measure was judged to

have good internal consistency if the majority of its subscales possessed Cronbach's alphas ≥ .70 and ≤ .90.

2.3.2 Inter-rater reliability

Inter-rater reliability is a measure of the extent of agreement between two or more raters measuring the same construct. Put another way, it represents the extent to which different raters are able to use a measurement tool to consistently perceive and rate the target construct for the same individual. It is important in this context as it signifies the extent to which variation in the scores obtained on an outcome measure between individuals is representative of differences between those individuals rather than differences between the raters. Historically, inter-rater reliability was measured by considering the correlation between raters' scores (e.g. using Pearson's r), with a correlation of >.70 generally being said to equate to good inter-rater reliability. More recently, it has been identified that desirable measures of reliability (between raters or time points) should reflect both correlation and agreement, as it would be possible for two raters to achieve a high correlation between the scores they gave, but for their level of agreement between ratings to be poor. Intra-class correlation coefficient (ICC) is such a measure and has been widely accepted as a desirable and appropriate measure of reliability (e.g. Koo & Li, 2016). It is important to select the correct form of ICC in order to measure inter-rater reliability successfully. McGraw and Wong (1996) defined 10 types of ICC, which vary according to model, type and definition (see Appendix H). The form of ICC used to calculate inter-reliability will be dependent on the aim of the researcher, therefore a model and type was not be specified in the quality appraisal tool. Koo and Li (2016) do specify however that absolute agreement should always be chosen over consistency when considering inter-rater reliability, as the latter is similarly problematic to the use of Pearson's *r*. Koo and Li (2016) highlight that different forms

of ICC rely on different assumptions and therefore the interpretations that can be drawn from them vary. They specify the importance of reporting the model, type and definition selections along with ICC estimates and their 95% confidence intervals. They go on to specify that the 95% confidence interval of the ICC estimate (rather than the ICC estimate itself) should be judged. For the purposes of this review, a measure was deemed to have good inter-rater reliability when the 95% confidence interval of the absolute agreement ICC was ≥ 0.75 and the model of ICC used was reported and justified.

2.3.3 Test-retest reliability

'Test-retest reliability can be defined as a measure of the reproducibility of the scale, that is, the ability to provide consistent scores over time in a stable population' (Paiva *et al*., 2014; p.2). Test-retest reliability is an important property for mental health outcome measures, which are typically designed to measure change over the course of therapy. A good level of test-retest reliability would indicate that any change in score between a rating at time-1 and time-2 is a reflection of change in the construct being measured, rather than the inconsistency of the rater. As with inter-rater reliability, historical trends of utilising Pearson correlation coefficients have been critiqued for failing to take systematic differences into account (Streiner & Norman, 2003). The ICC is now the most commonly used measure of reliability for continuous measures (Terwee *et al*. 2007). In a similar manner to inter-rater reliability, consideration should be given to the form of ICC used; however, there has been greater consensus for good practice for this domain, with a two-way mixed effects model with an absolute agreement definition being regarded as most appropriate for rating test-retest reliability (Koo & Li, 2016; Shrout & Fleiss, 1979). The 95% confidence interval ICC cut-offs also apply to the measurement of test-retest reliability. For the purposes of

this review, a measure was said to have good test-retest reliability when the 95% confidence interval of the absolute agreement ICC was ≥ 0.75 and where a two-way mixed effects model was employed.

2.3.4 Construct Validity

Construct validity refers to the extent to which scores on one measure relate to scores on another measure in a manner that is theoretically expected (Kirschner & Guyatt, 1985). This is an important quality in an outcome measure as it increases confidence that the measure is able to capture the construct it is designed to measure and adds weight to the conclusions that can be drawn from it. It is important that hypotheses regarding the nature of the relationships between the measure under examination and related measures are pre-defined and specific (Terwee *et al*. 2007). At present, there is no agreed upon method of hypothesis testing for the purpose of assessing construct validity, though ascertaining convergent (the degree to which two measures of constructs which should theoretically be related are related in reality) or divergent (the degree to which two measures of constructs which should be theoretically be unrelated are unrelated in reality) validity is a common example. Terwee *et al*. (2007) propose that a positive rating for construct validity should be given where hypotheses are specified in advance and the results gained correspond with these hypotheses at least 75% of the time. Although a rationale was not provided, their criteria have been employed in several studies (e.g. Molland *et al*., 2018) and were utilised for the present review to make decisions about if a paper provided evidence of good construct validity. Where correlation analysis was used, Cohen's (1998) criteria were employed and $r > .5$ (said to denote a large correlation) was set as the cut-off. Where the measure assessed was an adapted version of a previous measure (for example a shortened version of an original measure), comparison between the adapted measure and the original was not considered adequate evidence of construct validity.

2.3.5 Responsiveness

Husted *et al*. (2000) suggested that responsiveness should be separated into two constituent parts: 'internal' and 'external'. They defined internal responsiveness as 'the ability of a measure to change over a particular pre-specified time frame' in response to an intervention (p.1). External responsiveness is said to 'reflect the extent to which changes in a measure over a specified time frame relate to corresponding changes in a reference measure' (p.1), with the emphasis placed on the relationship between the change in the measure and the change in the external standard being emphasised. The ability of an outcome measure to detect change is clearly very important. Husted *et al.* (2000) conducted a critical review of the methods for assessing responsiveness. They concluded that repeated measures analyses (e.g. t-tests, Wilcoxon Signed rank) were problematic due to their reliance upon sample size causing problems for comparisons between studies, and instead favoured, for internal responsiveness, the standardised response mean (SRM; a type of effect size). SRM values of .80, .50 and .20 have been suggested to represent large, moderate and small responsiveness respectively (e.g. Beaton *et al*., 1997). Husted *et al*. (2000) proposed correlation analyses were well suited for assessing external responsiveness. For the purposes of this review, a measure was said to be responsive if it reported SRM values of ≥ 0.5 (for internal responsiveness) or a correlation between change in the measure and change in an external standard of ≥ 0.5 (for external responsiveness). In any cases of conflict between these two methods of rating in the same article, the more favourable evidence was rated.

2.3.6 Scoring

Measures were awarded a maximum score of one for each category outlined above. The maximum score any measure could achieve was therefore five. Measures were scored '1' in a category if any of the articles pertaining to their psychometric properties reported data meeting the criteria specified above. Measures were scored '0' in a category if there was evidence that the measure did not meet the minimum criteria. Measures were coded 'AE' (Absence of Evidence) in a category if there was an absence of good quality evidence from which it could be determined if they did or did not meet minimum criteria. In addition to the numerical scoring, the tables that follow in the Results section use a colour-coding system to make it more evident whether a study indicated: Evidence a quality criteria had been met (green); evidence that a measure did not meet the minimum criteria (red); an absence of evidence (amber). In addition, the psychometric properties of the SDQ and ASEBA (as stated in the aforementioned reviews of these measures) are reported alongside the measures included in this review for purposes of comparison. Data pertaining to these measures are highlighted in blue.

3. Results

3.1 Internal Consistency

Twenty-four of the 25 measures had studies reporting internal consistency data. The TSR did not provide any internal consistency data. Of the 24 measures, 15 were able to demonstrate adequate internal consistency according to the criteria. The articles pertaining to the Y-OQ, SFSS, BSMED-C and BASC-2 did not report Cronbach's alpha, while the HoNOSCA, CABA, CSI-4, CEAS and NCBRF-TIQ reported Cronbach's alphas outside of the acceptable margins, most commonly above .90. Overall, 15 measures had adequate levels of internal-consistency and were awarded one point. The ASEBA demonstrated evidence of adequate internal consistency according to the criteria. The SDQ also had evidence of adequate internal consistency for the teacher rated measure, but not for the parent rated version. There was no observable change in the number of articles reporting adequate internal consistency over time, indicating that there had not been an improvement in the quality of the measures with regards to their internal consistency over time. Please see Table 3 for a summary of these findings.

Table 3: Internal Consistency Findings

Measure	Article	Meet criteria (yes/no)	Rationale	Point for measure (1/0/AE)
Assessment Checklist for Adolescents (ACA)	Tarren- Sweeny (2013)	Yes	Scale alphas for 7 of the 9 scales fell within the acceptable range. For two scales (negative self-image and social instability) α > .90	1
Assessment Checklist for Children (ACC)	Tarren- Sweeny (2007)	Yes	Scale alphas ranged from α = .70 - .89	1
Behaviour Assessment for Children – 2 (Parent Rating Scale) (BASC-2 – PRS)	Gabrielli *et al*. (2015)	No	Cronbach's alpha not reported	AE
Behavioural and Emotional Rating Scale 2 (BERS-2)	Buckley *et al*. (2006)	No	Cronbach's alpha not reported	1
	Sointu *et al*. (2015)	Yes	Scale alphas for 3 of the 5 scales fell within the acceptable range. For 2 scales (interpersonal strengths and school functioning) α > .90	
Brief Assessment Checklist for Adolescents (BAC-A)	Tarren- Sweeny (2013)	Yes	Unidimensional measure. α = .87	1
	Denton (2016)	Yes	Unidimensional measure. α = .89	
	Goemans *et al*. (2018)	Yes	Unidimensional measure. α = .87	
Brief Assessment Checklist for Children (BAC-C)	Tarren- Sweeny (2013)	Yes	Unidimensional measure. α = .89	1

	Goemans *et al*. (2018)	Yes	Unidimensional measure. α = .89	
	Frogley (2016)	Yes	Unidimensional measure. α = .89	
Brief Problem Monitor (BPM)	Piper *et al*. (2014)	Yes	Scale alphas ranged from α = .79 - .87	1
	Richter (2014) *Norwegian version*	Yes	Scale alphas ranged from α = .71 - .77	
Brief Psychiatric Rating Scale for Children (BPRS-C)	Lachar *et al*. (2001)	Yes	Cronbach's alpha for 3 scales (Thinking Disturbance; Psychomotor Excitation and Anxiety) fell below the cut off. For the remaining 7 scales, α = .71 - .87	1
	Shafer (2013)	Yes	Cronbach's alpha for 1 scale (anxiety) fell below the cut-off. For the remaining scales, α = .73 - .85	
Brief Screening Measure of Emotional Distress in children (BSMED-C)	Parker *et al*. (2001)	No	Cronbach's alpha not reported	AE
Child Adjustment and Parent Efficacy Scale (CAPES)	Morawska *et al*. (2014)	Yes	Cronbach's alpha for 1 scale (self-efficacy) was above the cut off. For the remaining 2 scales, α = .70 & .90	1
	Mejia *et al*. (2016) *Spanish version*	No	Unable to report Cronbach's alpha due to assumption violations	
Child and Adolescent Behaviour Assessment (CABA)	Morin *et al*. (2016)	No	Cronbach's alpha for 1 scale (risk) fell below the cut off and for 1 scale (externalising). For the remaining 1 scale, α = .87	0

Child Symptom Inventory-4	Sprafkin *et al.* (2002)	No	Cronbach's alpha for 4 scales (MDD, dysthymic; schizophrenia & aspergers) was below the cut off. For a further 4 scales, Cronbach's alpha was above the cut off (ADHD:I ADHD:C ADHD:HI and ODD) For the remaining 4 scales, α = .73 - .79	0
Children's Emotional Adjustment Scale (CEAS)	Thorlacius & Gudmundsson (2014)	No	Scale alphas ranged from α = .92 - .95, above the cut off.	0
Devereux Scales of Mental Disorders (DSMD)	Gimpel & Nagle (1999)	Yes	For 5-12 year olds, 1 scale alpha was above the cut off (conduct). For the remaining 5 scales, α = .79 - .88. For 13-18 year olds, 2 scale alphas were above the cut off (conduct and depression). For the remaining 4 scales, α = .76 - .87.	1
The Devereux Student Strengths Assessment (DESSA)	LeBuffe *et al.* (2018)	Yes	Scale alphas ranged from α = .82 - .89 for parent raters	1
Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA)	Tiffin & Rolling (2012)	No	Cronbach's alpha for 1 out of 5 within adequate range	0
	Harnett *et al.* (2015)	No	Cronbach's alpha range from 0.10 – 0.48.	
	Von Wyl *et al.* (2017) *(German)*	No	Individual scale alphas not reported	
	Ballesteros *et al.* (2018) *(Spanish)*	No	Individual scale alphas not reported	

Nisonger Child Behaviour Rating Form: typical IQ version (NCBRF-TIQ)	Aman *et al.* (2008)	No	Four out of six scale alphas were above the cut off. The remaining two scale alphas were .90 (hyperactive) and .83 (overly sensitive)	0
Paediatric Symptom Checklist (PSC-17)	Gardner *et al.* (1999)	Yes	Scale alphas ranged from α = .79 - .83	1
	Murphey *et al.* (2016)	Yes	Scale alphas ranged from α = .78 - .82	
Primary Care Mental Health Screening Tool (PCMHS)	Hartung & Lefler (2010)	Yes	Cronbach's alpha for 4 scales fell within desired range. For remaining 4 scales (inattention, hyperactivity, oppositionality and learning problems), α > .90	1
The Revised Child Anxiety and Depression Scale - Parent Version (RCADS-P)	Ebesutani *et al.* (2009)	Yes	Scale alphas ranged from α = .81 - .84	1
	Ebesutani *et al.* (2010)	Yes	Scale alphas ranged from α = .71 - .84	
	Park *et al.* (2016) (Spanish version)	Yes	Scale alphas ranged from α = .75- .86	
	Ebesutani, et al. (2015)	Yes	'Younger group' scale alphas ranged from α = .70- .90. 'Older group' scale alphas ranged from α = .76- .89. 'Post institutionalized group' scale alphas ranged from α = .79- .88.	
Revised Ontario Child Health Study scales (OCHS)	Boyle *et al.* (1993)	Yes	Among 6-11 year olds, 2 scale alphas fell below the cut off	1

			(Conduct & Overanxious). The remaining 4 scale alphas ranged from α = .72- .90. Among 12-16 year olds, scale alphas ranged from α = .70- .89.	
SFSS Short forms (A and B) (SFSS-SF)	Gross *et al*. (2015)	No	Unidimensional measure. Cronbach's alpha = .78 and .82 for short forms A and B respectively.	1
	Duppong Hurley *et al*. (2015)	Yes	Unidimensional measure. Cronbach's alpha = .78 and .82 for short forms A and B respectively.	
Symptoms and Functioning Severity Scale (SFSS)	Athay *et al*. (2012)	No	Cronbach's alpha not reported	AE
Youth Outcome Questionnaire (Y-OQ)	Dunn *et al*. (2005)	No	Cronbach's alpha for subscales not reported for parent measure separate from self-report measure	AE
SDQ	Stone *et al*. (2010)	For the parent version, 4 of the 5 subcale weighted mean alphas fell below α = .70. For the teacher version, only one subscale alpha (Peer Problems) fell outside the cut off.		
ASEBA	Achenbach & Rescorla (2001)	For the parent rated measure (Child behaviour checklist; CBCL), of the 19 subscales, 3 had alphas below the cut off. For the teacher rated measure (Teacher's report form; TRF), 2 of the 19 scales had alphas below the cut off.		

3.2 Inter-rater reliability

Of the 25 measures, six had articles reporting data concerning their inter-rater reliability (BERS-2; BPRS-C; BSMED-C; HoNOSCA; TSR and Y-OQ), yet no study provided evidence to satisfy the criteria. Four of the 11 articles (Dunn *et al*., 2005; Gale *et al*., 1986; Gonzalez, J. *et al*., 2006; Lachar *et al*., 2001) did not use ICC to calculate inter-rater reliability. Where ICC was used, only one article (Barber *et al*., 2002; TSR) reported the type and model used. Only one article reported 95% confidence (Ballesteros *et al*., 2018; HoNOSCA). No measure was judged to provide evidence of good inter-rater reliability per the criteria. Three measures had articles indicative or poor inter-rater reliability (BPRS-C; BSMED-C and HoNOSCA) and were given a score of zero. Stone *et al*. (2010) reported data taken from studies addressing the inter-rater reliability of the SDQ using correlation analysis. Achenbach and Rescorla (2001) provided evidence of adequate inter-rater reliability for the CBCL as assessed using ICC, however confidence intervals were not reported. There was no observable change in the number of articles reporting adequate IRR over time, indicating that there had not been an improvement in the quality of the measures with regards to their IRR over time. Please see Table 4 for a summary of these findings.

Table 4: Inter-rater reliability findings

Measure	Article	Meet criteria (yes/no)	Rationale	Point for measure (1/0/AE)
Behavioural and Emotional Rating Scale 2 (BERS-2)	Gonzalez, J. *et al.* (2006)	No	Reported SEM rather than ICC	AE
Brief Psychiatric Rating Scale for Children (BPRS-C)	Gale *et al.* (1986)	No	Reported Spearman-Brown correlations, not ICC	0
	Lachar *et al.* (2001)	No	Reported *r*. Method of analysis unclear. Not ICC	
	Mullins *et al.* (1985)	No	ICC model and type not reported. 95% confidence intervals not reported. ICCs were reported for each factor, of the seven factors four had ICCs below the cut-off.	
Brief Screening Measure of Emotional Distress in children (BSMED-C)	Parker *et al.* (2001)	No	ICC model and type not reported. 95% confidence intervals not reported. Mean ICC for measure was .54, below the cut off.	0
Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA)	Ballesteros *et al.* (2018) *(Spanish)*	No	ICC calculated with an absolute agreement definition and confidence intervals were reported. However , model was not reported. Aside from this, ICC .898 -.962 was found at time 1 and .936 - .937 at time 2. All scores were above the cut-off.	0

	Brann *et al.* (2001)	No	Absolute agreement ICC was .52 – below the desired cut-off. No confidence intervals reported	
	Gerralda *et al.* (2000)	No	ICC was only above the cut off for the 'psychiatric symptoms' scale. ICC model and type not reported. 95% confidence intervals not reported	
	Gowers *et al.* (1999)	No	ICC model and type not reported. 95% confidence intervals not reported. Aside from this, ICC > .75 was reported for 11 out of 14 subscales.	
The Target Symptom Rating (TSR)	Barber *et al.* (2002)	No	ICC model and type was reported, however 95% confidence intervals not reported. Mean ICC for measure was .74, just below the .75 cut off. Measured separately, the ICCs for the two scales were above the cut off (.78 and .87).	AE
Youth Outcome Questionnaire (Y-OQ)	Dunn *et al.* (2005)	No	Reported Pearson's *r*. Was above cut-off for *r* of .70.	AE
SDQ	Stone *et al.* (2010)	Reported weighted mean of Pearson's *r* for total difficulties score as .44*		
ASEBA	Achenbach & Rescorla (2001)	Reported for CBCL only. ICC model and type was reported. Confidence intervals not reported. ICC for the total competencies score was reported to be .93. ICC for the total problems score was reported to be .96.		

*Inter-rater agreement between teachers and parents. For a meta-analysis of such data, a cut off of $r = .27$ is widely accepted as a rule of thumb. Achenbach (1987).

3.3 Test-retest reliability

Of the 25 measures, nine had articles reporting data concerning their test-retest reliability: BERS-2; CAPES; CSI-4; DESSA; HoNOSCA; PSC-17; RCADS-P; OCHS and Y-OQ. Of the nine measures, none had an article demonstrating an adequate level of test-retest reliability according to the criteria. Only four of the 12 articles used ICC to assess inter-rater reliability (Ballesteros *et al.*, 2018, HoNOSCA; Jacobson *et al*., 2018, PSC-17; Mejia *et al*., 2016, CAPES; Murphy *et al*., 2016, PSC-17). Six of the remaining articles reported Pearson's *r,* while in the case of two articles (Ebesutani *et al*., 2010, RCADS-P & LeBuffe *et al.,* 2018, DESSA) the method of analysis was unclear. While ICC was not used by Gerralda *et al.* (2000) in assessing the test-retest reliability of the HoNOSCA, the finding of $r < .70$ can still be taken to be indicative of poor test-retest reliability among the sample considered. Similarly, Jacobson et al. (2018; PSC-17) used ICC with a consistency definition and found ICC to be below the cut off, taken as indicative of poor test-retest reliability. Given the more stringent requirements of ICC with an absolute agreement definition, if acceptable inter-rater reliability could not be demonstrated using correlational analyses measuring consistency only, this finding would be reinforced if using ICC with an absolute agreement definition. Where ICC was used, only Mejia *et al.* (2016; CAPES) reported the type and model used in addition to the ICC confidence intervals. Unfortunately, in this instance, and in the case of Ballesteros *et al.* (2018; HoNOSCA) a consistency definition was used. No measure was judged to provide evidence of good test-retest reliability per the criteria. The HoNOSCA and PSC-17 were given a score of zero. Stone *et al*. (2010) reported data taken from studies addressing the test-retest reliability of the SDQ using correlation analysis. Achenbach and Rescorla (2001) provided

evidence of adequate test-retest reliability for the CBCL as assessed using ICC, however confidence intervals were not reported. There was no observable change in the number of articles reporting adequate test-retest reliability over time, indicating that there had not been an improvement in the quality of the measures with regards to their test-retest reliability over time. However, it was noted that there appeared to be a move towards using ICC to explore test-retest reliability over time, with only one article published since 2016 not using this method of analysis. Unfortunately, issues regarding the reporting of type, definition and confidence intervals pervaded. Please see Table 5 for a summary of these findings.

Table 5: Test-retest reliability findings

Measure	Article	Meet criteria (yes/no)	Rationale	Point for measure (1/0/AE)
Behavioural and Emotional Rating Scale 2 (BERS-2)	Mooney *et al.* (2005)	No	Pearson's *r.* used instead of ICC. (All scales above .70 cut off for *r*)	AE
Child Adjustment and Parent Efficacy Scale (CAPES)	Mejia *et al.* (2016)	No	ICC used. Model, type and 95% confidence intervals stated, however consistency rather than absolute agreement was used. Only 'Behavioural and Emotional Problems' scale had 95% confidence interval ICC of > .75.	AE
Child Symptom Inventory-4 (CSI-4)	Sprafkin *et al.* (2002)	No	Pearson's *r.* used instead of ICC. (Between 1 and 4 months, *r* for the majority of scales <.70).	AE
The Devereux Student Strengths Assessment DESSA	LeBuffe, P. *et al.* (2018)	No	ICC not used. Method of analysis unclear.	AE
Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA	Ballesteros *et al.* (2018)	No	ICCs calculated using a consistency definition. Average ICC was above cut off.	0
	Harnett *et al.* (2015)	No	Pearson's *r.* used instead of ICC. (*r* above .70 cut off)	
	Gerralda *et al.* (2000)	No	Pearson's *r.* used instead of ICC. *r* = .69, just below accepted cut off	
Paediatric Symptom Checklist (PSC-17)	Jacobson *et al.* (2018)	No	ICC model and type specified but 95% confidence intervals not specified. ICCs fell below cut-off.	0

	Murphy *et al*. (2016)	No	ICC model and type not specified. 95% confidence intervals not reported. Aside from this, ICCs above cut-off.	
Revised Child Anxiety and Depression Scale - Parent Version (RCADS-P)	Ebesutani *et al*. (2010)	No	ICC not used. Method of analysis unclear.	AE
Revised Ontario Child Health Study scales (OCHS)	Boyle *et al*. (1993)	No	Pearson's *r.* used instead of ICC. For 6-11 year olds, >.70 for 3 subscales. For 12-16 year olds, *r* <.70 for 4 out of 6 scales.	AE
Youth Outcome Questionnaire (Y-OQ)	Dunn *et al*. (2005)	No	Pearson's *r.* used instead of ICC. *r* was above cut off of .70.	AE
SDQ	Stone *et al*. (2010)	Pearson's r was reported. Weighted mean for total difficulties was *r* = .76 for parent version and *r* = .84 for teacher version.		
ASEBA	Achenbach & Rescorla (2001)	ICC model and type reported. Confidence intervals were not reported. Test-retest reported for CBCL only. ICC for total competencies score was reported to be 1.00 ICC for total problems score was .95.		

3.4 Construct Validity

Twenty of the 25 measures had articles reporting data concerning their construct validity. The Y-OQ, SFSS; OCHS, NCBRF-TIQ and BSMED-C did not have any articles pertaining to their construct validity. Of the 20 measures, 11 had articles demonstrating evidence of adequate construct validity according to the criteria. These were: ACA (Tarren- Sweeny, 2013); ACC Tarren- Sweeny, 2007); BAC-A (Denton, 2016 & Goemans *et al*. 2018)); BAC-C (Frogley, 2016 & Goemans *et al*. 2018); BASC-2 (reported in Lefler *et al*., 2012); BERS-2 (Mooney, *et al*., 2005); CSI-4 (reported in Lefler *et al*., 2012); DSMD (Curry & Ilardi, 2000; Smith & Reddy 2002); PCMHS (Lefler *et al*., 2012) and RCADS-P (Ebesutani, et al., 2015 & Park *et al*.,2016); SFSS-SF (Lambert *et al*., 2015). Of note, the BASC-2-PRS and CSI-4 had evidence of adequate construct validity as a result of being used as comparison measures in the exploration of the PCMHS (please see 'points carried forward column in Table 6). Hypotheses not being stated in advance of analysis was the most common reason for articles not meeting the criteria. Correlation coefficients falling below the criteria cut-off was an issue for the HoNOSCA and TSR in particular. Where a measure was a shorter or adapted version of an original measure, a common issue was the use of the original measure as a comparison. The SDQ and ASEBA had evidence of adequate construct validity according to the criteria employed here. There was no observable change in the number of articles reporting adequate construct validity over time, indicating that there had not been an improvement in the quality of the measures with regards to their construct validity over time. Please see Table 6 for a summary of these findings.

Table 6: Construct Validity findings

Measure	Article	Comparison measure(s)	Met criteria (yes/no)	Rationale if not meeting criteria	Point carried forward	Point for measure (1/0/AE)
The Assessment Checklist for Adolescents (ACA)	Tarren- Sweeny (2013)	CBCL	Yes	-	-	1
The Assessment Checklist for Children (ACC)	Tarren- Sweeny (2007)	CBCL	Yes	-	-	1
Brief Assessment Checklist for Adolescents (BAC-A)	Denton (2016)	SDQ	Yes	-	-	1
	Goemans *et al.* 2018)	SDQ	Yes	-		
	Tarren- Sweeny (2013)	ACA	No	Used ACA as comparison measure		
Brief Assessment Checklist for Children (BAC-C)	Frogley (2016)	SDQ	Yes	-		1
	Goemans *et al.* (2018)	SDQ	Yes	-		
	Tarren- Sweeny (2013)	ACC	No	Used ACC as comparison measure		

Behaviour Assessment for Children – 2 (Parent Rating Scale) (BASC-2 – PRS	See Lefler *et al* (2012) and	PCMHS	Yes	-	1	1
Behavioural and Emotional Rating Scale 2 (BERS-2)	Lambert, *et al.* (2015)	CBCL Columbia Impairment Scale (CIS)	No	Hypotheses not stated in advance	-	1
	Mooney, *et al.* (2005)	SSRS (Social Skills Rating System)	Yes	-		
Brief Problem Monitor (BPM)	Piper *et al.* (2014)	CBCL	No	BMP is an abbreviated version of the CBCL, TRF and YSR. CBCL was used as a comparison measure.	-	AE
	Richter (2014)	CBCL TRF	No	BMP is an abbreviated version of the CBCL, TRF and YSR. CBCL and TRF were used as comparison measures.		
Brief Psychiatric Rating Scale for Children (BPRS-C)	Shafer (2013)	CBCL	No	Hypotheses not stated in advance	-	AE
Child and Adolescent Behaviour Assessment (CABA)	Morin *et al.* (2016)	**BPRS-C**	No	Hypotheses not stated in advance	-	AE

Child Adjustment and Parent Efficacy Scale (CAPES)	Mejia *et al*. (2016)	SDQ	No	Hypotheses not stated in advance	-	AE
Children's Emotional Adjustment Scale (CEAS)	Thorlacius *& Gudmundsson* (2014)	SDQ **RCADS-P**	No	Hypotheses not stated in advance	-	AE
Child Symptom Inventory-4 (CSI-4)	See Lefler *et al* (2012)	**PCMHS**	Yes	-	1	1
	Sprafkin *et al.* (2002)	CBCL	No	Hypotheses not stated in advance		
The Devereux Student Strengths Assessment (DESSA)	LeBuffe *et al.* (2018)	Rating Scale of Impairment	No	Hypotheses not stated in advance	-	AE
	Nickerson & Fishman (2009)	BERS-2 **BASC-2 PRS**	No	Hypotheses not stated in advance		
Devereux Scales of Mental Disorders (DSMD)	Curry & Ilardi. (2000)	CBCL	Yes	-	-	1
	Smith & Reddy (2002)	CBCL BASC	Yes	-		
Health of the Nation Outcome Scales for Children and Adolescents	Ballesteros *et al.* (2018)	CGAS	No	Hypotheses not stated in advance	-	0
	Harnett *et al.* (2015)	Paddington Complexity Scale (PCS)	No	Hypotheses not stated in advance and correlation coefficients		

(HoNOSCA)		Frequency of 'risk incidents'		below cut off.		
	Von Wyl *et al.* (2017)	SDQ	No	Correlation coefficients below cut-off		
	Yates *et al.* (1999)	PCS SDQ CGAF	No	Hypotheses not stated in advance and correlation coefficients below cut off in several instances		
Primary Care Mental Health Screening Tool (PCMHS)	Lefler *et al* (2012)	Computerized Diagnostic Interview Schedule for Children–IV (C-DISC-IV) **BASC-2 PRS** **CSI-4**	Yes	-	-	1
Paediatric symptom checklist PSC-17	Jacobson *et al.* (2018)	SCARED	No	Correlation coefficients below cut off in several instances	-	0
SFSS short forms A and B (SFSS-SF)	Gross *et al.* (2015)	Short forms A&B SFSS	No	Used SFSS as comparison measure. Correlated both versions of the short form	-	1
	Lambert *et al.* (2015)	CBCL	Yes	-		

The Revised Child Anxiety and Depression Scale - Parent Version (RCADS-P)	Ebesutani, et al. (2015)	CBCL	Yes	-	-	1
	Park *et al.* (2016)	Brief Problem Checklist (BPC) SDQ	Yes	-		
The Target Symptom Rating (TSR)	Barber *et al.* (2002)	CAFAS CBCL	No	Correlation coefficients below cut-off 50% of the time.	-	0
SDQ	Stone *et al.* (2010)	Correlation of .76 for parent and teacher versions between SDQ and CBCL total scores. Other correlations between subscales were as expected and all above the $r = .5$ cut off. Many other correlations with other measures were discussed. Only correlations below cut off reported were those with HoNOSCA.				
ASEBA	Achenbach & Rescorla (2001)	CBCL and TRF subscales correlated with corresponding Conners subscales, all correlations were above the $r = .5$ cut off used here. CBCL and TRF subscales also corresponded with expected BASC subscales. Again, all correlations were greater than $r = .5$.				

*Note: Comparison measures in **bold** are those which are being reviewed here. Where a measure uses another reviewed measure as a comparison, any evidence of convergent validity is also attributed to the comparison measure and points are noted in in the 'points carried forward' column.*

3.5 Responsiveness

Of the 25 measures, three had articles reporting data concerning their responsiveness: BPRS-C (one article); HoNOSCA (six articles) and TSR (one article). The two articles pertaining to the BPRS-C (McLlhaney *et al*., 2008) and TSR (Barber *et al*., 2002) did not use methods of analysis compatible with the criteria; this was also the case for one of the articles reporting on the HoNOSCA (Brann & Coleman, 2010). Two of the articles concerning the responsiveness of the HoNOSCA appeared to use methods of analysis that corresponded with Husted *et al.* (2000)'s recommendations for assessing external responsiveness; however, a lack of clarity in the reporting of these analyses made it impossible to judge whether they met the suggested cut-off. Brann *et al*. (2001) found the external responsiveness of the HoNOSCA (with clinician rated change as a reference measure) to fall just below the $r = .5$ cut off ($r = .46$). In contrast, Gerralda *et al*. (2000) and Yuan (2015) found acceptable levels of responsiveness for the HoNOSCA according to the criteria. The HoNOSCA was therefore awarded one point. Neither the review concerning the SDQ nor the ASEBA had evidence pertaining to their responsiveness. There was no observable change in the number of articles reporting adequate responsiveness over time, indicating that there had not been an improvement in the quality of the measures with regards to their responsiveness over time. Please see Table 7 for a summary of these findings.

Table 7: Responsiveness findings

Measure	Article	Meet criteria (yes/no)	Rationale if not meeting criteria	Point for measure (1/0/AE)
Brief Psychiatric Rating Scale for Children (BPRS-C)	McLlhaney *et al*. (2008)	No	Used ANOVA to measure internal responsiveness.	AE
Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA)	Brann *et al*. (2001)	No	Measures correlation with a clinician rated measure of change and change in HoNOSCA score between time 1 and time 2. There was a significant correlation, however, r = .46, below the cut off.	1
	Brann & Coleman (2010)	No	Used MANOVA to assess relationship between treatment status, time and HoNOSCA score.	
	Gerralda *et al*. (2000)	Yes	A significant correlation between change as measured by the CGAS and HoNOSCA (r = .51, p > .001).	
	Gowers *et al*. (1999)	No	Analysis of correlation between change in HoNOSCA score between time 1 and time 2 and clinician rated global judgement of change, however no correlation coefficient reported.	
	Harnett *et al*. (2015)	No	For internal responsiveness, measured mean change in score over time. For external responsiveness, spoke of correlating change in HoNOSCA with change in clinician rated measure of global change, however results of this analysis were not reported.	

	Yuan (2015)	Yes	A significant correlation was found between change in HoNOSCA scores and clinician rated improvement scores (r =0.916, p < .001	
The Target Symptom Rating (TSR)	Barber *et al*. (2002)	No	Measured internal responsiveness using "a doubly multivariate repeated measures analysis of variance"	AE
SDQ	Stone *et al*. (2010)	-		
ASEBA	Achenbach & Rescorla (2001)	-		

3.6 Total Scores

The highest score achieved by any measure was two. A score of two was awarded to: ACA; ACC; BAC-A; BAC-C; DSMD; PCMHS and RCADS-P. Eleven measures were given a score of one: BASC-2; BERS-2; BPM; BPRS-C; CAPES; CSI-4; DESSA; HoNOSCA; PSC-17; SSFS-SF and OCHS. The remaining seven measures were given a score of zero: BSMED-C; CABA; CEAS NCBRF-TIQ; SFSS; TSR and Y-OQ. Please see Table 8 for a summary of these findings. The measure with evidence spanning the most number of domains was the HoNOSCA, however there was evidence of inadequacy in four of the domains considered, resulting in only one point being awarded to this measure. Eight other measures (BPRS-C; BSMED-C; CABA; CEAS; CSI-4; NCBRF-TIQ; PSC-17 and TSR) demonstrated evidence of poor quality in one of the five domains. Two measures (SFSS and Y-OQ) had an absence of evidence from which to assess quality across all five domains considered.

Table 8: A summary of total measure scores

Measure	Internal consistency	Inter-rater reliability	Test-retest reliability	Construct Validity	Responsive-ness	Total Score (no. of domains with evidence)
The Assessment Checklist for Adolescents (ACA)	1	AE	AE	1	AE	2 (2)
The Assessment Checklist for Children (ACC)	1	AE	AE	1	AE	2 (2)
Brief Assessment Checklist for Adolescents (BAC-A)	1	AE	AE	1	AE	2 (2)
Brief Assessment Checklist for Children (BAC-C)	1	AE	AE	1	AE	2 (2)
Behaviour Assessment for Children – 2 (Parent Rating Scale) (BASC-2 – PRS)	AE	AE	AE	1	AE	1 (1)
Behavioural and Emotional Rating Scale 2 (BERS-2)	1	AE	AE	1	AE	1 (1)
Brief Problem Monitor (BPM)	1	AE	AE	AE	AE	1 (1)

Brief Psychiatric Rating Scale for Children (BPRS-C)	1	0	AE	AE	AE	1 (2)
Brief Screening Measure of Emotional Distress in children (BSMED-C)	AE	0	AE	AE	AE	0 (1)
Child and Adolescent Behaviour Assessment (CABA)	0	AE	AE	AE	AE	0 (1)
Child Adjustment and Parent Efficacy Scale (CAPES)	1	AE	AE	AE	AE	1 (1)
Children's Emotional Adjustment Scale (CEAS)	0	AE	AE	AE	AE	0 (1)
Child Symptom Inventory-4 (CSI-4)	0	AE	AE	1	AE	1 (2)
The Devereux Student Strengths Assessment (DESSA)	1	AE	AE	AE	AE	1 (1)
Devereux Scales of Mental Disorders (DSMD)	1	AE	AE	1	AE	2 (2)
Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA)	0	0	0	0	1	1(5)
The Nisonger Child Behaviour Rating Form: typical IQ version (NCBRF-TIQ)	0	AE	AE	AE	AE	0(1)
Primary Care Mental Health Screening Tool (PCMHS)	1	AE	AE	1	AE	2 (2)

Paediatric symptom checklist (PSC-17)	1	AE	0	0	AE	1 (3)
Symptoms and Functioning Severity Scale (SFSS)	AE	AE	AE	AE	AE	0 (0)
SFSS short forms A and B (SFSS-SF)	1	AE	AE	AE	AE	1 (1)
The Revised Child Anxiety and Depression Scale - Parent Version (RCADS-P)	1	AE	AE	1	AE	2 (2)
Revised Ontario Child Health Study scales (OCHS)	1	AE	AE	AE	AE	1 (1)
The Target Symptom Rating (TSR)	AE	AE	AE	0	AE	0 (1)
Youth Outcome Questionnaire (Y-OQ)	AE	AE	AE	AE	AE	0 (0)
SDQ	1	AE	AE	1	AE	2 (2)
ASEBA	1	AE	AE	1	AE	2 (2)

4. Discussion

This paper aimed to review the psychometric properties of informant rated outcome measures of children and young people's mental health to support clinicians and researchers in making informed decisions about their use of such measures in gathering practice based evidence and conducting research to inform clinical practice respectively. To meet this aim, a systematic review of the evidence base pertaining to the psychometric properties of informant rated outcome measures of children and young people's mental health was conducted and the results assessed using a novel quality appraisal tool. This review demonstrated that there is a significant number of carer-rated general mental health outcome measures designed for use with young people. The literature concerning the psychometric properties of these measures is sizeable, however its quality is varied.

Internal consistency was the most widely researched psychometric property for the measures included in this review, with only one measure (the TSR) having no literature reporting on this property. In comparison to the other psychometric properties assessed here, there appeared to be more agreement between authors concerning the approach to quantifying internal consistency, with the majority reporting Cronbach's alpha. Overall, the measures considered in this review had evidence of good internal consistency, with 15 of the 24 measures rated in this area meeting minimum criteria. Of the remaining nine measures, five (CABA; CEAS; CSI-4; NCBRF-TIQ and HoNOSCA) had evidence of poor internal consistency. Cronbach's alpha falling above the recommended cut-off of .90 was the most common cause of measures not meeting the criteria for good internal consistency. It is of note that this upper limit is contested, with some advocating a cut-off of .95 (e.g. Terwee *et al*., 2007). Articles pertaining to the remaining four measures failed to demonstrate adequate internal consistency due to

issues with methodology or reporting. The data pertaining to the internal consistency of the measures reviewed here was comparable to that reported by Stone *et al*. (2010) and Achenbach and Rescorla (2001) concerning the SDQ and ASEBA respectively.

There was a striking lack of literature pertaining to the inter-rater reliability of the measures included in this review, with data for only six of the 25 measures. Of these six measures, three (HoNOSCA, BPRS-C and BSMED-C) had evidence of poor inter-rater reliability, while the evidence for the remaining three measures was lacking. A failure to specify type and model of ICC used was commonplace. The quantity and quality, or lack thereof, of literature in this area is concerning. The picture with regards to test-retest reliability was similar, with no measure having evidence of adequate test-retest reliability. Not one article included in this review used the method of analysis specified in the criteria. Again, this finding is of concern when considering the use of mental health outcome measures in clinical practice. Test-retest reliability can be understood to be an important asset of an outcome measure when one considers its parallels with responsiveness. When attempting to capture change, it is important to know that scores on a measure remain stable when change has (theoretically) not occurred. This gives confidence to claims that when scores on such a measure do change, they do so because they are reflecting a change that has occurred in reality, as opposed to an error in measurement. Interestingly, the data concerning the test-retest reliability and inter-rater reliability of the SDQ as reported by Stone *et al*. (2010) was equally flawed according to the criteria employed here in regards to the use correlation analysis as opposed to ICC. Conversely, (aside from the lack of confidence intervals reported) there was evidence of good inter-rater and test-retest reliability for the CBCL, one component of the ASEBA (Achenbach & Rescorla, 2001).

Akin to internal consistency, construct validity was a highly researched area for the measures in this review. There was more consensus on methodology, however many articles failed to report hypotheses regarding nature of the target measure's relationship with the comparison measure(s). Drawing post-hoc conclusions from a large number of correlations between measures is somewhat different to generating hypotheses about the nature of such relationships based on theory, which data is then used to prove or disprove. The latter is clearly more closely aligned with Kirschner and Guyatt's (1985) definition of construct validity. Methodological limitations aside, three measures (HoNOSCA, PSC-17 and TSR) had associated literature suggestive of poor construct validity. Evidence of adequate construct validity was available for the SDQ and ASEBA, however similarly to the measures reviewed here, it was unclear whether the data analyses conducted were informed by theoretically derived hypotheses.

Finally, and somewhat surprisingly, the literature pertaining to the responsiveness of the outcome measures included in this review was small. Only three measures had published literature concerning their responsiveness, with one measure, the HoNOSCA, having evidence of good responsiveness. The articles related to the remaining two measures (BPRS-C and TSR) did not use methods of analysis in line with criteria used in this review. Findings here were in line with the evidence base reviewed by Stone *et al*. (2010) and Achenbach and Rescorla (2001) regarding the SDQ and ASEBA respectively, where no data concerning responsiveness was reported.

Aside from the psychometric properties of the measures reviewed, another point of interest was the frequency of the developers of the measures being involved in the measurement of their psychometric properties. This clearly represents a conflict of interest and provides a rationale for the potential withholding of evidence that represents a measure in a negative light, particularly in light of the possible financial

gain associated with the creation of such a measure. It is possible that this may account for some of the absence of evidence concerning the psychometric properties of these measures.

Overall, in terms of informant rated mental health outcome measures for young people, what was striking was an absence of evidence regarding their psychometric properties rather than evidence of poor psychometric properties. The exception to this was the HoNOSCA, which had evidence of poor quality across four of the five domains assessed. Whilst there is not an insignificant quantity of literature pertaining to the psychometric properties of such measures, this review found its quality is somewhat lacking. These findings mirrored those of Deighton *et al.* (2014) in their review of self-rated mental health outcome measure, which concluded that none of the reviewed measures had sufficient psychometric rigour to suggest that they were able to reliably measure both symptom severity and responsiveness. This review also identified widespread use of outdated methods of analysing measures' psychometric properties, particularly in the case of less recently developed measures.

Strengths and Limitations

This review was the first of its kind aiming to assess the psychometric properties of informant rated mental health outcome measures for children and young people. Furthermore, it is the first review to employ a novel quality appraisal tool with justified empirical standards to evaluate the psychometric properties of such measures. It is important however to note that this review is not without limitations. First, with regards to the search strategy employed, the restriction of search results to those published in English may have restricted the breadth of measures included in this

review and limits its relevance to English speaking countries. Second, as recognised by Humphrey *et al*. (2011), the publication bias inherent in systematic reviews of this nature may have negatively affected the inclusion of some measures used in clinical practice without published research. Of importance is the failure of this review to include studies published in the test manuals of specific measures in place of empirical journals as a result of the cost implications of accessing such data. This likely restricted the breadth of evidence available to this researcher when assessing the merit of each measure. Conversely, the exclusion of test-manuals likely reflects the conditions under which many clinicians are forced to make decisions about the use of outcome measures. With regards to the quality criteria employed in this review, whilst it is felt by the author to reflect current thinking on test psychometrics, this thinking is changeable and contentious. This author also notes a limitation with regards to the evaluation of construct validity and acknowledges that this review may have neglected to take into account evidence where a measure not meeting the inclusion criteria for this review used one of the 25 included measures as a comparison in the exploration of its own construct validity. This review further failed to take into account the feasibility of use of the measures included here. Finally, this author notes that this review failed to address the suitability of the reviewed measures for use across different cultures. Mushquash and Bova (2007) warn of the dangers of using outcome measures with cultural groups for where the psychometric properties of these measures have not been adequately assessed within the cultural group in question. The relevance of the findings of this review are therefore limited to the largely western European and American cultures where the included measures were developed and researched.

Implications

This review has implications for both clinical practice and future research. With regards to the first, the findings of this review demonstrate the importance of clinicians' awareness of the outcome measures that they and the services they work in use to inform their practice. It is important for clinicians to consider what they hope to achieve through the use of an outcome measure and to select a measure with evidence of corresponding strengths. On both an individual and an aggregate level, the users of outcome measures should ensure that the decisions and assertions they make based on data gathered through such measures are valid, based on the available evidence. The findings of this review indicate that caution should be taken when using the measures with an absence of evidence pertaining to their psychometric properties. If using such measures, it would be important to acknowledge and report the limitations of the measure used. This would be of particular importance when using such data to facilitate the collection of practice based evidence to inform service level decision making. This review would caution the use of BPRS-C, BSMED, CABA, CEAS, CSI-4, HoNOSCA, NCBRF-TIQ, PSC-17 and TSR, for which there is evidence of poor reliability and/or validity.

With regards to research, when selecting an outcome measure to operationalise a dependant variable, the findings of this literature review highlight the need to be aware of the evidence pertaining to the psychometric properties of the measure being used and the implications of this for the conclusions that can be drawn. This review highlights the need for the psychometric properties of routinely used informant rated outcome measures for young people to continue to be tested in line with the most up to date literature on test statistics. In particular, responsiveness and inter-rater reliability should be prioritised for further research given their importance to the

measurement of change over time. In addition, it would be beneficial for the inter-rater reliability of measures to be further explored, particularly for those measures which are likely to be completed by multiple informants.

References (Reviewed articles are marked with an asterisk *)

Achenbach, T. M., McConaughy, S. H. & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin*, *101*(2), 213.

Achenbach, T.M. & Rescorla, L.A. (2001). *Manual for the ASEBA School-Age Forms and Profiles*. Burlington, VT: University of Vermont Research Center for Children, Youth, & Families.

Aman, M., Leone, S., Lecavalier, L., Park, L., Buican, B. & Coury, D. (2008). The Nisonger child behavior rating form: Typical IQ version. *International clinical psychopharmacology*, *23*(4), 232-242. *

Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of physical medicine and rehabilitation*, *81*, 15-20.

Athay, M.M., Riemer, M. & Bickman, L. (2012). The Symptoms and Functioning Severity Scale (SFSS): Psychometric evaluation and discrepancies among youth, caregiver, and clinician ratings over time. *Administration and Policy in Mental Health and Mental Health Services Research*, *39*(1-2), 13-29. *

Ballesteros-Urpí, A., Pardo-Hernández, H., Ferrero-Gregori, A., Torralbas-Ortega, J., Puntí-Vidal, J., *et al*. (2018). Validation of the Spanish and Catalan versions of the Health of the Nation Outcome Scale for Children and Adolescents (HoNOSCA). *Psychiatry research*, *261*, 554-559. *

Barber, C.C., Neese, D.T., Coyne,L., Fultz,J. & Fonagy, P. (2002). The target symptom rating: A brief clinical measure of acute psychiatric symptoms in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *31*(2), 181-192. *

Bates, M.P. (2001). The Child and Adolescent Functional Assessment Scale (CAFAS): Review and Current Status. *Clinical Child and Family Psychology Review, 4*(1), 63-84.

Beaton, D. E., Hogg-Johnson, S. & Bombardier, C. (1997). Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *Journal of clinical epidemiology*, *50*(1), 79-93.

Beutler, L.E. (2009). Making science matter in clinical practice: Redefining psychotherapy. *Clinical Psychology: Science and Practice, 16*, 301-317.

Bickman, L., Kelley, S. D., Breda, C., de Andrade, A. R. & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services, 62*(12), 1423-1429.

Boyle, M.H., Offord, D.R., Racine, Y., Fleming, J.E., Szatmari, P. & Sanford, M. (1993). Evaluation of the revised Ontario child health study scales. *Journal of Child Psychology and Psychiatry*, *34*(2), 189-213. *

Brann, P. & Coleman, G. (2010). On the meaning of change in a clinician's routine measure of outcome: HoNOSCA. *Australian and New Zealand Journal of Psychiatry*, *44*(12), 1097-1104. *

Brann, P., Coleman, G. & Luk, E. (2001). Routine outcome measurement in a child and adolescent mental health service: an evaluation of HoNOSCA. *Australian and New Zealand Journal of Psychiatry*, *35*(3), 370-376. *

Buckley, J.A., Ryser, G., Reid, R. & Epstein, M.H. (2006). Confirmatory factor analysis of the behavioral and emotional rating scale–2 (BERS-2) parent and youth rating scales. *Journal of Child and Family Studies*, *15*(1), 27-37. *

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences, 2nd ed*. Hillsdale, NJ: Erlbaum.

Curry, J.F. & Ilardi, S.S. (2000). Validity of the Devereux Scales of Mental Disorders with adolescent psychiatric inpatients. *Journal of clinical child psychology*, *29*(4), 578-588. *

Das, J.K., Salam, R.A., Lassi, Z.S., Khan, M.N., Mahmood, W., Patel, V., *et al*. (2016). Interventions for adolescent mental health: An overview of systemic reviews. *The Journal of Adolescent Health, 59*(4), 49-60.

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P. & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health, 8*(14).

Denton, R. N. (2016). *The assessment of mental health in looked-after adolescents: an exploratory study of the brief assessment checklist for adolescents*. Unpublished doctoral thesis, University of Surrey. *

Department of Health (2007). *Every Child Matters*. London

Department of Health (2012). *Liberating the NHS: No decision about me without me. Government response*. London: Stationary Office

Dunn, T. W., Burlingame, G. M., Walbridge, M., Smith, J. & Crum, M. J. (2005). Outcome assessment for children and adolescents: psychometric validation of the Youth Outcome Questionnaire 30.1 (Y-OQ®-30.1). *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, *12*(5), 388-401.*

Duppong Hurley, K., Lambert, M.C. & Stevens, A. (2015). Psychometrics of the Symptoms and Functioning Severity Scale for high-risk youth. *Journal of emotional and behavioral disorders*, *23*(4), 206-214. *

Ebesutani, C., Chorpita, B. F., Higa-McMillan, C. K., Nakamura, B. J., Regan, J. & Lynch, R.E. (2011). A psychometric analysis of the Revised Child Anxiety and Depression Scales—Parent version in a school sample. *Journal of abnormal child psychology*, *39*(2), 173-185. *

Ebesutani, C., Chorpita, B.F., Higa-McMillan, C.K., Nakamura, B.J., Regan, J. & Lynch, R.E. (2011). A psychometric analysis of the Revised Child Anxiety and Depression Scales—Parent version in a school sample. *Journal of abnormal child psychology*, *39*(2), 173-185. *

Ebesutani, C., Tottenham, N., & Chorpita, B. (2015). The Revised Child Anxiety and Depression Scale-Parent version: Extended applicability and validity for use with younger youth and children with histories of early-life caregiver neglect. *Journal of Psychopathology and Behavioral Assessment*, *37*(4), 705-718. *

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R. & Conover, N.C. (1985). Age differences in the reliability of the psychiatric interview of the child. *Child development*, 265-275.

Frogley, C. L. (2016). *Assessing the mental health needs of looked after children: a study investigating the utility of the brief assessment checklist for children*. Unpublished doctoral thesis, University of Surrey. *

Gabrielli, J., Jackson, Y. & Brown, S. (2015). Measurement of behavioral and emotional outcomes of youth in foster care: Investigation of the roles of age and placement type. *Journal of psychopathology and behavioral assessment*, *37*(3), 422-431. *

Gale, J., Pfefferbaum, B., Suhr, M.A. & Overall, J.E. (1986). The brief psychiatric rating scale for children: A reliability study. *Journal of Clinical Child Psychology*, *15*(4), 341-345. *

Gardner, W., Lucas, A., Kolko, D. J. & Campo, J. V. (2007). Comparison of the PSC-17 and alternative mental health screens in an at-risk primary care sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(5), 611-618. *

Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., *et al*. (1999). The PSC-17: A brief pediatric symptom checklist with psychosocial problem subscales. A report from PROS and ASPN. *Ambulatory Child Health*, *5*(3), 225-236. *

Garralda, M.E., Yates, P. & Higginson, I. (2000). Child and adolescent mental health service use: HoNOSCA as an outcome measure. *The British Journal of Psychiatry*, *177*(1), 52-58.*

Gimpel, G.A. & Nagle, R. J. (1999). Psychometric properties of the Devereux scales of mental disorders. *Journal of Psychoeducational Assessment*, *17*(2), 127-144.*

Goemans, A., Tarren-Sweeney, M., van Geel, M. & Vedder, P. (2018). Psychosocial screening and monitoring for children in foster care: Psychometric properties of the Brief Assessment Checklist in a Dutch population study. *Clinical child psychology and psychiatry*, *23*(1), 9-24.*

Gonzalez, J.E., Ryser, G.R., Epstein, M.H. & Shwery, C.S. (2006). The behavioral and emotional rating scale: Parent rating scale (BERS-II PRS): A Hispanic cross-cultural reliability study. *Assessment for Effective Intervention*, *31*(3), 33-43.*

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *The Journal of Child Psychology and Psychiatry, 38*(5), 581-586.

Gowers, S.G., Harrington, R.C., Whitton, A., Lelliott, P., Beevor, A., Wing, J., *et al*. (1999). Brief scale for measuring the outcomes of emotional and behavioural disorders in children: Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA). *The British Journal of Psychiatry*, *174*(5), 413-416.*

Green, D. & Latchford, G. (2012). *Maximising the benefits of psychotherapy: A practice-based evidence approach*. John Wiley & Sons

Grietens, H.; Onghena, P.; Prinzie, P., Gadeyne, E., Van Assche, V., Ghesquiere, P., et al. (2004). Comparison of mothers', fathers', and teachers' reports on problem behavior in 5- to 6-year-old children. *Journal of Psychopathology and Behavioral Assessment, 26*, 137–146.

Gross, T.J., Hurley, K.D., Lambert, M.C., Epstein, M.H. & Stevens, A.L. (2015). Psychometric evaluation of the Symptoms and Functioning Severity Scale (SFSS) short forms with out-of-home care youth. *Child & youth care forum*, *44*(2), 239-249. *

Hall, C. L., Moldavsky, M., Taylor, J., Sayal, K., Marriott, M., Batty, M. J., *et al*. (2014). Implementation of routine outcome measurement in child and adolescent mental health services in the United Kingdom: A critical perspective. *European Child and Adolescent Psychiatry, 23*, 239–242.

Happell, B. (2008). Determining the effectiveness of mental health services from a consumer perspective: Part 2: Barriers to recovery and principles for evaluation. *International Journal of Mental Health Nursing, 17*(2), 123-130.

Harmon, S.C., Lambert, M.J., Smart, D.M., Hawkins, E., Nielsen, S.L., Slade, K., *et al*. (2007). Enhancing outcome for potential treatment failures: Therapist–client feedback and clinical support tools. *Psychotherapy research, 17*(4), 379-392

Harnett, P.H., Loxton, N.J., Sadler, T., Hides, L. & Baldwin, A. (2005). The Health of the Nation Outcome Scales for Children and Adolescents in an adolescent in-patient sample. *Australian and New Zealand Journal of Psychiatry*, *39*(3), 129-135. *

Hartung, C.M. & Lefler, E.K. (2010). Preliminary examination of a new mental health screener in a pediatric sample. *Journal of Pediatric Health Care*, *24*(3), 168-175.*

Hatfield, D., McCullough, L., Frantz, S. H. & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, *17*(1), 25-32.

Hodges, K. (1989). Child and Adolescent Functional Assessment Scale. Unpublished manuscript, Eastern Michigan University, Ypsilanti.

Humphrey, N., Kalambouka, A., Wigelsworth, M., Lendrum, A., Deighton, J., Wolpert, M. (2011). Measures of social and emotional skills for children and young people: a systematic review. *Educ Psychol Meas, 71*, 617–637.

Hunsley, J. & Mash, E.J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*, 29-51.

Hunter, J., Higginson, I. & Garralda, E. (1996). Systematic literature review: outcome measures for child and adolescent mental health services. *J Public Health Med,18 ,* 197–206.

Husted, J.A., Cook, R.J., Farewell, V.T. & Gladman, D.D. (2000). Methods for assessing responsiveness: a critical review and recommendations. *Journal of clinical epidemiology*, *53*(5), 459-468.

Jacobson, J.H., Pullmann, M.D., Parker, E.M. & Kerns, S.E. (2018). Measurement Based Care in Child Welfare-Involved Children and Youth: Reliability and Validity of the PSC-17. *Child Psychiatry & Human Development*, 1-14. *

Kirshner, B. & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of chronic diseases*, *38*(1), 27-36.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Kwan, B. & Rickwood, D.J. (2015). A systematic review of mental health outcome measures for young people aged 12 to 25 years. BMC Psychiatry, *15*, 1-19.

Lachar, D., Randle, S.L., Harper, R.A., Scott-Gurnell, K.C., Lewis, K.R., Santos, C.W, *et al*. (2001). The brief psychiatric rating scale for children (BPRS-C): Validity and reliability of an anchored version. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(3), 333-340.*

Lambert, M.C., Hurley, K.D., Gross, T.J., Epstein, M.H. & Stevens, A.L. (2015). Validation of the Symptoms and Functioning Severity Scale in residential group care. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(3), 356-362.*

Lambert, M.C., January, S.A.A., Epstein, M.H., Spooner, M., Gebreselassie, T. & Stephens, R.L. (2015). Convergent validity of the Behavioral and Emotional Rating Scale for youth in community mental health settings. *Journal of Child and Family Studies*, *24*(12), 3827-3832.*

Lambert, M. J., Whipple, J.L., Hawkins, E.J., Vermeersch, D.A., Nielsen, S.L. & Smart, D.W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice, 10*(3), 288-301.

LeBuffe, P. A., Shapiro, V. B. & Robitaille, J. L. (2018). The Devereux Student Strengths Assessment (DESSA) comprehensive system: Screening, assessing, planning, and monitoring. *Journal of Applied Developmental Psychology*, *55*, 62-70. *

Lefler, E.K., Hartung, C.M. & Fedele, D.A. (2012). Psychometric properties of a primary care mental health screening tool for young children. *Children's Health Care*, *41*(2), 79-96.*

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S, *et al.* (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice, 6*(1), 95-106.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., *et al.* (2002). The dodo bird verdict is alive and well—mostly. *Clinical Psychology: Science and Practice, 9*(1), 2-12.

Marsh, H. W., Debus, R. & Bornholt, L. (2005).Validating young children's self-concept responses: Methodological ways and means to understand their responses. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 138-160). Oxford: Blackwell.

McClendon, D. T., Warren, J. S., M. Green, K., Burlingame, G. M., Eggett, D. L. & McClendon, R. J. (2011). Sensitivity to change of youth treatment outcome measures: a comparison of the CBCL, BASC-2, and Y-OQ. *Journal of Clinical Psychology*, *67*(1), 111-125.*

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30.

McIlhaney, K.K., Henderson, L., Gunn, S. & Wasser, T.E. (2008). Use of the Brief Psychiatric Rating Scale-Children (BPRS-C) as an Outcomes Measure in the Residential and Foster Care Setting. *Residential Treatment For Children & Youth*, *24*(3), 243-259. *

Mejia, A., Filus, A., Calam, R., Morawska, A. & Sanders, M.R. (2016). Validation of the Spanish version of the CAPES: A brief instrument for assessing child psychological difficulties and parental self-efficacy. *International Journal of Behavioral Development*, *40*(4), 359-372. *

Molland, R.S., Diep, L.M., Brox, J.I., Stuge, B., Holm, I. & Kibsgard, T.J. (2018). Reliability and Construct Validity of the Adapted Norwegian Version of the Early-Onset Scoliosis 24-item Questionnaire. *JAAOS Global Research & Reviews*, *2*(7).

Mooney, P., Epstein, M.H., Ryser, G. & Pierce, C.D. (2005). Reliability and validity of the behavioral and emotional rating scale: Parent rating scale. *Children & Schools*, *27*(3), 147-155. *

Morawska, A., Sanders, M. R., Haslam, D., Filus, A. & Fletcher, R. (2014). Child adjustment and parent efficacy scale: Development and initial validation of a parent report measure. *Australian Psychologist*, *49*(4), 241-252. *

Morin, A.L., Miller, S.J., Smith, J.R. & Johnson, K.E. (2017). Reliability and Validity of the Child and Adolescent Behavior Assessment (CABA): A Brief Structured Scale. *Child Psychiatry & Human Development*, *48*(2), 200-213. *

Mullins, D., Pfefferbaum, B., Schultz, H. & Overall, J.E. The brief psychiatric rating scale for children: Quantitative scoring of medical records. *Psychiatry Res, 19*(1), 43-49.*

Murphy, J.M., Bergmann, P., Chiang, C., Sturner, R., Howard, B., Abel, M.R., *et al*. (2016). The PSC-17: subscale scores, reliability, and factor structure in a new national sample. *Pediatrics*, *138*(3). *

Mushquash, C.J. & Bova, D.L. (2007). Cross-cultural assessment and measurement issues. *Journal on Developmental Disabilities 13*(1), 53-66.

Nickerson, A. B. & Fishman, C. (2009). Convergent and divergent validity of the Devereux Student Strengths Assessment. *School Psychology Quarterly*, *24*(1), 48. *

Nunnally, J.C. and Bernstein, I.H. (1994) The Assessment of Reliability. *Psychometric Theory*, *3*, 248-292.

Paiva, C.E., Barrosco, E.M., Carneseca, E.C., Padua Souza, C., Santos, F.T., Lopez, R.V.M., *et al*. A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review. *BMC Medical Research Methodology, 14*(8).

Park, A.L., Ebesutani, C.K., Bose, D. & Chorpita, B.F. (2016). Psychometric properties of a Spanish translation of the revised child anxiety and depression scale–parent version. *Journal of Psychopathology and Behavioral Assessment*, *38*(2), 307-319. *

Parker, G., Yiming, C., Rutter, M. & Tan, S. (2001). The development of a brief screening measure of emotional distress in children. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 221-225.*

Piper, B.J., Gray, H.M., Raber, J. & Birkett, M.A. (2014). Reliability and validity of Brief Problem Monitor, an abbreviated form of the Child Behavior Checklist. *Psychiatry and clinical neurosciences*, *68*(10), 759-767.*

Reddy, L.A., Pfeiffer, S.I., & Files-Hall, T.M. (2007). Use of the devereux scales of mental disorders for children and adolescents with emotional disturbance. *Journal of Psychoeducational Assessment*, *25*(4), 356-372. *

Richter, J. (2015). Preliminary evidence for good psychometric properties of the Norwegian version of the Brief Problems Monitor (BPM). *Nordic journal of psychiatry*, *69*(3), 174-178.*

Shafer, A.B. (2013). Factor structure of the Brief Psychiatric Rating Scale for Children (BPRS-C) among hospital patients and community clients. *Personality and Individual Differences*, *55*(1), 41-46. *

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Siddique, H. (2018, 22 November). Mental health disorders on rise among children. *The Guardian*.

Smith, S.R. & Reddy, L.A. (2002). The concurrent validity of the Devereux Scales of Mental Disorders. *Journal of Psychoeducational Assessment*, *20*(2), 112-127. *

Sointu, E.T., Geležinienė, R., Lambert, M. C. & Nordness, P. D. (2015). Internal consistency and cross-informant agreement of the Lithuanian-Translated Behavioral and Emotional Rating Scale. *International Journal of School & Educational Psychology*, *3*(2), 135-141. *

Sprafkin, J., Gadow, K.D., Salisbury, H., Schneider, J. & Loney, J. (2002). Further evidence of reliability and validity of the Child Symptom Inventory-4: Parent checklist in clinically referred boys. *Journal of Clinical Child and Adolescent Psychology*, *31*(4), 513-524. *

Stone, L.L., Otten, R., Rutger, C.M.E., Vermulst, A.A. & Janssens, J.M.A.M. (2010). Psychometric Properties of the Parent and Teacher Versions of the Strengths and Difficulties Questionnaire for 4- to 12-Year-Olds: A Review. *Clinical Child and Family Psychology Review, 13*(3), 254-274.

Streiner, D.L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, *80*(1), 99-103.

Streiner, D.L. & Norman, G.R.(2003) *Health measurement scales: A practical guide to their development and use*. Oxford University Press: Oxford

Tarren-Sweeney, M. (2007). The Assessment Checklist for Children—ACC: A behavioral rating scale for children in foster, kinship and residential care. *Children and Youth Services Review*, *29*(5), 672-691.*

Tarren-Sweeney, M. (2013). The Assessment Checklist for Adolescents—ACA: A scale for measuring the mental health of young people in foster, kinship, residential and adoptive care. *Children and Youth Services Review*, *35*(3), 384-393.*

Tarren-Sweeney, M. (2013). The Brief Assessment Checklists (BAC-C, BAC-A): Mental health screening measures for school-aged children and adolescents in foster, kinship, residential and adoptive care. *Children and Youth Services Review*, *35*(5), 771-779.*

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., *et al*. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*, *60*(1), 34-42.

The Children's Society (2018). *The good child report 2018*. London.

Thorlacius, O. & Gudmundsson, E. (2014). Assessment of children's emotional adjustment: Construction and validation of a new instrument. *Child: care, health and development, 41*(5), 762-771.*

Tiffin, P.A. & Rolling, K. (2012). Structure of the Health of the Nation Outcome Scales for Children and Adolescents: an ordinal factor analysis of clinician ratings of a sample of young people referred to community mental health services. *Psychiatry research*, *197*(1-2), 154-162.*

Vaz, S., Cordier, R., Boyes, M., Parsons, R., Joosten, A., Ciccarelli, M., *et al*. (2016). Is using the strengths and difficulties questionnaire in a community sample the optimal way to assess mental health functioning? *PLoS One, 11*(1).

Von Wyl, A., Toggweiler, S., & Zollinger, R. (2017). HoNOSCA-D as a Measure of the severity of Diagnosed Mental Disorders in children and adolescents—Psychometric Properties of the German Translation. *Frontiers in psychiatry*, *8*, 186.*

Wampold, B.E. (2001). *The great psychotherapy debate: Models, methods and findings* (2nd edn*).* Lawrence Erlbaum Associates: Mahwah

Wolpert, M., Aitken, J., Syrad, H., Munroe, M., Saddington, C., Trustam, E., *et al*. (2009). *Review and recommendations for national policy for England for the use of mental health outcome measures with children and young people. DCSF Research Report 2008/56*. Nottingham: DCSF Publications.

Yates, P., Garralda, M.E. & Higginson, I. (1999). Paddington Complexity Scale and Health of the Nation Outcome Scales for children and adolescents. *The British Journal of Psychiatry*, *174*(5), 417-423. *

Yuan, J.M. (2015). HoNOSCA in an adolescent psychiatric inpatient unit: An exploration of outcome measures. *Psychiatria Danubina*, *27*(1), 357-363. *

Part two

The Structure and Psychometric Properties of the BERRI, an Outcome Measure for Looked After Children in Residential Care.

Abstract

Background: Looked After Children are (LAC) considered one of the most vulnerable groups in society due to difficult life experiences and subsequent poor outcomes within social, physical health and mental health domains. The use of mental health outcome measures has been demonstrated to have a positive impact on clinical outcomes. The use of such measures carries additional importance for LAC with regards to placement planning and the identification of need for therapeutic intervention. The BERRI is an outcome measure designed for use with LAC, developed out of concern regarding the lack of measures tailored to the needs of this population. However, the psychometric properties of the BERRI had not been explored.

Objective: This study aimed to: (i) Explore the psychometric properties of the BERRI in its current form for use with LAC in residential care; namely its inter-rater reliability (IRR), construct validity and internal consistency; (ii) explore whether these psychometric properties might be enhanced through the empirical extraction of factors.

Method: Data were collected from an online database where several residential children's homes routinely collect BERRI data for children in their care. A subgroup of residential children's homes were also asked to complete additional BERRIs and the Strengths and Difficulties Questionnaire for the purposes of exploring IRR and construct validity.

Results: Calculation of Cronbach's alpha indicated good internal consistency for all original scales. Exploration of the BERRI's IRR through the calculation of intra-class correlation coefficients demonstrated poor to moderate IRR for the 'Emotions', 'Relationships' and 'Indicators' scales and moderate to good IRR for the 'Behaviour' and 'Risk' scales, in addition to the BERRI total score. Evidence of good construct validity was found. An exploration of the structure of the BERRI using principal components analysis revealed a five component structure similar to the original scales. The psychometric properties of the BERRI were not improved through the empirical extraction of factors.

Discussion: Suggestions are made with regards to the item content of the BERRI. Consideration is also given to the clinical implications arising from the exploration of the measure's IRR and subsequently how IRR might be improved. Areas for future research were identified, including an exploration of the BERRIs test-retest reliability and responsiveness to change. Overall, the BERRI was felt to show promise as a targeted outcome measure for use with LAC in residential care.

1. Introduction

1.1 The needs of Looked After Children

The Children Act 1989 defines a young person as being 'looked after' when they are placed in the care of a local authority through the granting of a care order, or are accommodated by a local authority for a continuous period exceeding 24 hours. It is estimated that two thirds of children placed in care have experienced abuse or neglect (DfE, 2017). Given these early life experiences, it is of little surprise that Looked After Children (LAC) are considered among the most vulnerable groups in society (e.g. Iwaniec, 2006). Indeed, LAC have been found to have worse outcomes than their peers in the general population in a variety of domains, including physical health (Rodrigues, 2004), offending behaviour (DfE, 2016), homelessness (Broad, 1998) and educational attainment (DfE, 2016). Unsurprisingly, research indicates that this pattern extends to the mental health of LAC, with Fisher (2015) identifying a wide range of adverse life events often experienced by these children, prior to and following their entry into care, and the negative impact of these on their neurobiological and psychological development. In line with this, Sempik *et al*. (2008) found that in their sample of 648 LAC, 72% had indications of behavioural and emotional difficulties upon entry to care. This pattern was found by Teggart and Menary (2005) to continue post entry to care, with over 60% of four to 10 year-olds and two thirds of 11 to 16 year olds in their LAC sample meeting diagnostic criteria for a 'mental health difficulty'.

1.2 Mental health outcome measures for LAC

The benefits associated with the use of mental health outcome measures among the general population are widely reputed. These benefits include assessment of change over time (e.g. Hatfield *et al*., 2010) and increases in the speed and magnitude of positive change (Bickman *et al*., 2011; Harmon *et al*., 2007). The use of mental health outcome measurement for LAC carries additional importance. In 2017, the DfE reported that of the 72,670 children cared for by their local authority in England, approximately one third (31.7%) had experienced more than one placement breakdown during the preceding year. Of this proportion, just under half (48.5%) had experienced more than three placement breakdowns. Rubin *et al.* (2007) found (after controlling for existing difficulties upon entry to care) that children who experienced multiple placement breakdowns had poorer emotional and behavioural wellbeing than those with stable placements. This study is a clear example of the compounding impact of placement breakdown on the mental health of this already vulnerable population. Research into placement breakdown suggests that young people's difficulties with emotional and behavioural regulation as well as carers' perceptions of these difficulties and the level of challenge they present, are strongly associated with placement instability (Farmer *et al*., 2005). Carers' perceptions of risk and threat to their family, alongside child-carer relationship difficulties, have also been found to be associated with placement breakdown (Rock *et al*., 2013). Given the aforementioned research indicating that emotional and behavioural wellbeing contribute to and suffer as a result of placement breakdown, it is of little surprise that a recent report by the House of Commons (2016) stressed the important role played by mental health outcome measurement in appropriate placement planning and the identification of need for therapeutic intervention among the LAC population.

Despite the importance of mental health outcome measurement in the care of LAC, there is a paucity of measures designed to meet the unique needs of this population. In the UK, the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) is used by local authorities to identify the emotional and behavioural needs of LAC on the recommendation of the Department for Children, Schools and Families (2009). It is mandatory for the SDQ to be completed by the primary carers of children who have been 'looked after' for over a year and who are between the ages of four and 16. It is then required that local authorities report this data to the Department for Education on an annual basis (Goodman & Goodman, 2012). The SDQ assesses five domains: Conduct problems; emotional symptoms; peer relationship difficulties; prosocial behaviour; and hyperactivity/inattention. There is evidence that the SDQ is able to identify LAC meeting criteria for DSM-IV diagnoses (Goodman *et al*., 2004), screen for social and emotional difficulties among LAC in the USA (Jee *et al*., 2011) and discriminate between 'LAC' and 'non-LAC' (Marquis & Flynn, 2009). Despite this, the threats experienced by LAC and the range of behaviours they display to attempt to get their needs met are often either not covered by standardised diagnostic classification systems, or are made sense of in terms of 'symptoms' devoid of contextualised understanding of what a given presentation might represent for a child. There is often a complex relationship between presenting difficulties and a child's attempts to meet attachment needs or manage difficulties associated with developmental trauma (Iwaniec, 2006). As such, concerns have been raised that existing standardised measures, such as the SDQ, may not fully capture the complex needs of LAC (Achenbach, Dumenci & Rescorla, 2003; Tarren-Sweeny, 2007). Silver *et al*. (2016) identified that existing standardised assessment measures fail to take into account the frequency and complexity of life events that commonly occur for LAC after entering care (e.g. placement breakdown, school moves, entry of a new child into a

placement) and the impact of these on assessment scores when considering change over time.

In a recent systematic literature review of measures suitable for the assessment of children and adolescents with experiences of developmental trauma, Denton *et al*. (2016) identified The Assessment Checklist tools created by Tarren-Sweeny (2007, 2013a, 2013b) as the only measure suitable to capture the needs of this group. The Assessment Checklist tools include the Assessment Checklist for Children (ACC) and the Assessment Checklist for Adolescents (ACA), alongside shorter versions of these assessments named the Brief Assessment Checklist for Children (BAC-C) and the Brief Assessment Checklist for Adolescents (BAC-A). The ACC, ACA, BAC-C and BAC-A have been shown to have good internal consistency and construct validity, however it has been acknowledged that the factor structure of the brief checklists require further attention and the external reliability and responsiveness to change of each of the assessment checklist tools requires exploration (Denton, 2016; Frogley, 2016; Goemans *et al*., 2018; Tarren-Sweeny 2007, 2013a). The items included in the Assessment Checklist Tools were developed through consultation with professionals and carers supporting LAC, and a review of existing research and clinical assessment reports . As such, the content of these measurement tools more accurately reflects the complex needs of LAC. Unfortunately, the Assessment Checklist Tools, like the SDQ, fail to take into account contextual factors and carers' perceptions of the level of difficulty and challenge presented by the each of the rated constructs.

1.3 The development of the BERRI

In response to a series of consultations held by CPLAC (Clinical Psychologists working with Looked After Children), concerning the deficits in available outcome

measures for LAC, Silver (unpublished) developed the BERRI, a system of outcome measurement specifically designed for use with LAC. In developing the BERRI, Silver conducted a series of focus groups with foster carers, residential care staff, managers of children's homes and social workers in 2007. These groups discussed the primary difficulties experienced by LAC and the ways in which these were communicated to carers. The difficulties raised in the focus groups were consolidated to form an 87 item questionnaire. Items were clustered on face-value into five scales, before feedback was sought from 324 Clinical Psychologist with an interest in working with LAC and adopted children. The measure was designed for completion by care-staff, foster carers or mental health professionals, who are asked to rate both the frequency and perceived difficulty of each item. The information gathered is entered into an online system which allows for a young person's outcomes to be tracked as subsequent BERRIs are added. The BERRI includes a life-events scale whereby raters can indicate whether any of a series of commonly experienced adverse life events have occurred for a given young person, such that any change in score can be interpreted in the context of life events which have occurred over the rated period. From 2009 to 2014, the BERRI was used by clinicians working with children in and on the edge of care as part of a pilot phase; feedback was also collected regarding the utility of the BERRI. Following the pilot phase, the BERRI began to be used commercially and was introduced to 20 children's homes across five local authorities. The use of the BERRI in these settings was widely praised as an example of good practice and the residential care provider involved in the pilot study elected to introduce the BERRI to all of their homes nationally. From 2016, the BERRI has been introduced more widely into other residential care settings operated by different providers. It is currently being used in 165 children's homes, three secure children's homes and a small fostering agency.

To date, the psychometric properties of the BERRI have not been explored. Whilst the addition of a tool tailored to measuring the needs of LAC over time is welcome, it is important to ascertain that any outcome measure is valid and reliable before conclusions can be drawn from its use (Happell, 2008). As the BERRI is designed for completion by informants, it is likely that different informants will rate the same child over time. For this reason, it is important to assess whether the BERRI has good inter-rater reliability (IRR) in order that the extent to which differences between individuals, or the same individual rated by different people at different time points, are a result of differences related to the individual(s), as opposed to those rating them. It is also necessary to establish the BERRI's construct validity, or the extent to which it is measuring the constructs it is intending to. The structure of the BERRI was determined through the face-value clustering of items into scales, as such it is important to ascertain the internal consistency of the subscales, to ensure that the constituent items within each scale are each contributing to the assessment of a particular construct. It is further important to explore if the five subscales are empirically supported through a factor analysis of BERRI data.

1.4 Aims and objectives

This study aimed to: (i) Explore the psychometric properties of the BERRI as it has been utilised in services to date (namely its internal consistency, IRR and construct validity) (ii) Explore the structure of the BERRI through factor analysis; and (iii) to explore whether the psychometric properties of the measure would be improved if the factors extracted through factor analysis were utilised over the original subscales.

2. Method

2.1 Design

The study was cross-sectional in design. Naturalistic data were collected from the online BERRI database, with additional data generated for the purposes of exploring inter-rater reliability and construct validity. Internal consistency was determined through the calculation of Cronbach's alpha for each scale on the BERRI. In order to establish IRR, comparisons between sets of raters rating the same young people were made by calculating intraclass correlations (ICC), using an absolute agreement definition. ICC was calculated for each BERRI scale and the total BERRI score. Please see Appendix I for further discussion regarding the use of ICC to measure IRR and the type of ICC selected here. Construct validity was ascertained using correlation analysis to explore the degree to which scales on the BERRI converged with and diverged from scores on related measures in a manner that would be theoretically expected, known as convergent and divergent validity. An exploration of the BERRIs structure was undertaken using Principal Component Analysis (PCA). The internal consistency, IRR and construct validity of the BERRI's scales as indicated by PCA were then analysed. All statistical analyses were performed using SPSS v. 25.0 (IBM Corp., Armonk, NY, USA).

2.2 Measures

2.2.1 The BERRI

Each item on the BERRI (Appendix J) is given two scores: One pertaining to how frequently the item occurs and another concerning the extent to which it is

perceived as challenging by the rater. Frequency is rated on a five point Likert scale, where a score of '0' is used when a behaviour is thought to 'never occur', a score of '1' is given when it is deemed to occur 'less than once a week', a score of '2' indicates that a behaviour is judged to occur 'a few times a week', a score of '3' represents a frequency of 'once a day' and a score of '4' is used when a behaviour is judged to occur 'several times a day'. Difficulty is also rated on a five point Likert scale where a score of '0' is used when a behaviour is thought to 'never occur' (and therefore does not require 'managing'); a score of '1' is given when it is deemed to be 'a minor problem, not difficult to manage'; a score of '2' indicates that a behaviour is judged to be 'a moderate problem, fairly easy to manage'; a score of '3' represents that it is 'a major problem, fairly challenging to manage'; and a score of '4' is used when a behaviour is felt to be 'an extreme problem, almost impossible to manage'. The two scores are multiplied to produce a frequency x difficulty (FxD) score for each item. Paper and online versions of the BERRI include scoring instructions. Raters are asked to insert a score for the frequency of each behaviour (as per the scales outlined above) and where a behaviour occurs to also score the difficulty that it presents. Raters are reminded that the difficulty score is indicative of the amount of care and support the rater feels is required in relation to each occurring behaviour and not how well the setting is able to prevent it.

The 87 items on the BERRI are clustered into five scales *Behaviour*; *Emotional well-being (Emotions)*; *Risk to self and others (Risk)*; *Relationships*; and *Indicators of neuropsychological difficulties* (coined *Psychological indicators*), giving the measure its name. The Behaviour scale (19 items) is intended to measure behavioural difficulties, such as aggression and damage to property. The Emotions scale (18 items) is intended to identify difficulties associated with affect (e.g. 'Low

mood/sadness/crying') and self-concept (e.g. 'Lacks self-esteem/pride, has a poor self-image'). The Relationships scale (15 items) is intended to measure difficulties in relating to peers and carers (e.g. 'Trying to be in control of everyone around them'). The Risk scale (14 items) contains items pertaining to a young person's level of risk towards themselves and others (e.g. 'Running away/absconding'). Finally, the Indicators scale (21 items) includes items said to be possibly indicative of a neurodevelopmental 'condition' (e.g. 'Obsessions or narrow all-consuming interests').

2.2.2 The SDQ (parent version)

The SDQ (Appendix K) is a 25 item behavioural and emotional screening questionnaire for young people between the ages of two and 17 years. The parent version is designed to be completed by primary carers of a child. The questionnaire is comprised of five scales:

1. Emotional symptoms (five items; e.g. 'Often unhappy, down-hearted or tearful')
2. Conduct problems (five items; e.g. 'Often fights with other children or bullies them')
3. Hyperactivity/inattention (five items; e.g. 'Constantly fidgeting or squirming')
4. Peer relationship problems (five items; e.g. 'Rather solitary, tends to play alone')
5. Prosocial behaviour (five items; e.g. 'Considerate of other people's feelings').

The scores of scales one to four are added together to produce a 'total difficulties' score, for which clinical cut offs have been defined. The SDQ has been found to have acceptable internal consistency, good test-retest and inter-rater reliability, and moderate

to good construct validity (Stone *et al*., 2010). Further to this, a relationship has been identified between the 'total difficulties' score and the likelihood of a young person having a psychiatric diagnosis (Goodman, 2001). As the content of the SDQ and the BERRI share some similarities and the former is the mandated outcome measure for use with LAC, it was determined that the SDQ represented a good comparison measure for establishing construct validity.

2.2.3 The Novel Risk Questionnaire

The SDQ does not contain a scale comparable to the Risk scale in the BERRI and there is paucity of brief structured risk assessment tools for use with adolescents. Consequentially, to explore the construct validity of the BERRI Risk scale, a novel questionnaire (NRQ; Appendix L) was developed collaboratively with representatives from the participating residential care provider's senior management team. The NRQ contains a set of 'risk incidents' routinely monitored by the residential care provider for each child in their care for the purposes of monitoring and auditing.

1. Absconding/going missing
2. Self-harm
3. Suicidal behaviour
4. Assault towards others
5. Victim of assault
6. Allegations made (unfounded only)
7. Substance misuse
8. Concerns around Child Sexual Exploitation
9. Setting fires

Raters are required to detail the frequency of occurrence for each 'risk incident' over the rating period. As homes run by the participating care provider routinely complete the BERRI on a quarterly basis, it was decided that it would be most appropriate to collect the frequency of the above risk incidents over a three-month period. As the NRQ was created for the purpose of this study, its psychometric properties are unknown.

2.3 Data collection

The data used to evaluate the structure and psychometric properties of the BERRI were drawn from the two samples outlined below. For a detailed discussion of the ethical considerations taken with regards to data collection, please see Appendices M - N.

2.3.1 Naturalistic Sample

The BERRI is the routine outcome measurement tool of choice in several residential children's homes and has been implemented on a national level by the second largest provider of residential care in the UK. The BERRI is completed using an online system allowing trained users to track changes in scores over time. Data are anonymised at the point of entry, with each young person being assigned a randomly generated BERRI ID code. Data entered between December 2014 and March 2018 were extracted from the online system. This comprised 1569 BERRIs concerning 538 young people living in residential children's homes run by three separate providers. No demographic data were available for this sample.

2.3.2 Recruited Sample

For the purposes of establishing inter-rater reliability and construct validity, data were gathered from a number of children's homes (operated under one provider) who routinely use, and are trained in using, the BERRI. The members of staff taking part in the study would have been provided with training on how to use the BERRI either by the creator of the measure on its introduction to the care home or during their induction training if joining the residential care provider after this point. It is acknowledged however that a one hundred percent training rate could not be guaranteed. The care provider agreed to approach 40 of their children's homes and request that each member of care staff in these homes complete a BERRI and SDQ for each of the children in their care (Appendix O). Home managers were asked to complete the NRQ for each child in their home (Appendix P). The residential children's provider assigned a novel identification code to each child residing in one of the 40 target homes in order to allow the matching of questionnaires pertaining to the same young person. Care staff were instructed not to discuss their ratings with others. They were given a one-week period to complete the questionnaire pack to allow for shift rotations. As an incentive for questionnaire completion, staff and home managers were offered the opportunity to enter into a prize draw to win one of three £25 vouchers. Once the questionnaire packs had been completed, they were returned via post to the care provider's head office, where they were checked to ensure the absence of any identifiable information before being forwarded to the Chief Investigator.

Of the 40 children's homes, 32 returned questionnaire packs. A total of 616 BERRIs were completed for 92 children (a median of seven raters per child). BERRIs were excluded as a consequence of being incomplete (n = 6) or in instances where identical responses were provided for the same child by different raters (n = 10),

suggesting collaboration in completing questionnaires, which would confound estimates of inter-rater reliability. Of the 92 children, 53 and 72 children also had SDQ and NRQ data respectively. Demographic data were provided for 72 of the 92 children, who ranged in age from seven to 18 years (M = 13.91 years, SD = 2.74 years) and of whom 57 were male and 15 were female.

2.4 Criteria and hypotheses

2.4.1 Internal consistency

In line with Nunnally and Bernstein's (1994) recommendation, scales were deemed to have good internal consistency when Cronbach's alpha ranged from .70 to .90.

2.4.2 IRR

ICC confidence intervals were interpreted in line with Koo and Li's (2016) guidelines whereby ICC:

- < .5 = poor reliability;
- .5 - .74 = moderate reliability;
- .75 - .89 = good reliability;
- > .9 = excellent reliability.

2.4.3 Construct validity

Hypotheses were derived a priori concerning how each BERRI scale was expected to converge with and diverge from the NRQ and scales on the SDQ (Table 1). Tests of these hypotheses were conducted using Bonferroni adjusted alpha levels of .002 (Appendix Q) to control for the increased chances of Type I errors associated with

multiple correlations. Correlations were said to be indicative of convergent validity where $r > .05$, as per Cohen's (1998) criteria. Evidence of divergent validity was concluded where scales did not correlate significantly when it had been hypothesised that they would not.

Table 1: Convergent and divergent validity hypotheses

BERRI Scale	Convergent Validity Hypothesis	Divergent Validity Hypothesis	Justification
Behaviour	There will be a strong significant positive correlation with SDQ-Conduct problems	There will not be a significant correlation with SDQ-Emotional symptoms	BERRI-Behaviour and SDQ-Conduct problems scales both focus on observable behavioural difficulties, such aggression and bullying. This presentation of difficulty is considered different to that measured by SDQ-Emotional symptoms, which focuses on the emotional expression of low mood and anxiety.
Emotions	There will be a strong significant positive correlation with SDQ-Emotional symptoms	There will not be a significant correlation with SDQ-Conduct problems	BERRI-Emotions and SDQ- Emotional symptoms scales both focus on the emotional expression of low mood and anxiety. The former also considers self-concept. In contrast, SDQ-Conduct problems on the behavioural expression of difficulties.
Relationships	There will be a strong significant positive correlation with SDQ-Peer problems	There will not be a significant correlation with the NRQ	BERRI-Relationships and SDQ-Peer problems scales both focus on difficulties in forming and maintaining peer relationships. The former also considers carer relationships and wider relational patterns. The NRQ largely focuses on factors which place a young person at risk of harm from others as opposed to their level of risk in relation to others, as such it is felt that any correlation between this scale and SDQ-Relationships is unlikely to be significant.
Risk	There will be a strong significant positive correlation with the NRQ	There will not be a significant correlation with SDQ-Peer problems	The BERRI-Risk scale and NRQ both consider behaviours that place a young person at risk of harm, with the latter also considering the risk of harm by the young person to others. The BERRI-Risk scale largely focuses on factors which place a young person at risk of harm from others as opposed to their level of risk in relation to others, as such it is felt that any correlation between this scale and SDQ-Peer problems is unlikely to be significant.
Indicators	There will be a strong significant positive correlation with SDQ-Total difficulties	There will not be a significant correlation with SDQ-Pro-social	The BERRI-Indicators scale contains items relating to neurodevelopmental conditions. Young people with psychiatric diagnoses have been found to have higher SDQ-Total difficulties scores than those without. It follows that there should be a relationship between BERRI-Indicators and SDQ-Total difficulties. The SDQ-Prosocial scale measures positive behaviours towards others. There is no theoretical reason why this scale and BERRI-Indicators should be related.

3. Results

The psychometric properties of the BERRI as used are reported first, before the analysis of the measure's structure. The psychometric properties of the scales as devised through Principal Component Analysis (PCA) are then reported.

3.1 Internal consistency

Internal consistency was explored using data gathered from the naturalistic and recruited samples. FxD scores for each item were used. In order to control the impact of error associated with individuals' scoring profiles on the analyses, when a child had multiple completed BERRIs, one BERRI was chosen at random for each individual, resulting in a final sample of 630 (naturalistic sample n = 538, recruited sample n = 92). As a rule of thumb, sample sizes of over 300 are thought to be sufficient for the internal validation of psychiatric scales (Rouquette & Falissard, 2011).

Cronbach's alpha for all scales fell within the range indicative of good internal consistency, ranging between .836 and .875 (Table 2).

Table 2: Cronbach's alpha for original BERRI scales

Scale	Cronbach's Alpha	Number of items
Behaviour	.836	19
Emotions	.856	18
Relationships	.875	15
Risk	.838	14
Indicators	.859	21

3.2 IRR

In order to calculate inter-rater reliability it was required that a subsample be created such that each young person was rated by the same number of carers. It was calculated that a cut-off of six raters per individual would result in the optimum number of data points being retained (n = 59). Cases with fewer than six ratings were excluded from the analysis (n = 34). In instances where a child had been rated by more than six raters, excess ratings were excluded at random. The final sample used to ascertain inter-rater reliability consisted of 58 children, each rated by six staff members. As a rule of thumb, it is suggested that to establish IRR, a sample size of 30 individuals rated by at least three individuals should be obtained (Koo & Li, 2016).

The Behaviour and Risk scales, along with Total Score had ICC values indicative of 'moderate to good' IRR (taking into consideration 95% CIs). ICC for the Emotions, Relationships and Indicators scales was indicative of 'poor to moderate' IRR. See Table 3.

Table 3: Results of ICC Calculations Using Single-Rating, Absolute-Agreement, 1-Way Random-Effects Model for original measure

Scale	Single Measures ICC	95% Confidence Interval		F Test With True Value 0			
		Lower Bound	Upper Bound	Value	*df1*	*df2*	Sig
Behaviour	.676	.579	.767	13.500	57	290	.000
Emotions	.502	.390	.622	7.057	57	290	.000
Relationships	.599	.492	.705	9.946	57	290	.000
Risk	.735	.649	.814	17.631	57	290	.000
Indicators	.579	.471	.689	9.257	57	290	.000
Total Score	.696	.603	.784	14.768	57	290	.000

3.3 Construct Validity

In order to control for the impact of error associated with individual children's scoring profiles on the analyses, only one questionnaire set was retained for each child. In instances where only one rater had completed the SDQ for a young person, this rater's questionnaire set was chosen by default. In instances where no SDQ had been completed for a child (and as such only multiple BERRIs and the NRQ score were available) or where multiple raters had completed the SDQ in addition to the BERRI, the data for that child was ordered according to the BERRI total score from lowest to highest and the data set provided by the rater with the median BERRI score was retained. The final sample consisted of 87 BERRIs, 53 SDQs and 72 NRQs. An a priori power calculation using G*Power suggested that to identify correlations of 0.5, a sample size of 29 would be required.

The data were explored for outliers and deviation from normality through the visual inspection of histograms and p-p plots (Appendix R). On each of the BERRI

scales, including total score, the distributions were positively skewed. The distributions of scores on the SDQ subscales were mixed, with scores on the 'pro-social' subscale and total score appearing normal and scores on the remaining subscales deviating from normality (though less markedly than for the BERRI scales). Outliers were common among the BERRI and NRQ scores. These were retained as they were viewed to be representative of the population sampled, as opposed to measurement or inputting errors (Field, 2017). Whilst the sample sizes in this analysis would generally be considered large enough to trigger central limit theorem and therefore negate the need for a non-parametric test (i.e. N >30, Field, 2017), there is evidence to suggest that for heavy-tailed distributions (such as those seen for the BERRI scales) samples in excess of 160 may be required for central limit theorem to be applied (Wilcox, 2010). This, combined with the decision to retain outliers and data transformation failing to generate normally-distributed data, resulted in a decision to employ non-parametric correlations (Spearman's Rho) to test the a priori hypotheses.

There was evidence of convergent validity for the Behaviour, Relationships, Risk and Indicators scales, with each showing statistically significant strong correlations with the hypothesised scales. There was poorer evidence of the convergent validity of the Emotions scale with the correlation between BERRI-Emotions and SDQ-Emotions not exceeding the .5 cut-off as expected. The correlation was however statistically significant and moderate in size ($r_s = .494$, $p < .002$).

There was evidence of divergent validity for all BERRI scales, with none having a significant correlation ($p < .002$) where it was hypothesised that they would not.

Table 4: Construct validity analyses for original measure

			NRQ	SDQ Emot	SDQ Cond	SDQ Hyp	SDQ Peer	SDQ ProSoc	SDQ Total Diff
Spearman's rho	BERRI Behav.	Correl. Coeff.	.593*	**0.067**	**.635***	.545*	.327	-0.184	-
	BERRI Emot.	Correl. Coeff.	.433*	**.494***	**.281**	.473*	.271	-0.094	-
	BERRI Rel.	Correl. Coeff.	**.326**	.294	.487*	.455*	**.551***	-.285	-
	BERRI Risk.	Correl. Coeff.	**.659***	0.088	0.220	.507*	**-0.163**	.285	-
	BERRI Indicat.	Correl. Coeff.	0.172	.367	.434*	.566*	.400	**-0.199**	**.693***
		n	72	53	53	53	53	53	53

*Note: *. Correlation is significant at the p < .002 level (2-tailed). Coefficients in green denote correlations related to convergent validity hypotheses, those in red denote correlations related to divergent validity hypotheses.*

3.4 Structure

PCA rather than exploratory factor analysis was used as a consequence of the BERRI being designed to operationalise the need of LAC using several domains (as opposed to a single latent construct). Structure was explored using the same sample utilised for the exploration of internal consistency. Suitability of the data for PCA was assessed a priori: All items correlated with at least another item on the BERRI to at least r = .3, with the exception of four items that were excluded from the factor analysis ('Harm to animals'; 'Gender identity issues'; 'Suicidal thoughts/plans/talk of non-existence or death'; 'Repetitive behaviour or rituals'). The Kaiser-Meyer-Olkin (KMO) measure indicated that the sample size (n = 630) was 'marvellous', (KMO = .912) according to Kaiser and Rice's (1974) criteria. Through inspection of the anti-image correlation matrix, it was ascertained that all items had KMO values in excess of .67 (greater than the acceptable limit of .5; Kaiser & Rice, 1974). Bartlett's Test of

Sphericity was statistically significant ($\chi 2$ (3403) = 26188.47, $p < .001$) (Appendix S). The determinant of the correlation matrix was R = 1.148, indicating an absence of problematic multicollinearity (Field, 2017). A PCA was therefore conducted on the remaining 83 items. An orthogonal rotation (varimax) was used as the correlation between components could not be theoretically assumed.

There are known problems with sole reliance on eigenvalues for the extraction of components, especially with regards to overestimation of the number of components to retain (Field, 2017). Stevens (2002) suggests that the use of Kaiser's criterion (Kaiser, 1960) of retaining components with eigenvalues greater than 1 is most appropriate when the number of variables does not exceed 30 and the communalities after extraction are all greater than .7. In the case of the BERRI, the number of variables analysed exceeded 30 (83) and the communalities ranged from .47 to .79 (M = .63, SD = 0.07), thus not meeting the above criteria. Stevens (2002) recommends that where the criteria are not met and sample size exceeds N = 300, the use of a scree plot is more acceptable for decisions regarding factor extraction. Through visual inspection of the scree plot (Figure 1), two points of inflection were identified (at component four and component six) indicating that exploration of three and five component models was warranted (Field, 2017). As per Stevens' (2002) recommendation, component loadings with an absolute value over .4 were interpreted.

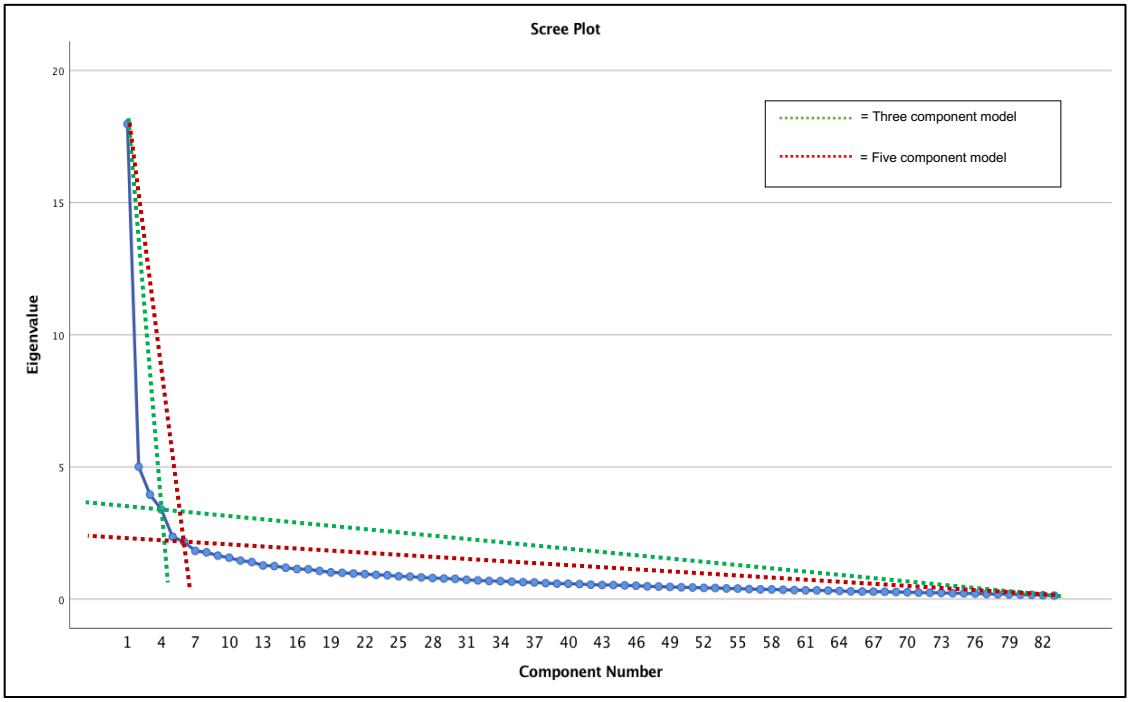


Figure 2: Scree plot from PCA

After consideration, the five component solution (Table 5) was considered most interpretable and explained more of the variance (39.5%) than the three component solution (32.4%; Appendix T). The five components were labelled 'Behavioural/deactivating'(20 items), 'Indicators'(15 items), 'Risk'(nine items), 'Emotional needs'(11 items) and 'Relational/hyperactivating'(six items). These scales are further outlined in the Discussion. The internal consistency of each scale was examined, at which point the removal of one item, 'Lethargy (including being up at night and sleeping in day)', from the Emotional needs scale was indicated (Appendix U). The removal of this item did not impact on the interpretability of the five component solution. Twenty-two items did not load significantly onto any component and as such were not retained, these are listed under 'other' in Appendix V.

Table 5: Summary of principal components analysis results for the BERRI (N = 630). Factor loadings under .4 are hidden.

Item	Rotated Component Loadings				
	Behavioural-deactivating	Indicators	Risk	Emotional Needs	Relational-hyperactivating
Verbal aggression	0.787				
Need to provoke chaos/winds up others/test their response	0.771				
Argumentative/winding others up	0.765				
Under/over eating, storing of food, self-induced vomiting	0.763				
Extreme emotional response/tantrums/anger/rage	0.758				
Physical aggression towards carers or other children	0.747				
Non-compliance/defiance	0.733				
Irritable/mood swings	0.668				
Trying to be in control of everyone around them	0.643				
Impulsiveness and impatience (e.g. interrupting)	0.627				
Damage to property	0.621				
Hyper arousal (always seems ready for fight/flight)	0.600				
Screaming/shouting/too loud	0.598				
Proud of negative characteristics (e.g. 'devil'/toughness)	0.518				
Cannot understand other people's thoughts and feelings	0.496	0.435			
Lacks empathy	0.495				
Seems to have no guilt	0.476				
Racism or other prejudice	0.420				
Interest in violence/death/gore	0.419				
Always alert for danger signs/agitated/can't settle	0.405				
Level of understanding poor/learning disability		0.628			
Difficulties with speech or understanding of language		0.620			
Problems with skills of daily life		0.574			
Poor grip on reality (e.g. bizarre beliefs/sees or hears things)		0.550			
Difficulties with motor co-ordination		0.547			
Literal understanding of language (e.g. can't get jokes or lies)		0.522			
Incoherent speech/makes noises		0.515			
Obsessions or narrow all-consuming interests		0.510			
Poor judge of personal space/ poor social judgement		0.506			
Lack of concern about how others see them		0.448			
Can't separate facts from fantasy (e.g. tells elaborate stories)		0.447			

Lacks concentration/distractible/poor attention span	0.425	0.445			
Has odd movements such as tics/rocking/flapping		0.433			
Struggles with change/has a rigid need for routine		0.416			
Echolalia (copies back what is said, like a parrot)		0.412			
Choosing unsafe peers/environments			0.806		
Getting involved in crime			0.794		
Placing self at risk of exploitation			0.783		
Running away/absconding			0.748		
Drug, solvent or alcohol abuse			0.708		
No fear, puts self in danger, recklessness, thrill seeking			0.658		
Sexually active in a risky way or sex working			0.575		
Cheating or stealing			0.558		
No cause-effect reasoning/can't predict consequences of actions			0.497		
Lack of joy/laughter/emotionally flat				0.687	
Low mood/sadness/crying				0.629	
Self-critical/can't take praise				0.615	
Lacks self-esteem/pride, has a poor self-image				0.610	
Withdrawn/uncommunicative				0.589	
Not able to show full range of feelings				0.575	
Lack of comfort–seeking (e.g. if hurt)				0.556	
Worries/phobias				0.552	
Self blame or unrealistic expectations of self				0.535	
Poor sense of own identity/culture				0.440	
Fear of normal situations/carers				0.437	
Makes indiscriminate, superficial, overly close relationships					0.513
Making unfounded disclosures					0.505
Attention seeking/clingy/needy/whingeing					0.494
Fictitious illness/ailments or hypochondria					0.481
Self harm: cutting/tying ligatures/overdosing					0.447
Self harm: biting/scratching/pulling hair/head banging/pica					0.404
Component Eigenvalues	17.975	5.006	3.950	3.393	2.354

3.5 Internal consistency of revised scales

Cronbach's alpha for the Indicators, Risk and Emotional needs scales fell within the range indicative of 'good' internal consistency. Cronbach's alpha for the Behavioural/deactivating scale (α = .940) was above the desired cut off, and for the Relational/hyperactivating scale it was slightly below (α = .629). Table 6. There were no instances where the deletion of an item would lead to an increase in Cronbach's alpha.

Table 6: Cronbach's alpha for revised BERRI scales

Scale	Cronbach's Alpha	Number of items
Behavioural/deactivating	.940	20
Indicators	.850	15
Risk	.884	9
Emotional needs	.837	11
Relational/hyperactivating	.629	6

3.6 IRR of revised measure

The Behavioural/deactivating scale and BERRI Total Score had ICC values indicative of 'moderate to good' IRR (taking into consideration 95% CIs). ICC for the Emotional needs, Relational/hyperactivating, Risk and Indicators scales was indicative of 'poor to moderate' IRR. See Table 6. IRR classification changed only for the Risk scale, which moved from 'moderate to good' to 'poor to moderate'.

Table 7: Results of ICC Calculations Using Single-Rating, Absolute-Agreement, 1-Way Random-Effects Model for revised measure

Scale	Single Measures ICC	95% Confidence Interval		F Test With True Value 0			
		Lower Bound	Upper Bound	Value	*df1*	*df2*	Sig
Behavioural/ deactivating	.666	.569	.760	12.988	57	290	.000
Indicators	.418	.305	.545	5.306	57	290	.000
Risk	.582	.475	.691	9.359	57	290	.000
Emotional needs	.386	.275	.514	4.772	57	290	.000
Relational/ hyperacti vating	.375	.264	.504	4.589	57	290	.000
Total Score	.672	.575	.765	13.312	57	290	.000

3.7 Construct validity of revised measure

Although the extracted factors could not have been predicted prior to analysis, because they were deemed to approximately map onto the original five BERRI subscales, the new scales were compared with the SDQ subscales and NRQ utilising the same hypotheses for convergent and divergent validity indicated in Table 1. There was no evidence of convergent validity for any of the revised scales, with none correlating significantly with the hypothesised scales. As few BERRI correlations were significant, it was not possible to take a non-significant correlation with a hypothesised scale as evidence of divergent validity. In the case of the Relational/hyperactivating scale, there was a significant correlation with the NRQ, contrary to the divergent validity hypothesis. It is likely that this is a consequence of the new Relationships scale containing items concerning self-harm and 'unfounded disclosures', which closely

mirror some NRQ items. As such, the correlation between the two scales could be expected.

Table 8: Construct validity analyses for revised measure

			NRQ	SDQ Emot	SDQ Cond	SDQ Hyp	SDQ Peer	SDQ ProSoc	SDQ Total Diff
Spear. rho	BERRI Behav.	Correl. Coeff.	.390*	**0.156**	**0.172**	0.228	0.183	-0.076	-
	BERRI Indicat.	Correl. Coeff.	.284	0.104	0.219	0.102	0.264	**-0.092**	**0.251**
	BERRI Risk.	Correl. Coeff.	**.278**	-0.012	0.077	.502*	**-0.023**	0.106	-
	BERRI EmNee.	Correl. Coeff.	.330	**0.083**	**-0.132**	-0.033	0.024	0.218	-
	BERRI RelHyp.	Correl. Coeff.	**.493***	0.023	0.033	0.172	**0.204**	-0.057	-
		n	72	53	53	53	53	53	53

*Note: *. Correlation is significant at the p <.002 level (2-tailed). Coefficients in green denote correlations related to convergent validity hypotheses, those in red denote correlations related to divergent validity hypotheses.*

4. Discussion

The aim of this study was to explore the psychometric properties of the BERRI as it is currently being used and to examine whether these properties might be improved by the empirical exploration of its structure through Principal Component Analysis (PCA) and subsequent extraction of components. The internal consistency of the original BERRI's scales was measured through the calculation of Cronbach's alpha. An exploration of this measure's inter-rater reliability (IRR) was then conducted through the calculation of intra-class correlation coefficients. This was followed by an assessment of construct validity through the use of correlation analysis to ascertain the original BERRI's convergent and divergent validity. Finally, an exploration of the BERRIs structure was conducted through the use of PCA and the aforementioned analyses were repeated using the suggested structure. Key findings will be summarised, in addition to further consideration of the structure indicated by the PCA and how attachment theory might be used as a framework to make sense of the extracted factors. A discussion of this study's strengths and limitations and its implications for future research and clinical practice will follow.

4.1 Original BERRI internal consistency

This study indicated that internal consistency of the BERRI as used was 'good', with Cronbach's alpha for each scale as used falling between .70 and .90. In contrast, a meta-analysis of the psychometric properties of the parent SDQ found Cronbach's alpha to fall below .70 for four out of the five subscales (Stone *et al*., 2010).

4.2 Original BERRI IRR

This study indicated that the IRR of the BERRI as used was 'moderate to good' for the Behaviour and Risk scales, in addition to the total score. The IRR of the Relationships, Emotions and Indicators scales was less promising and fell within the 'poor to moderate' range. It is important to consider the IRR of the BERRI in the context that its scoring system attributes significant weight to a rater's perception of how difficult each 'problem' is to manage and then multiplies this with their assessment of frequency. Difference between raters' perceptions is therefore is multiplied through the current FxD scoring system. Raters' perceptions of difficulty are likely to be influenced by many factors, from their experience of caring for LAC to their own personalities and life experiences, whilst ratings of frequency could also be subjective and influenced by what had been witnessed and recalled from time on shift. In this context, the IRR of the BERRI as measured here could be viewed to be better than might be expected, particularly given that the IRR data is in keeping with, and exceeds in some cases, that of widely used outcome measures for young people (e.g. the Brief Psychiatric Rating Scale for Children, as measured by Mullins *et al*., 1985; and the HoNOSCA, as measured by Brann *et al*., 2001; Garralda *et al*., 2000). Parent-teacher inter-rater reliability for the SDQ was found by Stone et al. (2010) in their meta-analysis to be adequate, however this was measured using Pearson's r rather than ICC and as such only took into account consistency rather than absolute agreement.

There are several possible reasons for the poorer IRR of the Relationships, Emotions and Indicators scales. Considering the Relationships scale, high rates of children with disorganised attachment styles have been found among LAC in residential care (Bifulco *et al*. 2017). By definition, children with disorganised attachment styles have not been able to develop an organised strategy to promote feelings of relational safety as a consequence of the unpredictability of their early attachment figures (e.g. Hesse & Main, 2000). As such, they often relate differently to different adults in order

to ensure that their needs are met (Silver, 2013). The Relationships scale on the BERRI asks raters to assess how a young person relates to others in their environment. Different raters' perceptions of this are likely to vary for the reasons discussed above, in addition to the impact of the carer's own attachment style on their relationship with the young person (Berlin & Cassidy, 2001) and the attachment relationship between a given child and carer. Many of the items on the Emotions scale relate to the young person's internal mood state or feelings towards themselves (e.g. self-blame, low mood, poor sense of own identity). In a similar manner to the Relationships scale, a young person's propensity to share such information with a rater is likely dependent on their relationship. The poorer IRR for the indicators scale was somewhat more surprising, considering that the content could be presumed to be more objective at face-value. It is however possible that raters may differ in the level of experience they have of identifying factors that could be considered indicative of 'autism spectrum disorder' (ASD), 'attention deficit hyperactivity disorder' (ADHD) and 'intellectual disability' (ID), which feature heavily in this scale. Additionally, there is significant debate concerning the reliability of diagnoses such as ASD (e.g. Mallet & Timimi, 2016) and ADHD (e.g. Timimi & Leo, 2009) using the current systems of classification, indicating that disagreement between raters on items on the Indicators scale may be unsurprising.

In addition to considering why some scales of the BERRI may have poorer IRR, it is important to consider the value of inter-rater *dis*agreement for the BERRI. Systemic approaches consider differences between perspectives to be useful information (Tomm, 1987). This, combined with the aforementioned evidence suggesting that a large proportion of LAC are likely to interact differently with different adults, suggests that the difference in the ratings of these adults could be argued to be a source of valuable information when formulating the needs of a young person and to

facilitate reflective discussions between staff members. Equally, differences in scores could provide further useful information when considering where intervention or support might be most beneficial in the systems surrounding LAC.

4.3 Original BERRI construct validity

This study provided evidence of the BERRIs construct validity through the measurement of convergent and divergent validity. The Behaviour, Relationships, Risk and Indicators scales all converged with the scales on the SDQ as hypothesised. The Emotions scale correlated with the SDQ Emotions scale as hypothesised, however this correlation fell very slightly below the .5 cut off. This could be due to the inclusion of items in the BERRI Emotions scale that are concerned with a young person's feelings towards themselves, whereas the focus is on emotional affect in the SDQ subscale. There was evidence of divergent validity for all BERRI scales, with none correlating significantly with any of the SDQ subscales or NRQ where it was hypothesised that there would be no significant association. The SDQ has been found to have similarly good construct validity, with Stone et al. (2010) reporting evidence of good convergent and divergent validity for the SDQ scales when compared with the CBCL.

4.4 Structure

Through factor analysis, a structure for the BERRI was extracted, which was similar to that devised at face-value during the measure's development. Five scales were retained: 'Behavioural/deactivating'; 'Emotional needs'; 'Risk'; 'Relational/hyperactivating'; and 'Indicators'. Several of these scales can be helpfully understood in the context of attachment theory. Bowlbly (1969) theorised that where a child's internal working model predicts that caregivers will be unpredictable or inconsistent in meeting their needs, they develop secondary attachment strategies,

which may be *hyperactivating* or *deactivating* in nature. Deactivating strategies can be understood as those which attempt to shut down attachment feelings, as opposed to hyperactivating strategies, which escalate them (Dallos, 2006).

In consideration of the items loading onto the Behavioural/deactivating scale, this scale might be best described as including behaviours which are representative of deactivating strategies. Such strategies often make sense for a young person when active attachment seeking is seen as non-viable or threatening, resulting in a perceived need for self-reliance and withdrawal from relationships (Dallos, 2006). This pattern of relating is also described as an 'avoidant attachment style' (Ainsworth *et al*., 1978). Items in this scale share a common theme of behaviours that keep others at a distance, both physically (e.g. 'Physical aggression…') and emotionally (e.g. 'Lacks empathy'). Items in this scale could also be understood as serving to support self-reliant emotional regulation (e.g. 'Under/over eating…') and maintenance of feelings of safety (e.g. 'Trying to be in control of everyone around them').

The Emotional needs scale could be understood to reflect young people's feeling states, including affect and their feelings towards themselves. These might be formulated as being associated with the internal working models held by young people about themselves (e.g. 'Self-critical, can't take praise') in additional to emotional reactions to adverse life experiences (e.g. 'Low mood, sadness, crying').

The Risk scale could be understood as containing items which might represent hyperactivating and deactivating strategies which have a propensity to lead to unintended harmful consequences. For example, a young person being 'sexually active in a risky way…' could be formulated as a method of seeking attachment relationships. Similarly, 'drug, solvent or alcohol use' could be understood as a self-reliant method of regulating emotional feelings to protect against the threat of emotional overwhelm. It is

important to note here the dangers of locating 'vulnerability' to risk and exploitation within an individual, rather than those who present a risk to them (Boyle, 2003). As such, the items 'Placing self at risk of exploitation' and 'No fear, puts self in danger…' could potentially benefit from rewording.

The Relational/hyperactivating scale could be understood to contain items representative of hyperactivating strategies. Such strategies make sense for a young person when attuned care is (or has been) inconsistently available and as such care must be actively sought to maximise the probability of their needs being met. With regards to the loadings of both questionnaire items referring to self-harm onto this component, it is acknowledged that function served by self-harm varies between individuals (e.g. Edmondson *et al*., 2016) and could theoretically link onto other scales, such as Emotional needs and Risk. It is postulated that these items loading most strongly onto the Relational-hyperactivating scale might be reflective of the BERRI being carer-rated. Self-harm that comes to the attention of raters might be more likely to serve an attachment seeking function than self-harm primarily intended, for example, as a strategy for emotional regulation, with the latter perhaps being more concealed from carers.

The Indicators scale is very similar in composition to the original BERRI and contributing items may best be understood as being related to neurodevelopmental 'conditions', such as ASD, ADHD and ID. With regards to the item, 'Poor grip on reality (e.g. bizarre beliefs/sees or hears things)', it is acknowledged that its placement in this scale could be viewed to support a much contested (e.g. Deacon, 2013) biomedical model of distress. The importance of understanding this item in the context of the rated individuals' lived experiences is emphasised.

Twenty-six items were not included in the final model. Some of the unretained items, including 'seeking punishment', 'seeking restraint' and 'gender identity issues', could be viewed as having the potential to be read and used in a pejorative and pathologizing manner. As such the measure likely benefits from their exclusion. No clear link was identified between the unretained items.

4.5 Psychometric properties of the revised BERRI

Internal consistency was found to be 'good' for three of the five new scales (Indicators, Risk and Emotional needs), compared to all five of the original scales having 'good' levels of internal consistency. Cronbach's alpha is known to be impacted by the number of items in a scale, with alpha increasing as the number of items increases (e.g. Abdelmoula 2015). The small number of items on the Relational/hyperactivating scale may have contributed to its alpha falling below the desired range. Similarly, the Behavioural/deactivating scale's large number of items may have contributed to its Cronbach's alpha falling above the recommended cut off. A very high Cronbach's alpha (> .90) is thought by some to be indicative of items on a scale being too highly correlated and thus redundant (e.g. Nunnally and Bernstein, 1994). Some debate exists in the literature with regards to the recommended upper limit indicative of 'good' internal consistency, with some researchers (e.g. Terwee *et al*., 2007) advocating a cut off of $\alpha = 0.95$, higher than that employed in the present study. Under these criteria, the Behavioural/deactivating scale would be considered to have good internal consistency.

The IRR of the revised BERRI was not substantially different from that of the original. Only the Risk scale moved with regards to classification, changing from

‘moderate-good’ to ‘poor-moderate’. No evidence was found in support of the revised measure’s construct validity. It is possible that the use of the SDQ as a comparison measure was less appropriate for the revised measure compared to the original scales. Although the original five-scale structure was largely supported by the factor analysis, it could be argued that the content of each scale changed slightly, such that it would be unexpected for the revised scales to converge with and diverge from the SDQ and NRQ in the same manner as the original scales.

There was no evidence to suggest that the psychometric properties of the BERRI were improved through the exploration of its structure and subsequent extraction of components. As such, it could be concluded that the original structure should be retained with further consideration given to the contribution of the items excluded through PCA and item wording as discussed.

4.6 Strengths and limitations

This study is the first to assess the psychometric properties and structure of the BERRI. The study benefited from large sample sizes and the use of rigorous, empirically supported criteria. Limitations of this study include the sample being restricted to children in residential care only. This sample also only included ratings given by care staff, rather than SGO carers, foster carers or mental health professionals. The psychometric properties of the BERRI may differ for LAC in foster care or residing with family members under special guardianship orders (SGOs), or when ratings are provided by adults other than care staff. As such the results of the study cannot be generalised beyond LAC in residential care rated by care staff. In addition, demographic information was not available for the naturalistic sample, meaning that the scope for generalising the findings of this study is unclear. Finally, although the NRQ

was generated because of the absence of a suitable comparison measure, the use of this tool was problematic due to its own psychometric properties remaining unexplored.

4.7 Implications for research and clinical practice

The BERRI in its current from has poor-moderate to moderate-good inter-rater reliability, and good convergent validity. This suggests that it measures what it intends to measure and that scores on some subscales and the total score can be accurately interpreted when completed by different raters. In order to improve the measure's IRR, the weighting of the difficulty score in the overall score for each item may benefit from being revised, particularly in settings where multiple raters are likely such as in residential care homes. This might be achieved by adding, rather than multiplying, frequency and difficulty scores. Alternatively, a means of collecting differences in raters' scores could be developed, such that these differences could be used as prompts to support reflection and discussion amongst staff teams concerning their work with an individual young person (e.g. Mason, 1991). Exploration of the structure of the BERRI suggests that the original structure is psychometrically superior but may benefit from the revision of items identified as problematic through PCA. Additionally a review of the language used in the BERRI with a service user group would likely be beneficial in order to ensure that items maximise the understanding of the rated behaviours as serving a function, rather than being indicative of 'abnormalities'. Such information could usefully also be communicated to carers when being trained on use of the BERRI.

Overall, the BERRI shows significant promise as an outcome measure for LAC in residential care. In order to further understand the ability of the BERRI to assess change over time, it would be important to next explore the measure's test-retest reliability and responsiveness to change. This is vital for understanding the extent to which the measure can contribute to placement planning and monitoring. The BERRI's

'life events' scale would likely be pivotal to measuring it's responsiveness. As such, it would be important to fully explore the validity and reliability of this scale in its own right. As this study demonstrated that the BERRI's psychometric properties were not improved after ascertaining and empirically derived factor structure, further analysis of the measure's psychometric properties at an item level may be beneficial. This would allow the removal of psychometrically weaker items, which may serve to improve the psychometric robustness of the measure in addition to creating a shorter version more suited to routine clinical use. An item level analysis would also benefit from the consideration of the clinical utility of each item being assessed using qualitative feedback from clinicians and quantitatively assessing the frequency of use for each item.

References

Abdelmoula, M., Chakroun, W. & Akrout, F. (2015). The effect of sample size and the number of items on reliability coefficients: Alpha and rho: A meta-analysis. *International Journal of Numerical Methods and Applications*, *13*(1), 1-20.

Achenbach, T. M., Dumenci, L. & Rescorla, L. A. (2003). DSM-Oriented and empirically based approaches to constructing scales from the same item pools. *Journal of Clinical and Adolescent Psychiatry*, *32*, 328–340.

Ainsworth, M.D.S. (1973). The development of infant-mother attachment, in Caldwell, B.M. & Ricuti, H. *Review of Child Development Research, Vol. 3.* Chicago: Chicago University Press.

Berlin, L. J. & Cassidy, J. (2001). Enhancing early child parent relationships: Implications of adult attachment research. *Infants and Young Children , 14*(2), 64-77.

Bickman, L., Kelley, S. D., Breda, C., de Andrade, A. R. & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services*, *62*(12), 1423-1429.

Bifulco, A., Jacobs, C., Ilan-Clarke, Y., Spence, R. & Oskis, A. (2016). Adolescent attachment style in residential care: the attachment style interview and vulnerable attachment style questionnaire. *British Journal of Social Work, 47*(7), 1870-1883.

Bowlby, J. (1969). *Attachment, Attachment and Loss, Vol. 1.* New York: Basic Books.

Boyle, M. (2003). The dangers of vulnerability. *Clinical Psychology*, *24*(4), 27-30.

Brann, P., Coleman, G. & Luk, E. (2001). Routine outcome measurement in a child and adolescent mental health service: an evaluation of HoNOSCA. *Australian and New Zealand Journal of Psychiatry*, *35*(3), 370-376.

Broad, B. (1998). *Young people leaving care: life after the Children Act 1989*. London: Jessica Kingsley Publishers.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dallos, R. (2006). *Attachment narrative therapy: Integrating systemic, narrative and attachment approaches.* London: Open University Press.

Deacon, B. J. (2013). The biomedical model of mental disorder: A critical analysis of its validity, utility, and effects on psychotherapy research. *Clinical psychology review*, *33*(7), 846-861.

Denton, R. N. (2016). *The assessment of mental health in looked-after adolescents: an exploratory study of the brief assessment checklist for adolescents*. Unpublished doctoral thesis, University of Surrey.

Department for Children, Schools and Families. (2009). *Guidance notes for the completion of SSDA903 records: Children looked after by local authorities in England*. London: The Stationery Office.

Department for Education. (2016). *Children looked after in England (including adoption) year ending 31 March 2016.* London: The Stationery Office.

Department for Education. (2017). *Children looked after in England (including adoption), year ending 31 March 2017*. London, England: The Stationery Office.

Edmondson, A. J., Brennan, C. A. & House, A. O. (2016). Non-suicidal reasons for self-harm: a systematic review of self-reported accounts. *Journal of Affective Disorders*, *191*, 109-117.

Farmer, E., Lipscombe, J. & Moyers, S. (2005). Foster carer strain and its impact on parenting and placement outcomes for adolescents. *British Journal of Social Work*, *35*(2), 237-253.

Fisher, A. (2015). Review: Adoption, fostering, and the needs of looked-after and adopted children. *Child and Adolescent Mental Health, 20*(1), 5-12.

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). London: Sage.

Frogley, C. L. (2016). *Assessing the mental health needs of looked after children: a study investigating the utility of the brief assessment checklist for children*. Unpublished doctoral thesis, University of Surrey.

Garralda, M.E., Yates, P. & Higginson, I. (2000). Child and adolescent mental health service use: HoNOSCA as an outcome measure. *The British Journal of Psychiatry*, *177*(1), 52-58.

Goemans, A., Tarren-Sweeney, M., van Geel, M. & Vedder, P. (2018). Psychosocial screening and monitoring for children in foster care: Psychometric properties of the Brief Assessment Checklist in a Dutch population study. *Clinical child psychology and psychiatry*, *23*(1), 9-24.

Goodman, R., Ford, T., Corbin, T. & Meltzer, H. (2004). Using the Strengths and Difficulties Questionnaire (SDQ) multi-informant algorithm to screen looked-after children for psychiatric disorders. European child & adolescent psychiatry, 13(2), 25-31.

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *The Journal of Child Psychology and Psychiatry, 38*(5), 581-586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337-1345.

Goodman, A. & Goodman, R. (2012). Strengths and Difficulties Questionnaire scores and mental health in looked after children. *The British Journal of Psychiatry*, *200*(5), 426-427.

Happell, B. (2008). Determining the effectiveness of mental health services from a consumer perspective: Part 2: Barriers to recovery and principles for evaluation. *International Journal of Mental Health Nursing, 17*(2), 123-130.

Harmon, S. C., Lambert, M. J., Smart, D. W., Hawkins, E. J., Nielsen, S. L., Slade, K., *et al*. (2007). Enhancing outcome for potential treatment failures: Therapist/client feedback and clinical support tools. *Psychotherapy Research, 17*, 379 –392.

Hatfield, D., McCullough, L., Frantz, S. H. & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, *17*(1), 25-32.

Hesse, E. & Main, M. (2000). Disorganized infant, child, and adult attachment: Collapse in behavioral and attentional strategies. *Journal of the American Psychoanalytic Association, 48*(4), 1097-1127.

House of Commons Education Committee (2016). *Mental health and well-being of looked-after children*. London: The Stationery Office.

Iwaniec, D., Larkin, E. & Higgins, S. (2006). Research review: Risk and resilience in cases of emotional abuse. *Child & Family Social Work, 11*(1), 73-82.

Jee, S. H., Halterman, J. S., Szilagyi, M., Conn, A. M., Alpert-Gillis, L. & Szilagyi, P. G. (2011). Use of a brief standardized screening instrument in a primary care setting to enhance detection of social-emotional problems among youth in foster care. *Academic Pediatrics*, *11*(5), 409-413.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaiser, H. F. & Rice, J. (1974). Little jiffy, mark 4. *Educational and Psychological Measurement, 34*(1), 111-117.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Mallet, R. & Timimi, S. (2016). *Re-thinking autism: Diagnosis, identity and equality*. London: Jessica Kingsley Publishers.

Mason, B. (1991). *Handing Over: Developing Consistency Across Shifts in Residential and Health Settings: Developing Consistency Across Shifts in Residential and Health Care Settings (Systemic Thinking & Practice)*. London: Routledge.

Marquis, R. A. & Flynn, R. J. (2009). The SDQ as a mental health measurement tool in a Canadian sample of looked-after young people. *Vulnerable Children and Youth Studies*, 4(2), 114-121.

Mullins, D., Pfefferbaum, B., Schultz, H. & Overall, J.E. The brief psychiatric rating scale for children: Quantitative scoring of medical records. *Psychiatry Res, 19*(1), 43-49.

Nunnally, J.C. and Bernstein, I.H. (1994) The Assessment of Reliability. *Psychometric Theory*, *3*, 248-292.

Rock, S., Michelson, D., Thomson, S. & Day, C. (2013). Understanding foster placement instability for looked after children: A systematic review and narrative synthesis of quantitative and qualitative evidence. *British Journal of Social Work*, *45*(1), 177-203.

Rodrigues, V.C. (2004). Health of children looked after by the local authorities. *Public Health, 118*, 370-376.

Rouquette, A., & Falissard, B. (2011). Sample size requirements for the internal validation of psychiatric scales. *International Journal of Methods in Psychiatric Research, 20*(4), 235-249.

Rubin, D. M., O'Reilly, A. L. R., Luan, X. & Localio, A. R. (2007). The impact of placement stability on behavioral well-being for children in foster care. *Pediatrics*, 119, 336–44.

Sempik, J., Ward, H. & Darker, I. (2008). Emotional and behavioural difficulties of children and young people at entry into care. *Clinical child psychology and psychiatry*, *13*(2), 221-233.

Silver, M. (2013). *Attachment in common sense and doodles: A practical guide*. London: Jessica Kingsley Publishers.

Silver, M., Graham, R., Tucker, R. & Swann, R. *et al.* (2016). *Outcome measurement with children who are looked after in public care.* London and South East: CYP-IAPT Learning Collaborative.

Stevens, J.P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Earlbaum.

Stone, L.L., Otten, R., Rutger, C.M.E., Vermulst, A.A. & Janssens, J.M.A.M. (2010). Psychometric Properties of the Parent and Teacher Versions of the Strengths and Difficulties Questionnaire for 4- to 12-Year-Olds: A Review. *Clinical Child and Family Psychology Review, 13*(3), 254-274.

Tarren-Sweeney, M. (2007). The Assessment Checklist for Children—ACC: A behavioral rating scale for children in foster, kinship and residential care. *Children and Youth Services Review*, *29*(5), 672-691.

Tarren-Sweeney, M. (2013a). The Assessment Checklist for Adolescents—ACA: A scale for measuring the mental health of young people in foster, kinship, residential and adoptive care. *Children and Youth Services Review*, *35*(3), 384-393.

Tarren-Sweeney, M. (2013b). The Brief Assessment Checklists (BAC-C, BAC-A): Mental health screening measures for school-aged children and adolescents in foster, kinship, residential and adoptive care. *Children and Youth Services Review*, *35*(5), 771-779.

Teggart, T. & Menary, J. (2005). An investigation of the mental health needs of children looked after by Craigavon and Banbridge health and social services trust. *Child Care in Practice*, *11*(1), 39-49.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., *et al*. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*, *60*(1), 34-42.

Timimi, S. & Leo, J. (2009). *Rethinking ADHD: From brain to culture*. London: Palgrave.

Tomm, K. (1987). Interventive interviewing: Part II. Reflexive questioning as a means to enable self-healing. *Family process*, *26*(2), 167-183.

Wilcox R.R. (2010) Hypothesis Testing and Small Sample Sizes. In: Fundamentals of Modern Statistical Methods. Springer, New York, NY

Appendix A - Guidelines for submission to British Journal of Clinical Psychology (the target journal for the literature review)*

AIMS AND SCOPE

The *British Journal of Clinical Psychology* publishes original research, both empirical and theoretical, on all aspects of clinical psychology:

- clinical and abnormal psychology featuring descriptive or experimental studies
- aetiology, assessment and treatment of the whole range of psychological disorders irrespective of age group and setting
- biological influences on individual behaviour
- studies of psychological interventions and treatment on individuals, dyads, families and groups.

The Journal is catholic with respect to the range of theories and methods used to answer substantive scientific problems. Studies of samples with no current psychological disorder will only be considered if they have a direct bearing on clinical theory or practice.

The following types of paper are invited:

- papers reporting original empirical investigations;
- theoretical papers, provided that these are sufficiently related to empirical data;
- review articles, which need not be exhaustive, but which should give an interpretation of the state of research in a given field and, where appropriate, identify its clinical implications;
- Brief Reports and Comments.

MANUSCRIPT CATEGORIES AND REQUIREMENTS

Articles should be no more than 5000 words (excluding the abstract, reference list, tables and figures) and any papers that are over this word limit will be returned to the authors. Appendices are included in the word limit; however online appendices are not included.

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.

All systematic reviews must be pre-registered.

PREPARING THE SUBMISSION

Contributions must be typed in double spacing. All sheets must be numbered.

Cover Letters

Cover letters are not mandatory; however, they may be supplied at the author's discretion. They should be pasted into the 'Comments' box in Editorial Manager.

Parts of the Manuscript

The manuscript should be submitted in separate files: title page; main text file; figures/tables; supporting information.

Title Page

You may like to use **this template** for your title page. The title page should contain**:**

i. A short informative title containing the major key words. The title should not contain abbreviations (see Wiley's best practice SEO tips);
ii. A short running title of less than 40 characters;
iii. The full names of the authors;
iv. The author's institutional affiliations where the work was conducted, with a footnote for the author's present address if different from where the work was conducted;
v. Abstract;
vi. Keywords;
vii. Practitioner Points;
viii. Acknowledgments.

Authorship

Please refer to the journal's Authorship policy in the Editorial Policies and Ethical Considerations section for details on author listing eligibility. 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.

Abstract

Please provide a structured abstract of up to 250 words under the headings: Objectives, Methods, Results, Conclusions. Articles which report original scientific research should also include a heading 'Design' before 'Methods'. The 'Methods' section for systematic reviews and theoretical papers should include, as a minimum, a description of the methods the author(s) used to access the literature they drew upon. That is, the abstract should summarize the databases that were consulted and the search terms that were used.

Keywords

Please provide appropriate keywords.

Practitioner Points

All articles must include Practitioner Points – these are 2-4 bullet points, following the abstract, with the heading 'Practitioner Points'. These should briefly and clearly outline the relevance of your research to professional practice. (Please include the 'Practitioner Points' in your main document but do not submit them to Editorial Manager with your abstract.)

Acknowledgments

Contributions from anyone who does not meet the criteria for authorship should be listed, with permission from the contributor, in an Acknowledgments section. Financial and material support should also be mentioned. Thanks to anonymous reviewers are not appropriate.

Main Text File

As papers are blind peer reviewed, the main text file should not include any information that might identify the authors.

The main text file should be presented in the following order:

i. Title
ii. Main text
iii. References
iv. Tables and figures (each complete with title and footnotes)
v. Appendices (if relevant)

Supporting information should be supplied as separate files. Tables and figures can be included at the end of the main document or attached as separate files but they must be mentioned in the text.

- As papers are double-blind peer reviewed, the main text file should not include any information that might identify the authors. Please do not mention the authors' names or affiliations and always refer to any previous work in the third person.
- The journal uses British/US spelling; however, authors may submit using either option, as spelling of accepted papers is converted during the production process.

References

References should be prepared according to the *Publication Manual of the American Psychological Association* (6th edition). This means in text citations should follow the author-date method whereby the author's last name and the year of publication for the source should appear in the text, for example, (Jones, 1998). The complete reference list should appear alphabetically by name at the end of the paper. Please note that for journal articles, issue numbers are not included unless each issue in the volume begins with page 1, and a DOI should be provided for all references where available.

Tables

Tables should be self-contained and complement, not duplicate, information contained in the text. They should be supplied as editable files, not pasted as images. Legends should be concise but comprehensive – the table, legend, and footnotes must be understandable without reference to the text. All abbreviations must be defined in footnotes. Footnote symbols: †, ‡, §, ¶, should be used (in that order) and *, **, *** should be reserved for P-values. Statistical measures such as SD or SEM should be identified in the headings.

Figures

Although authors are encouraged to send the highest-quality figures possible, for peer-review purposes, a wide variety of formats, sizes, and resolutions are accepted.

Click here for the basic figure requirements for figures submitted with manuscripts for initial peer review, as well as the more detailed post-acceptance figure requirements.

Legends should be concise but comprehensive – the figure and its legend must be understandable without reference to the text. Include definitions of any symbols used and define/explain all abbreviations and units of measurement.

Colour figures. Figures submitted in colour may be reproduced in colour online free of charge. Please note, however, that it is preferable that line figures (e.g. graphs and charts) are supplied in black and white so that they are legible if printed by a reader in black and white. If an author would prefer to have figures printed in colour in hard copies of the journal, a fee will be charged by the Publisher.

Supporting Information

Supporting information is information that is not essential to the article, but provides greater depth and background. It is hosted online and appears without editing or typesetting. It may include tables, figures, videos, datasets, etc.

Click here for Wiley's FAQs on supporting information.

Note: if data, scripts, or other artefacts used to generate the analyses presented in the paper are available via a publicly available data repository, authors should include a reference to the location of the material within their paper.

General Style Points

For guidelines on editorial style, please consult the **APA Publication Manual** published by the American Psychological Association. The following points provide general advice on formatting and style.

- **Language:** Authors must avoid the use of sexist or any other discriminatory language.
- **Abbreviations:** In general, terms should not be abbreviated unless they are used repeatedly and the abbreviation is helpful to the reader. Initially, use the word in full, followed by the abbreviation in parentheses. Thereafter use the abbreviation only.
- **Units of measurement:** Measurements should be given in SI or SI-derived units. Visit the Bureau International des Poids et Mesures (BIPM) website for more information about SI units.
- **Effect size:** In normal circumstances, effect size should be incorporated.
- **Numbers:** numbers under 10 are spelt out, except for: measurements with a unit (8mmol/l); age (6 weeks old), or lists with other numbers (11 dogs, 9 cats, 4 gerbils).

Appendix B – Guidelines for submission to Children and Youth Service Review (the target journal for the empirical study)*

Children and Youth Services Review is an interdisciplinary forum for critical scholarship regarding service programs for children and youth. The journal will publish full-length articles, current research and policy notes, and book reviews. The Journal's audience includes: Social Workers; Sociologists; Educators; and Psychologists.

References

There are no strict requirements on reference formatting at submission. References can be in any style or format as long as the style is consistent. Where applicable, author(s) name(s), journal title/ book title, chapter title/article title, year of publication, volume number/book chapter and the article number or pagination must be present. Use of DOI is highly encouraged. The reference style used by the journal will be applied to the accepted article by Elsevier at the proof stage. Note that missing data will be highlighted at proof stage for the author to correct.

Formatting requirements

There are no strict formatting requirements but all manuscripts must contain the essential elements needed to convey your manuscript, for example Abstract, Keywords, Introduction, Materials and Methods, Results, Conclusions, Artwork and Tables with Captions.
If your article includes any Videos and/or other Supplementary material, this should be included in your initial submission for peer review purposes.

Divide the article into clearly defined sections.

Figures and tables embedded in text

Please ensure the figures and the tables included in the single file are placed next to the relevant text in the manuscript, rather than at the bottom or the top of the file. The corresponding caption should be placed directly below the figure or table.

Article structure

Subdivision - numbered sections

Divide your article into clearly defined and numbered sections. Subsections should be numbered 1.1 (then 1.1.1, 1.1.2, ...), 1.2, etc. (the abstract is not included in section numbering). Use this numbering also for internal cross-referencing: do not just refer to 'the text'. Any subsection may be given a brief heading. Each heading should appear on its own separate line.

Introduction

State the objectives of the work and provide an adequate background, avoiding a detailed literature survey or a summary of the results.

Material and methods

Provide sufficient details to allow the work to be reproduced by an independent researcher. Methods that are already published should be summarized, and indicated by a reference. If quoting directly from a previously published method, use quotation marks and also cite the source. Any modifications to existing methods should also be described.

Theory/calculation

A Theory section should extend, not repeat, the background to the article already dealt with in the Introduction and lay the foundation for further work. In contrast, a Calculation section represents a practical development from a theoretical basis.

Results

Results should be clear and concise.

Discussion

This should explore the significance of the results of the work, not repeat them. A combined Results and Discussion section is often appropriate. Avoid extensive citations and discussion of published literature.

Conclusions

The main conclusions of the study may be presented in a short Conclusions section, which may stand alone or form a subsection of a Discussion or Results and Discussion section.

Appendices

If there is more than one appendix, they should be identified as A, B, etc. Formulae and equations in appendices should be given separate numbering: Eq. (A.1), Eq. (A.2), etc.; in a subsequent appendix, Eq. (B.1) and so on. Similarly for tables and figures: Table A.1; Fig. A.1, etc.

Appendix C – Statement of epistemological position*

A ‘critical-realist epistemological stance was adopted by the researcher. Critical realism emerged in the 1970s as an alternative to positivism and constructionism (Denzin & Lincoln, 2011). Critical realism attempts to occupy a middle position between positivism, which has been critiqued for reducing ‘reality’ to what can be empirically known, and constructionism which has been critiques for reducing ‘reality’ to a construct entirely formed through and within human discourse and knowledge (Fletcher, 2017). In contrast, critical realism does not deny the existence of a real social world we can attempt to access through science, but also recognises that some knowledge can be closer to this reality than other knowledge (Danermark *et al*. 2002).

Through the literature review and research report, a stance is taken that the psychometric properties of measures and the constructs they attempt to measure are realities that can be measured and accessed. However it is acknowledged that these measurements are understood through human interpretation (Fletcher, 2017), which impacts, for example, what can be viewed as a ‘good’ measure or as ‘adequate’ evidence of a psychometric property.

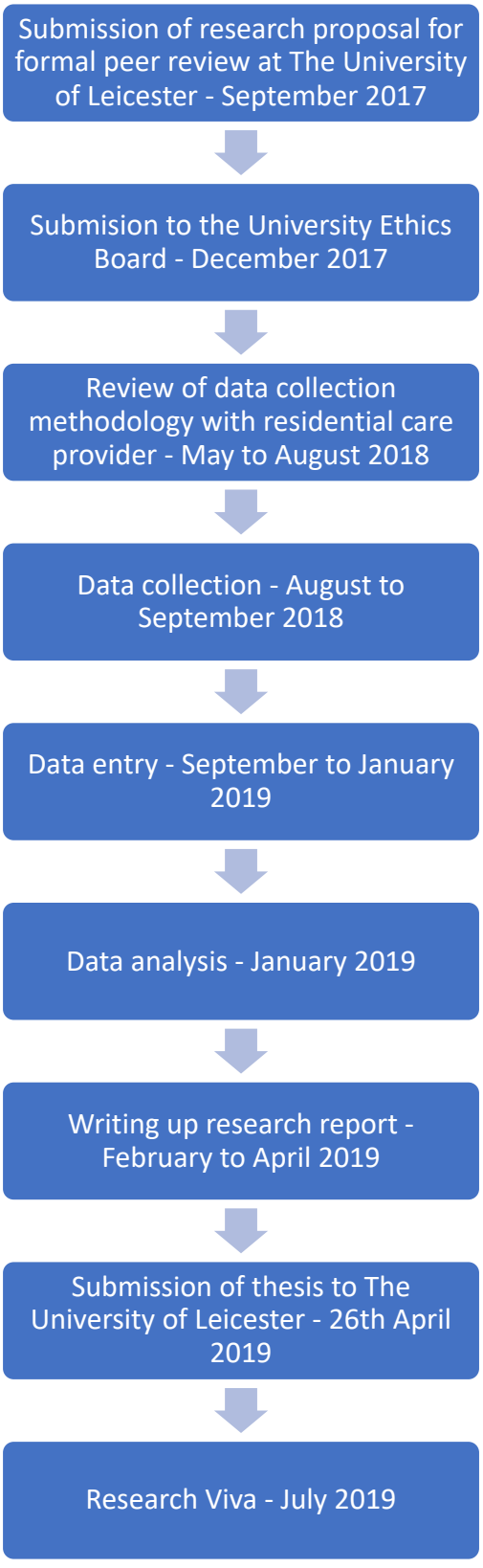
References

Danermark, B., Ekström, M., Jakobsen, L., & Karlsson, J. C. (2002). *Explaining society: An introduction to critical realism in the social sciences*. London: Routledge.

Denzin, N. K. & Lincoln, Y. S. (2011). *The Sage handbook of qualitative research*. Thousand Oaks, CA: Sage

Fletcher, A. J. (2017). Applying critical realism in qualitative research: methodology meets method. *International Journal of Social Research Methodology*, *20*(2), 181-194.

Appendix D – Chronology of research process*


Submission of research proposal for formal peer review at The University of Leicester - September 2017
Submision to the University Ethics Board - December 2017
Review of data collection methodology with residential care provider - May to August 2018
Data collection - August to September 2018
Data entry - September to January 2019
Data analysis - January 2019
Writing up research report - February to April 2019
Submission of thesis to The University of Leicester - 26th April 2019
Research Viva - July 2019

Appendix E – Coursework handbook Appendix D*

	Checked in Executive Summary/Abstract/ Overview (if included in assignment)	**Checked in main text**	**Checked in appendices**
Pseudonym or false initials used	✓	✓	✓
Reference to pseudonym/false initials as a footnote	✓	✓	✓
Removed any reference to names of Trusts/hospitals/clinics/services (including letterhead if including letters in appendices)	✓	✓	✓
Removed any reference to names/specific dates of birth/specific date of clinical appointments/addresses/ location of client(s), participant(s), relatives, caregivers, and supervisor(s). [For research thesis – supervisors can be named in the research thesis "acknowledgements" section]	✓	✓	✓
Removed/altered references to client(s) jobs/professions/nationality where this may potentially identify them. [For research thesis – removed potential for an individual research participant to be identifiable (e.g., by a colleague of the participant who might read the thesis on the internet and be able to identify a participant using a combination of the participants specific job title, role, age, and gender)]	✓	✓	✓
Removed any information that may identify the trainee (consult with course staff if this will detract from the points the trainee is making)	✓	✓	✓
No Tippex or other method has been used to obliterate the original text – unless the paper is subsequently photocopied and the trainee has ensured that the obliterated text cannot be read	✓	✓	✓
The "find and replace" function in word processing has been used to check the assignment for use of client(s) names/other confidential information	✓	✓	✓

Appendix F – Justification for database selection

PsycInfo is a comprehensive database for psychological articles; it was selected due to there being a high likelihood of it successfully retrieving articles related to measures used psychologically informed clinical practice and research. The Medline database is an extensive source of literature in the biomedical field and was selected to capture any MHMs being used in primary physical health care (e.g. general practitioners) and in the field of psychiatry. Web of Science is a large multidisciplinary database and was chosen to capture any research outside of the fields covered by PsycInfo and Medline.

Appendix G – Quality appraisal tool

There is much debate within the literature concerning what defines a measure with 'good' or 'adequate' psychometric properties. Rosenkoetter and Tate (2018) conducted a literature review concerning this debate and identified six quality appraisal tools suitable for appraising psychometric studies. Of these, only the 'Terwee tool' (Terwee *et al*., 2007), specifies numerical standards for assessing the statistical outcomes of psychometric studies. On review of the Terwee tool, the present author recognised that while numerical standards had been provided by Terwee *et al*. (2007), the rationales provided for the selection of many of these standards were thin. Another appraisal tool of interest was the 'Andresen tool' (Andresen, 2000). This tool was adapted by Tsang and Wong (2012) for use in their systematic review of instruments designed to measure adolescents' well-being. Whilst some attempt was made to provide numerical standards, this was not consistent throughout the tool, and little justification was provided when numerical standards were given. The appraisal tool created for use in this study is an amalgamation of the domains considered by the aforementioned tools and the numerical standards from the Terwee tool where robust. Where necessary, additional research was undertaken to ascertain appropriate numerical standards for the domains.

References

Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of physical medicine and rehabilitation*, *81*, 15-20.

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), 103-118.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., *et al*. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*, *60*(1), 34-42.

Tsang, K.L.V., Wong, P. Y. H. & Lo, S.K. (2012). Assessing psychosocial well-being of adolescents: A systematic review of measuring instruments. *Child: care, health and development*, *38*(5), 629-646.

Appendix H – Intra Class Correlation model, definition and type

Model selection concerns whether a one-way random-effects model (where each subject is rated by a different set of randomly chosen raters), two-way random-effects model (for use when raters are randomly selected from a larger pool with similar characteristics, with the aim of generalising results to others with the same characteristics) or two-way mixed-effects model (for use when the selected raters are the only raters of interest with no intention to generalise to others) is most appropriate. Type selection concerns whether the researcher is interested in the reliability of the mean ratings of multiple raters (in which case the 'mean of *k* raters' type should be used) or the reliability of a single rater (in which case the 'single rater' type should be selected). Definition selection concerns whether the researcher views 'absolute agreement' or 'consistency' between raters to be more desirable.

Appendix I – Intra Class Correlation as a measure of inter-rater reliability

Pearson's correlation has been historically used to assess inter-rater reliability, however this form of analysis only provides information pertaining to the consistency of the relationship between the two ratings, as opposed to the absolute agreement between the raters. A perfect correlation could thus be achieved if one rater's scores systematically differed from another by a consistent amount, despite there being no absolute agreement between the raters. ICCs however incorporate the magnitude of any disagreement between the raters in order to compute the inter-rater reliability estimates, with larger magnitude disagreements resulting in lower ICCs.

Shrout and Fleiss (1979) propose four factors central to determining what variant of ICC should be used.

1. **Is a one-way or a two-way model most appropriate?**

   In this instance a one-way model was most appropriate because the design was not fully crossed (i.e. all subjects weren't rated by the same set of coders because the required n exceeded the number of children in one care home).

2. **Would good inter-rater reliability be characterised by absolute agreement or consistency in the rating?**

   It was important to consider the absolute agreement of ratings rather than consistency as in practice it is important that two raters are able to rate similarly to one another rather than simply in a consistent manner.

3. **Is ICC intended to qualify the reliability of the ratings based on averages of ratings provided by more than one person or based on the ratings of a single coder?**

   The single measures ICC was appropriate in this instance as it in practice single ratings on the BERRI are taken into account, rather than an average of several ratings.

4. **Are the coders selected for the study considered to be random or fixed effects?**

   Shrout and Fleiss (1979) suggest that if the raters in a study are selected from a larger population and their ratings are intended to generalize to that population, a random effects model should be used. A random effects model was employed as coders were selected from a wider population of care-staff and their scores.

Appendix J – The BERRI*

BERRI: A checklist to explore Behaviour, Emotional wellbeing, Relationships, Risk and Indicators of psychological conditions in children and young people.

Name .....

Date of Birth..... **Ethnic group**..... **Current Date**.....

Address.....
.....

Placement: Residential home ☐ Foster ☐ Kin ☐ With Parents ☐ Adopted ☐

Form completed by:..... Role: .....

Life Event	Tick if occurred in last 6 months
Change of placement	
Change of carer (including one parent leaving household)	
Change in birth family (eg another child born or removed)	
Change in contact arrangements	
Major conflict with (birth) family or peer group	
House move (with same carers)	
Victim or perpetrator of crime with police involvement	
Other (please specify)	

For each item, please mark how often the child/young person shows evidence of each problem or behaviour and how difficult a problem it is when they show that behaviour, according to the following scales.

Frequency Scores:
0 = never
1 = less than once a week
2 = a few times a week
3 = once a day
4 = several times a day

Difficulty Scores:
0 = does not occur
1 = a minor problem, not difficult to manage
2 = a moderate problem, fairly easy to manage
3 = a major problem, fairly challenging to manage
4 = an extreme problem, almost impossible to manage

Grey areas are for administrative use only.

BEHAVIOUR	Frequency	Difficulty	FxD
Physical aggression towards carers or other children			
Truancy or resistance to attending school			
Sleep disturbance/nightmares			
Argumentative/winding others up			
Racism or other prejudice			
Bullying/threatening others			
Inappropriate toileting: wetting/soiling			
Sexual risk to others/inappropriate touch or comments			
Damage to property			
Cheating or stealing			
Public masturbation			
Gender identity issues			
Lying			
TOTAL BEHAVIOUR			

EMOTIONAL WELLBEING	Frequency	Difficulty	FxD
Lacks self-esteem/pride, has a poor self-image			
Low mood/sadness/crying			
Interest in violence/death/gore			
Fictitious illness/ailments or hypochondria			
Irritable/mood swings			
Worries/phobias			
Lethargy (including being up at night and sleeping in day)			
Extreme emotional response/tantrums/anger/rage			
Fear of normal situations/carers			
Always alert for danger signs/agitated/can't settle			
Proud of negative characteristics (e.g. 'devil'/toughness)			
Poor sense of own identity/culture			
TOTAL EMOTIONAL			

With the emotional wellbeing and relationships categories, please use the following scoring:

Frequency Scores:	Difficulty Scores:
0 = never	0 = does not occur
1 = rarely notice this	1 = a minor problem, not difficult to manage
2 = sometimes notice this	2 = a moderate problem, fairly easy to manage
3 = notice this daily	3 = a major problem, fairly challenging to manage
4 = this is true at all times	4 = an extreme problem, almost impossible to manage

RELATIONSHIPS	Frequency	Difficulty	FxD
Does not make and sustain friendships			
Makes indiscriminate, superficial relationships			
Attention seeking/clingy/needy/whingeing			
Need to provoke chaos			
Shy/timid/bossed about/victim of bullies/isolated			
Not able to show full range of feelings			
Cannot express needs appropriately			
Cannot understand other people's thoughts and feelings			
Lacks empathy			
Seems fake or to be playing a role			
Seems to have no guilt			
TOTAL RELATIONSHIPS			

RISK	Frequency	Difficulty	FxD
Running away/absconding			
Placing self at risk of exploitation			
Self harm: biting/scratching/pulling hair/head banging/pica			
Self harm: cutting/tying ligatures/overdosing			
Choosing unsafe peers/environments			
Getting involved in crime			
No cause-effect reasoning/can't predict consequences of actions			
Making unfounded disclosures			
No fear, puts self in danger, recklessness, thrill seeking			
Setting fires			
TOTAL RISK			

PSYCHOLOGICAL INDICATORS	Frequency	Difficulty	FxD
Lacks concentration/distractible/poor attention span			
Impulsiveness and impatience (e.g. interrupting)			
Incoherent speech/makes noises			
Level of understanding poor/learning disability			
Difficulties with motor co-ordination			
Problems with skills of daily life			
Has odd movements such as tics/rocking/flapping			
Poor grip on reality (e.g. bizarre beliefs/sees or hears things)			
Paranoid			
Literal understanding of language (e.g. can't get jokes or lies)			
Repetitive behaviour or rituals (e.g. checking/washing)			
Hyper arousal (always seems ready for fight/flight)			
Flashbacks or intrusive thoughts/images from trauma			
Echolalia (copies back what is said, like a parrot)			
Can't separate facts from fantasy (e.g. tells elaborate stories)			
TOTAL PSYCHOLOGICAL INDICATORS			

Scoring instructions:

Insert a score for the frequency of the behaviour, and where a behaviour is present, score the difficulty that it presents. Remember that the difficulty score should indicate the amount of care and support required, and not how well the setting is able to prevent or compensate for the issue.

Appendix K – SDQ*

Strengths and Difficulties Questionnaire

For each item, please mark the box for Not True, Somewhat True or Certainly True. It would help us if you answered all items as best you can even if you are not absolutely certain or the item seems daft! Please give your answers on the basis of the child's behaviour over the last six months or this school year.

Child's Name ..... Male/Female

Date of Birth.....

	Not True	Somewhat True	Certainly True
Considerate of other people's feelings	☐	☐	☐
Restless, overactive, cannot stay still for long	☐	☐	☐
Often complains of headaches, stomach-aches or sickness	☐	☐	☐
Shares readily with other children (treats, toys, pencils etc.)	☐	☐	☐
Often has temper tantrums or hot tempers	☐	☐	☐
Rather solitary, tends to play alone	☐	☐	☐
Generally obedient, usually does what adults request	☐	☐	☐
Many worries, often seems worried	☐	☐	☐
Helpful if someone is hurt, upset or feeling ill	☐	☐	☐
Constantly fidgeting or squirming	☐	☐	☐
Has at least one good friend	☐	☐	☐
Often fights with other children or bullies them	☐	☐	☐
Often unhappy, down-hearted or tearful	☐	☐	☐
Generally liked by other children	☐	☐	☐
Easily distracted, concentration wanders	☐	☐	☐
Nervous or clingy in new situations, easily loses confidence	☐	☐	☐
Kind to younger children	☐	☐	☐
Often lies or cheats	☐	☐	☐
Picked on or bullied by other children	☐	☐	☐
Often volunteers to help others (parents, teachers, other children)	☐	☐	☐
Thinks things out before acting	☐	☐	☐
Steals from home, school or elsewhere	☐	☐	☐
Gets on better with adults than with other children	☐	☐	☐
Many fears, easily scared	☐	☐	☐
Sees tasks through to the end, good attention span	☐	☐	☐

Signature ..... Date .....

Parent/Teacher/Other (please specify:)

Thank you very much for your help

Appendix L – NRQ*

Please complete 1 table below for **each young person** in your care to detail **how many times** the following incidents have occurred over the past **3 months**. Please make sure to provide each young person's **unique identification** code

1. This young person's **ID code** is: _____
   Their age is: _____
   Their gender is: _____

	Number of times occurred in the last **3 months**
Absconding/going missing	
Self-harm	
Suicidal behaviour	
Assault towards others	
Victim of assault	
Allegations made (unfounded only)	
Substance misuse	
Incidents of child sexual exploitation	
Setting fires	

Appendix M– Ethical Considerations

The identity of the young people for whom the BERRIs were completed were unknown to the chief investigator. Each child was assigned a unique ID code by the residential care provider to facilitate matching of questionnaires completed by different raters concerning the same young person; the chief investigator was blind to this process.

Care-staff participating in the inter-rater reliability and convergent/divergent validity aspects of the study were provided with a participant information sheet explaining the purpose of the study and outlining that their participation was voluntary. Care-staff were reassured that the measure, rather than their ability as raters, was being assessed and the clinical implications of the study were outlined. Participating care staff were offered the opportunity to enter into a prize draw to win one of three £25 amazon vouchers.

The study was co-designed with managers from the residential care provider to ensure that engagement in the study did not detract from the quality of care provided to the young people. It was agreed that home managers would determine the most appropriate time for completion of the questionnaires.

Feedback will be given to the participating staff members at the residential care homes. This feedback will be given via email or verbal presentation.

Permission was sought to use the SDQ

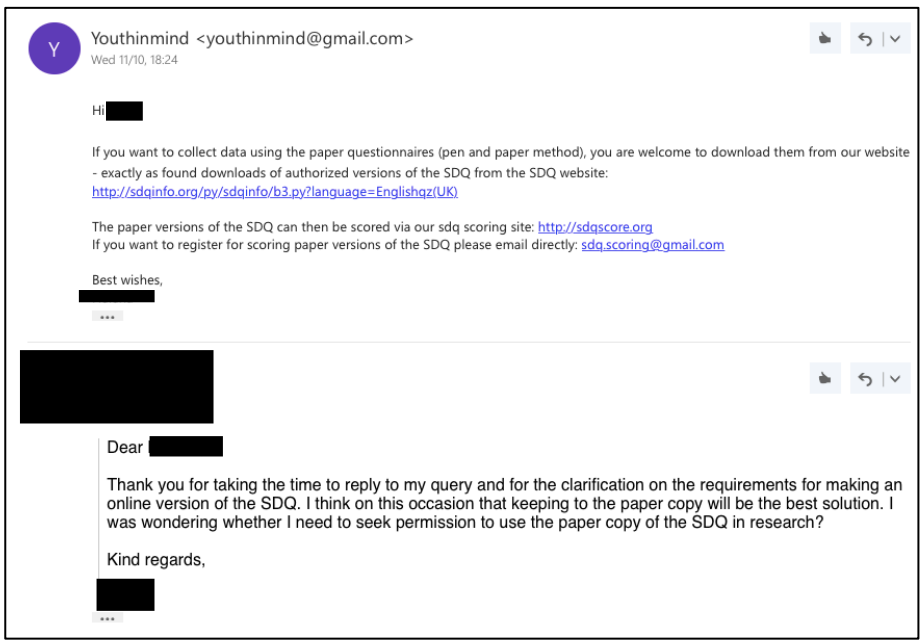


Figure 1: Permission for use of SDQ

Appendix N– Ethical approval confirmation*

The study was granted ethical approval by The University of Leicester first on 13th December 2017 (Figure 2) and then again after an amendment on 30th May 2018 (Figure 3).

University Ethics Sub-Committee for Psychology

13/12/2017

Ethics Reference: 13155-[redacted]-neuroscience,psychologyandbehaviour

TO:
Name of Researcher Applicant: [redacted]
Department: Psychology
Research Project Title: The validity and reliability of the BERRI for use with Looked After Children in residential care

Dear [redacted]

RE: Ethics review of Research Study application

The University Ethics Sub-Committee for Psychology has reviewed and discussed the above application.

1. Ethical opinion

The Sub-Committee grants ethical approval to the above research project on the basis described in the application form and supporting documentation, subject to the conditions specified below.

2. Summary of ethics review discussion

The Committee noted the following issues:
OK

3. General conditions of the ethical approval

The ethics approval is subject to the following general conditions being met prior to the start of the project:

As the Principal Investigator, you are expected to deliver the research project in accordance with the University's policies and procedures, which includes the University's Research Code of Conduct and the University's Research Ethics Policy.

If relevant, management permission or approval (gate keeper role) must be obtained from host organisation prior to the start of the study at the site concerned.

4. Reporting requirements after ethical approval

You are expected to notify the Sub-Committee about:
- Significant amendments to the project
- Serious breaches of the protocol
- Annual progress reports
- Notifying the end of the study

5. Use of application information

Details from your ethics application will be stored on the University Ethics Online System. With your permission, the Sub-Committee may wish to use parts of the application in an anonymised format for training or sharing best practice. Please let me know if you do not want the application details to be used in this manner.

Best wishes for the success of this research project.

Yours sincerely,

Prof. Panos Vostanis
Chair

Figure 2: Initial ethical approval letter

University Ethics Sub-Committee for Psychology

13/12/2017

Ethics Reference: 13155-[redacted]-neuroscience.psychologyandbehaviour

TO:
Name of Researcher Applicant: [redacted]
Department: Psychology
Research Project Title: The validity and reliability of the BERRI for use with Looked After Children in residential care

Dear [redacted]

RE: Ethics review of Research Study application

The University Ethics Sub-Committee for Psychology has reviewed and discussed the above application.

1. Ethical opinion

The Sub-Committee grants ethical approval to the above research project on the basis described in the application form and supporting documentation, subject to the conditions specified below.

2. Summary of ethics review discussion

The Committee noted the following issues:
The issue of ownership of the dataset has been clarified in writing.

3. General conditions of the ethical approval

The ethics approval is subject to the following general conditions being met prior to the start of the project:

As the Principal Investigator, you are expected to deliver the research project in accordance with the University's policies and procedures, which includes the University's Research Code of Conduct and the University's Research Ethics Policy.

If relevant, management permission or approval (gate keeper role) must be obtained from host organisation prior to the start of the study at the site concerned.

4. Reporting requirements after ethical approval

You are expected to notify the Sub-Committee about:

- Significant amendments to the project
- Serious breaches of the protocol
- Annual progress reports
- Notifying the end of the study

5. Use of application information

Details from your ethics application will be stored on the University Ethics Online System. With your permission, the Sub-Committee may wish to use parts of the application in an anonymised format for training or sharing best practice. Please let me know if you do not want the application details to be used in this manner.

Best wishes for the success of this research project.

Yours sincerely,

Prof. Panos Vostanis
Chair

Figure 3: New ethical approval letter

Appendix O – Participant information sheet (care staff)*

We would like to invite you to take part in a study about the BERRI. You have been invited to take part in this study because you use the BERRI on a regular basis. This study aims to assess the reliability and validity of the BERRI as an outcome measure for use with Looked After Children in residential care. The study is being undertaken by [redacted] and forms part of her doctoral level training in Clinical Psychology at the University of Leicester.

We would like you to complete a BERRI and an SDQ, for each young person in your care. Please **do not** write the names of the young people anywhere on the questionnaires. Please use their **unique identification** code instead of their name. A list of unique identification codes will be provided by your home manager We are doing this to measure how good the BERRI is, **we are in no way assessing your ability as a rater or trying to catch you out! Other people will not be made aware of your ratings. Also, please do not discuss your ratings with other staff members; this is really important for the study.**

We anticipate that the questionnaires will take no more than 10 minutes to complete per child. We hope that the results of this study, will help us to consider what changes we might make to the BERRI in order that it is of the most help to you and the young people you care for. We do not anticipate that taking part in this study will cause you any discomfort or disadvantage. Should you choose to participate in the study, it is really important that you **complete the questionnaire pack within 1 week of the date given by your home manager**. As a way of saying thank you for completing the questionnaires, for each full set of questionnaires you complete **within one week of the date given by your home manager** you have the option to enter into a prize draw to win a £25 amazon voucher. There are three £25 vouchers available. If you would like to be entered into the prize draw, please give your email address when prompted.

Your participation in this study is voluntary and declining to take part will have no adverse consequences for you.

If you do decide to take part in the study, you have the right to withdraw at any time until the responses from the questionnaires have been analysed, after which point it will not be possible to remove your data. In order that we can identify your data should you choose to withdraw from the study, we will ask you to create a unique identification code. If you decide to withdraw from the study, please contact [redacted] using the contact information below and provide your identification code.

The findings of the research project this survey is contributing to will be written into a thesis and submitted to the University of Leicester in partial fulfilment of the requirements of the doctoral level training in Clinical Psychology. It is also possible that the findings of the research project may be submitted for publication in an academic journal.

Completed questionnaires will be stored by the University of Leicester for a period of five years following the completion of the study, in accordance with the Data Protection Act.

Should you have any questions regarding this survey, please contact [redacted] (Trainee Clinical Psychologist) at [redacted]@leicester.ac.uk, [redacted] (Clinical

Psychologist and project supervisor) at [redacted]@leicester.ac.uk or [redacted]
(Clinical Psychologist and field supervisor at [redacted]@gmail.com.

- I confirm that I have read and understand the above information about the survey
- I am aware that I can contact [redacted] or [redacted] if I have any questions about the survey
- I understand that all personal information will remain confidential
- I understand that data gathered in this study will be stored anonymously and securely
- I understand that my participation is voluntary and that I am free to withdraw at any time without giving a reason.
- I agree to take part in this study

If you would like to be entered into a prize draw to win one of three £25 amazon vouchers, please write your email address below

_____

In order that we can find your data should you wish to withdraw from the study, please create a unique 5 character code. The first part of the code is the **day** of your birthday, for example if your birthday is 22nd August, the day will be **22.** The second part of the code is the **first three letters of your mother's maiden name**. For example if her maiden name is Smith, the answer would be **SMI**

My unique 5 character identification code is: _____

Appendix P - Participant information sheet (home managers)*

Dear

Thank you for agreeing to be one of the chosen [redacted] homes to participate in the BERRI validation study. The study aims to assess the reliability and validity of the BERRI as an outcome measure for use with Looked After Children in residential care. This research is being undertaken by [redacted] and forms part of her doctoral level training in Clinical Psychology at the University of Leicester.

We would like to ask each member of your care staff to complete one of the enclosed questionnaire packs for every child currently residing in the home. We anticipate that each pack will take no more than 10 minutes to complete. **It is really important that the packs are complete within 1 week of the date of this letter**.

In order to support your care staff in completing the questionnaire packs, we would be most grateful if you could complete the tables below for each child in your home. It is really important that the information remains anonymous, therefore we would please ask that you **use each young person's unique identification code rather than their name**. A list of unique identification codes will be included in this pack

As a way of saying thank you for participating in the study, you and your care staff have the option to enter into a prize draw to win a £25 amazon voucher. There are three £25 vouchers available. If you would like to be entered into the prize draw, please give your email address when prompted.

The findings of the research project this survey is contributing to will be written into a thesis and submitted to the University of Leicester in partial fulfilment of the requirements of the doctoral level training in Clinical Psychology. It is also possible that the findings of the research project may be submitted for publication in an academic journal.

Completed questionnaires will be stored by the University of Leicester for a period of five years following the completion of the study, in accordance with the Data Protection Act.

Should you have any questions regarding this survey, please contact [redacted] (Trainee Clinical Psychologist) at [redacted]@leicester.ac.uk, [redacted] (Clinical Psychologist and project supervisor) at [redacted]@leicester.ac.uk or [redacted] (Clinical Psychologist and field supervisor at [redacted]@gmail.com.

Once all of your care staff have completed their questionnaire packs, please post them along with the completed tables below to

[redacted]

Many thanks in advance,

[redacted]

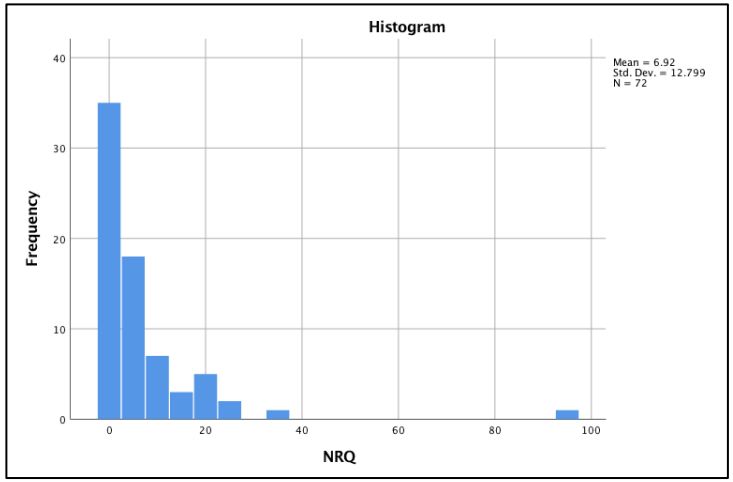
Appendix Q– Bonferroni correction

To account for the calculation of multiple correlations and the consequential increased likelihood of type II errors, the alpha level was adjusted using a Bonferroni correction (Bonferroni, 1936). Alpha (0.5) was divided by the number of unique correlations. For the assessment of the construct validity of the BERRI 'behaviour', 'emotions', 'relationships' and 'risk' scales, seven variables were analysed resulting in 21 (7(7-1)/2) unique correlations. For the BERRI 'indicators' scale, eight variables were analysed resulting in 28 (8(8-1)/2) unique correlations. Alpha (0.5) divided by the number of unique correlations in both instances (when rounded to two decimal places) equalled .002.

References

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3-62.

Appendix R – Distribution plots for construct validity



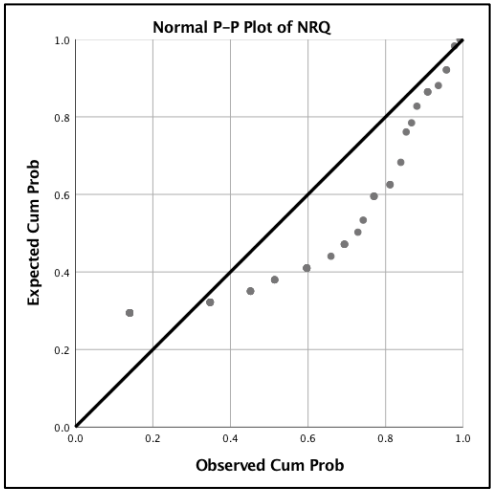
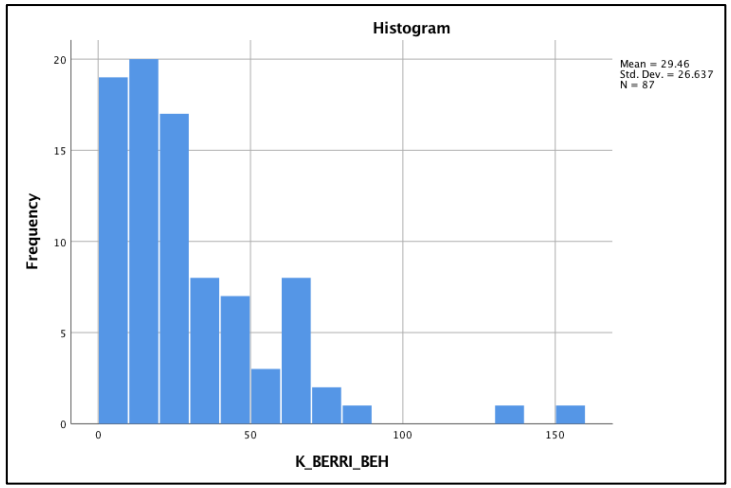


Figure 4: Histogram and PP plot for NRQ



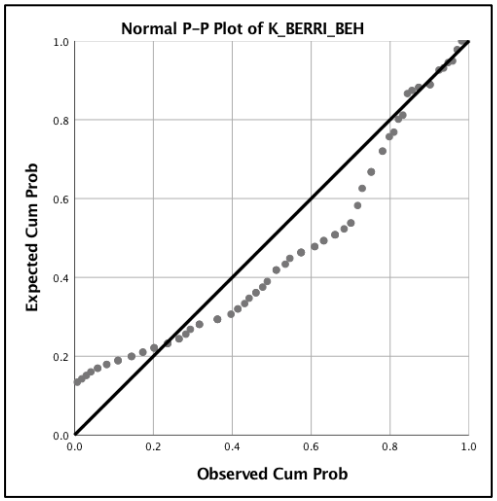
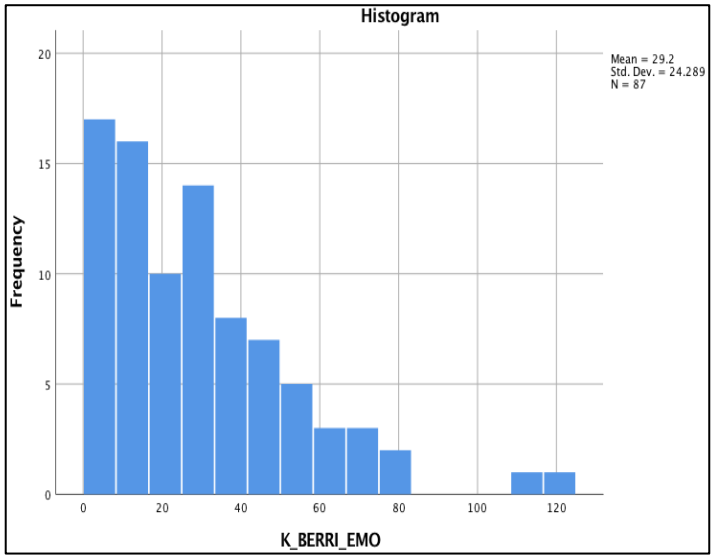


Figure 5: Histogram and PP plot for BERRI 'behaviour' scale



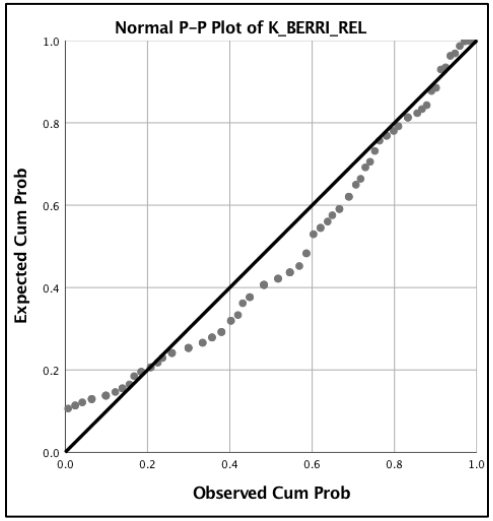
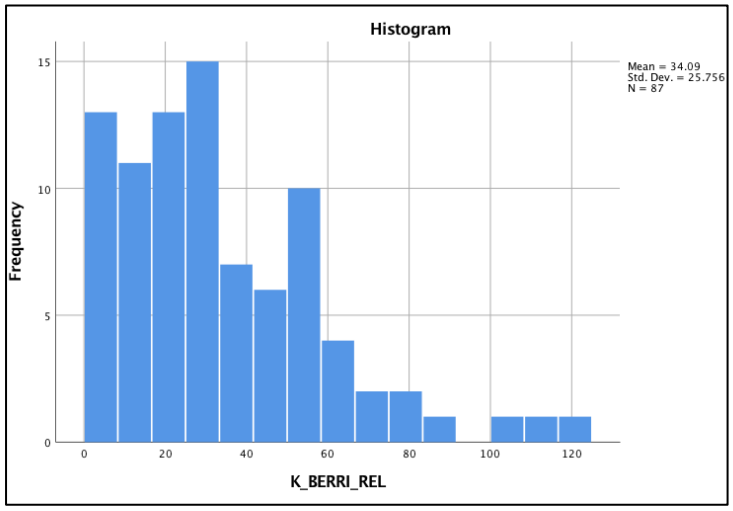


Figure 6: Histogram and PP plot for BERRI 'emotions'' scale



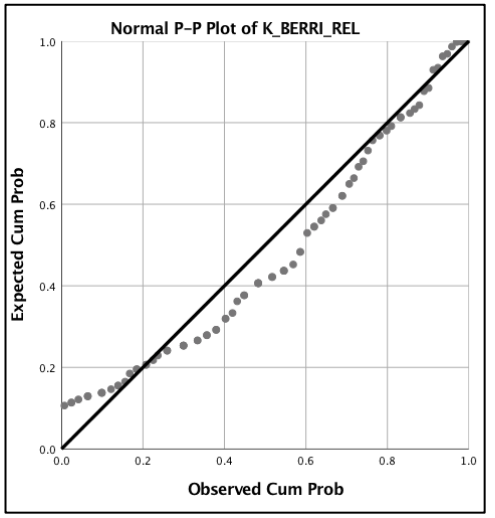
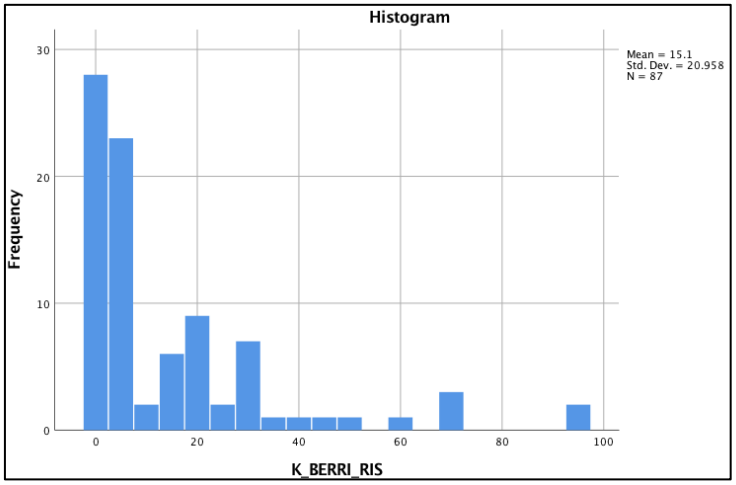


Figure 7: Histogram and PP plot for BERRI 'relationships' scale



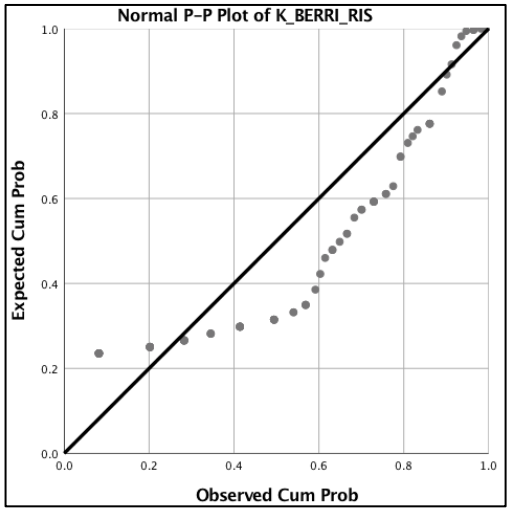
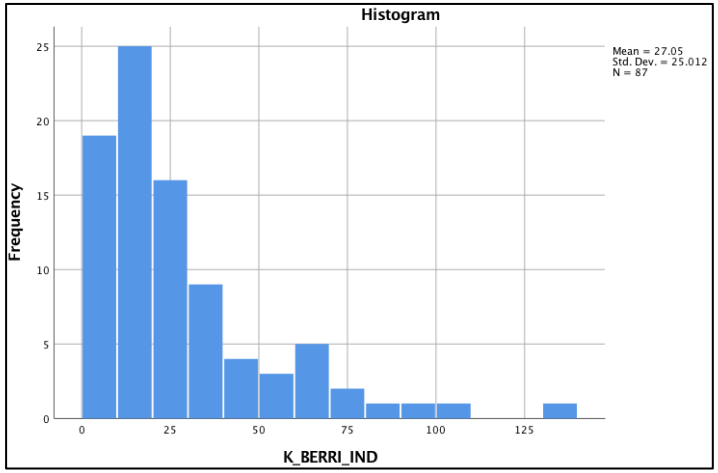


Figure 8: Histogram and PP plot for BERRI risk scale



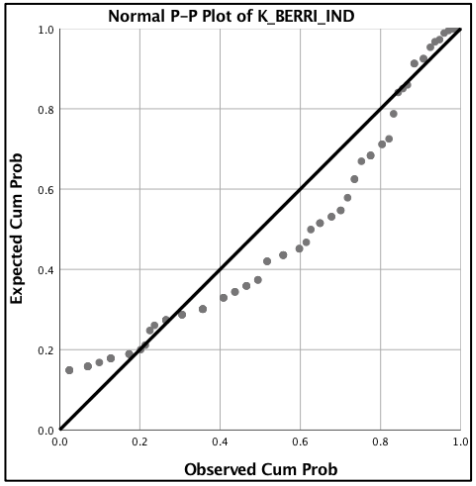
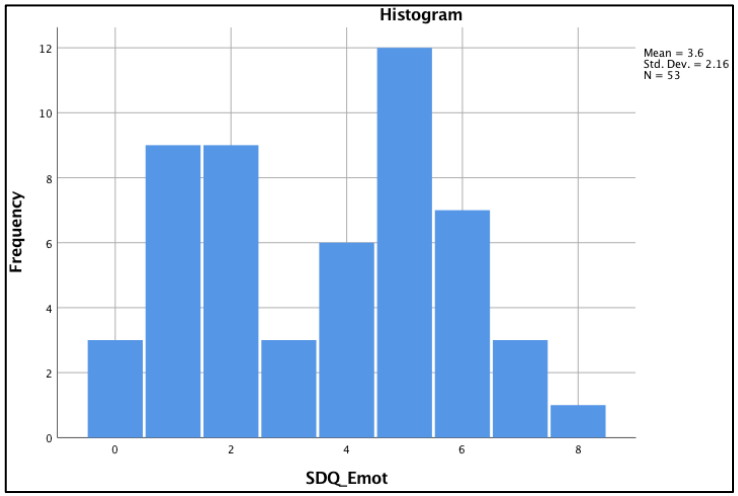


Figure 9: Histogram and PP plot for BERRI 'indicators' scale



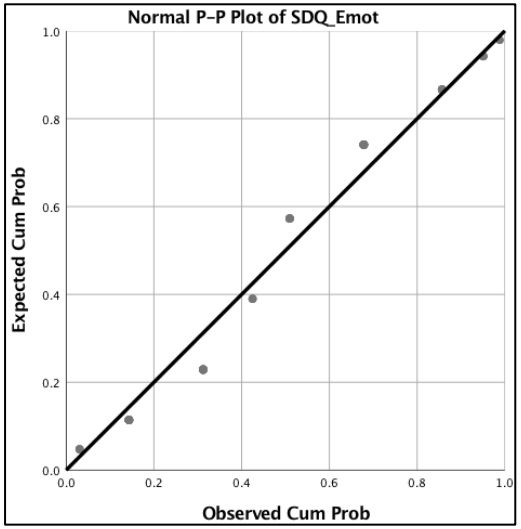
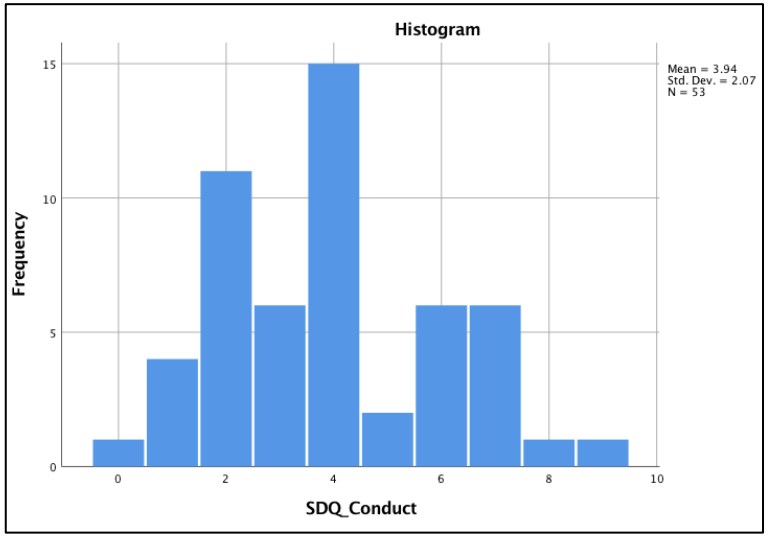


Figure 10: Histogram and PP plot for SDQ 'emotions' scale



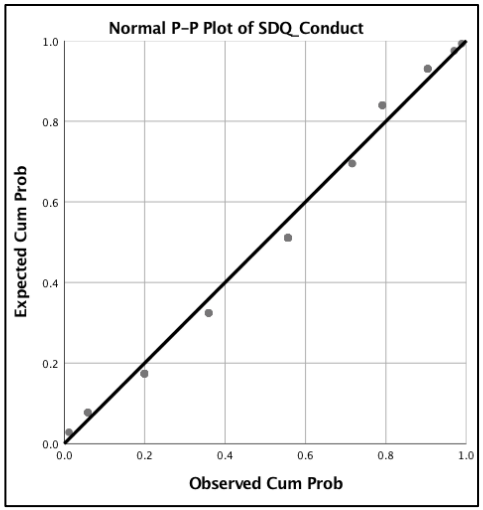
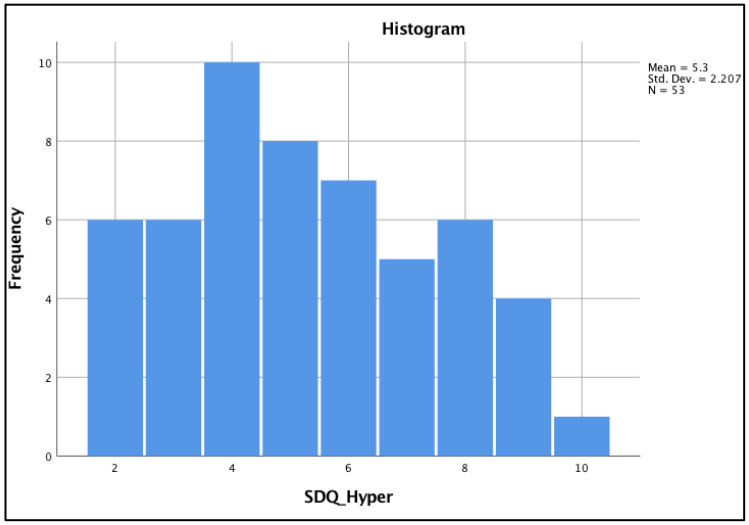


Figure 11: Histogram and PP plot for SDQ 'conduct' scale



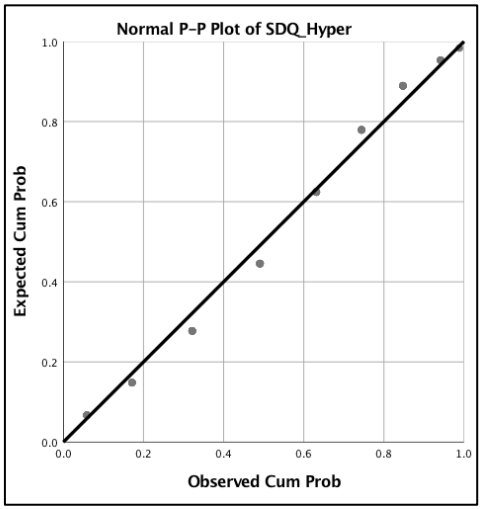
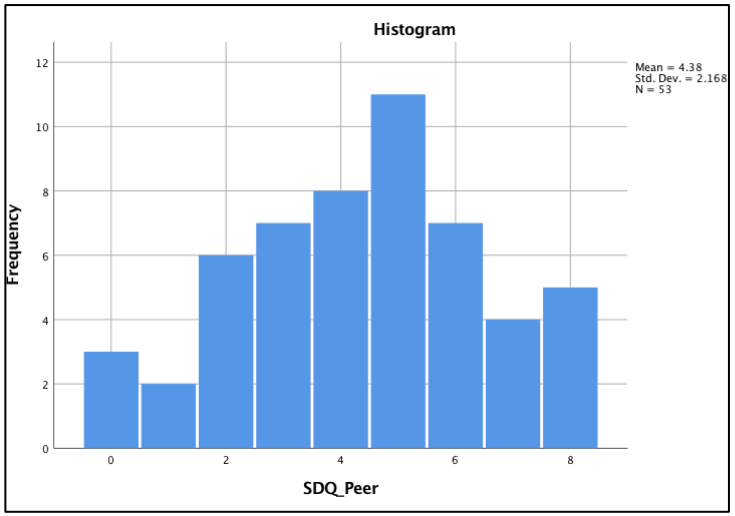


Figure 12: Histogram and PP plot for SDQ hyperactivity scale



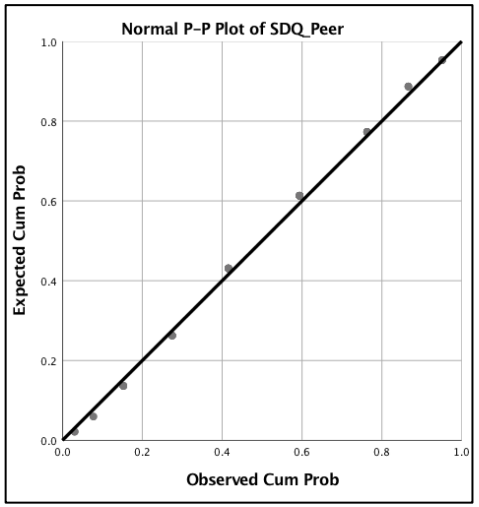
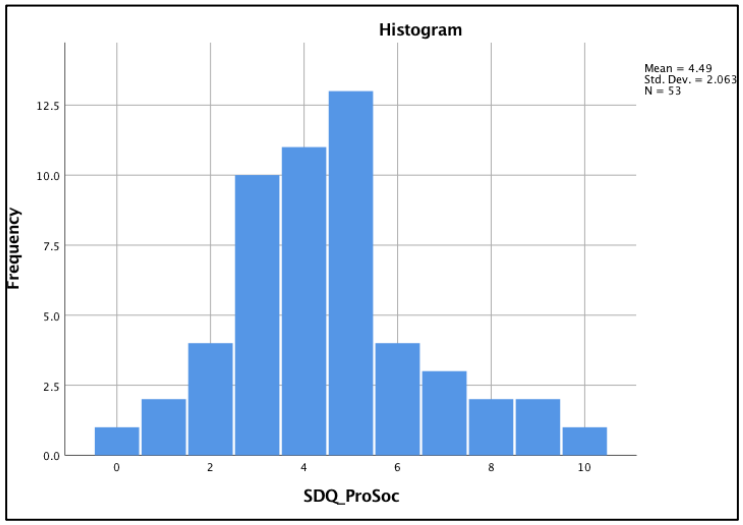


Figure 13: Histogram and PP plot for SDQ 'peer relationships scale



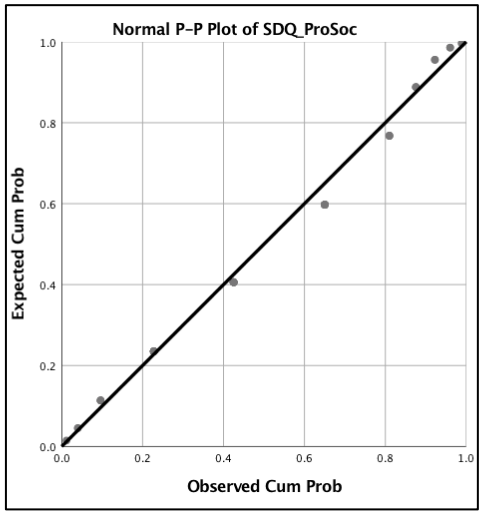
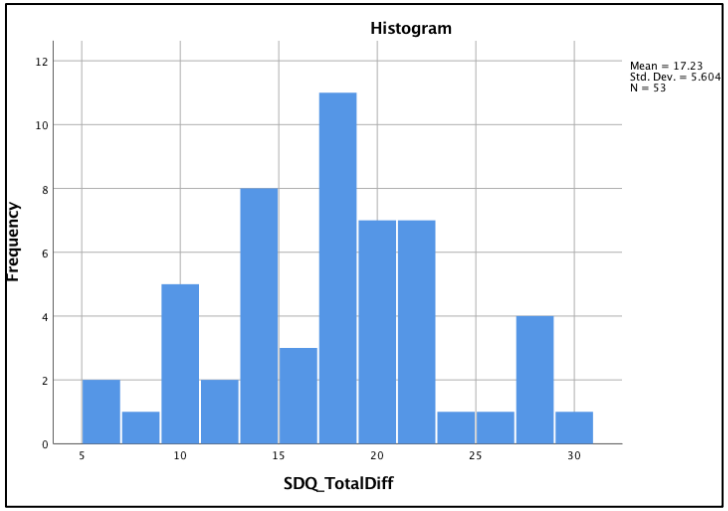


Figure 14: Histogram and PP plot for SDQ 'Pro-social' scale



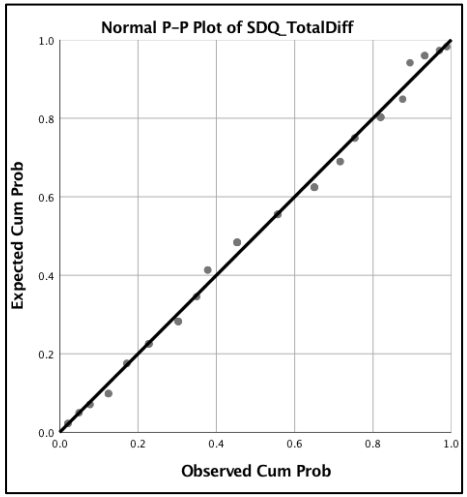


Figure 15: Histogram and PP plot for SDQ total difficulties score

Appendix S – KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.912
Bartlett's Test of Sphericity	Approx. Chi-Square	26188.468
	df	3403
	Sig.	0.000

Figure 16: SPSS output depicting KMO and Bartlett's test

Appendix T – Three component model

Rotated Component Matrix			
	Component		
	1	2	3
Verbal aggression	0.752		
Extreme emotional response/tantrums/anger/rage	0.750		
Need to provoke chaos/winds up others/test their response	0.736		
Physical aggression towards carers or other children	0.733		
Under/over eating, storing of food, self-induced vomiting	0.722		
Argumentative/winding others up	0.701		
Non-compliance/defiance	0.698		
Impulsiveness and impatience (e.g. interrupting)	0.696		
Hyper arousal (always seems ready for fight/flight)	0.657		
Irritable/mood swings	0.646		
Trying to be in control of everyone around them	0.643		
Screaming/shouting/too loud	0.635		
Cannot understand other people's thoughts and feelings	0.614		
Damage to property	0.602		
Lacks empathy	0.590		
Proud of negative characteristics (e.g. 'devil'/toughness)	0.568		
Seems to have no guilt	0.565		
Lacks concentration/distractible/poor attention span	0.557		
Attention seeking/clingy/needy/whingeing	0.512		
Poor judge of personal space/ poor social judgement	0.502	0.436	
No cause-effect reasoning/can't predict consequences of actions	0.498		
Racism or other prejudice	0.461		
Always alert for danger signs/agitated/can't settle	0.455	0.407	
Lack of concern about how others see them	0.452		

Interest in violence/death/gore	0.444		
Incoherent speech/makes noises	0.405		
Struggles with change/has a rigid need for routine	0.403		
Seeking restraint			
Sexual risk to others/inappropriate touch or comments			
Seeking punishment			
Seems fake or to be playing a role			
Spitting			
Lying			
Public Masturbation			
Inappropriate toileting: wetting/soiling			
Lacks self-esteem/pride, has a poor self-image		0.632	
Worries/phobias		0.625	
Low mood/sadness/crying		0.584	
Does not make and sustain friendships		0.559	
Self blame or unrealistic expectations of self		0.545	
Not able to show full range of feelings		0.530	
Poor sense of own identity/culture		0.514	
Shy/timid/bossed about/victim of bullies/isolated		0.510	
Self-critical/can't take praise		0.510	
Poor grip on reality (e.g. bizarre beliefs/sees or hears things)		0.495	
Lack of joy/laughter/emotionally flat		0.478	
Fear of normal situations/carers		0.463	
Flashbacks or intrusive thoughts/images from trauma		0.447	
Not able to 'click' with anyone		0.442	
Obsessions or narrow all-consuming interests		0.413	
Cannot express needs appropriately		0.403	
Problems with skills of daily life		0.402	
Withdrawn/uncommunicative			
Lack of comfort–seeking (e.g. if hurt)			
Makes indiscriminate, superficial, overly close relationships			

Can't separate facts from fantasy (e.g. tells elaborate stories)			
Level of understanding poor/learning disability			
Paranoid			
Making unfounded disclosures			
Self harm: cutting/tying ligatures/overdosing			
Literal understanding of language (e.g. can't get jokes or lies)			
Sleep disturbance/nightmares			
Fictitious illness/ailments or hypochondria			
Lack of self-care/hygiene			
Self harm: biting/scratching/pulling hair/head banging/pica			
Lack of imagination/self-directed play			
Bullying/threatening others			
Echolalia (copies back what is said, like a parrot)			
Has odd movements such as tics/rocking/flapping			
Inappropriate toileting: smearing faeces or urinating in room			
Choosing unsafe peers/environments			0.779
Placing self at risk of exploitation			0.736
Getting involved in crime			0.729
Running away/absconding			0.711
Drug, solvent or alcohol abuse			0.705
Sexually active in a risky way or sex working			0.586
No fear, puts self in danger, recklessness, thrill seeking	0.412		0.572
Cheating or stealing			0.471
Truancy or resistance to attending school			0.436
Lethargy (including being up at night and sleeping in day)			
Setting fires			
Difficulties with speech or understanding of language			
Difficulties with motor co-ordination			

Appendix U – Internal consistency of 'Emotional needs' scale

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
0.827	0.832	11

Figure 17: SPSS output demonstrating reliability statistics for the 'Emotional needs' scale prior to deletion of 'Lethargy'

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Lacks self-esteem/pride, has a poor self-image	23.15	389.818	0.626	0.505	0.800
Self-critical/can't take praise	24.55	410.967	0.604	0.447	0.803
Low mood/sadness/crying	24.28	426.277	0.567	0.392	0.808
Lack of joy/laughter/emotionally flat	25.31	435.976	0.559	0.449	0.810
Worries/phobias	24.30	426.363	0.495	0.314	0.813
Self blame or unrealistic expectations of self	25.04	435.122	0.468	0.288	0.816
Lethargy (including being up at night and sleeping in day)	24.49	436.508	0.334	0.169	0.830
Withdrawn/uncommunicative	24.77	439.053	0.462	0.348	0.816
Poor sense of own identity/culture	24.97	432.762	0.426	0.223	0.820
Lack of comfort–seeking (e.g. if hurt)	25.66	453.473	0.438	0.290	0.819
Not able to show full range of feelings	22.83	391.346	0.568	0.356	0.807

Figure 18: SPSS output depicting item-total statistics for 'Emotional needs' scale, indicative of the need to remove 'lethargy' from the scale

Appendix V– Items excluded from PCA

Table 9: Items excluded from PCA

	Rotated Component Loadings				
Item	Behaviour	Indicators	Risk	Emotional Needs	Relationship (attachment) seeking
Seeking punishment	0.399	0.077	-0.078	0.206	0.089
Seeking restraint	0.392	0.124	-0.092	0.112	0.024
Seems fake or to be playing a role	0.340	0.123	0.189	0.200	0.312
Public masturbation	0.228	0.112	0.136	-0.048	0.026
Lack of imagination/self-directed play	0.184	0.380	-0.005	0.173	0.095
Lack of self-care/hygiene	0.107	0.347	0.133	0.199	0.069
Cannot express needs appropriately	0.318	0.341	0.056	0.319	0.049
Inappropriate toileting: wetting/soiling	0.162	0.271	-0.074	0.000	0.014
Inappropriate toileting: smearing faeces or urinating in room	0.042	0.150	-0.006	0.113	0.124
Truancy or resistance to attending school	0.231	-0.077	0.374	0.298	0.014
Setting fires	0.263	0.039	0.367	0.025	-0.061
Spitting	0.268	0.149	0.324	0.039	-0.176
Shy/timid/bossed about/victim of bullies/isolated	-0.112	0.185	0.036	0.398	0.255
Does not make and sustain friendships	0.258	0.327	-0.010	0.369	0.331
Sleep disturbance/nightmares	0.230	0.117	0.156	0.339	0.077
Paranoid	0.172	0.248	0.082	0.335	-0.006
Not able to ‘click’ with anyone	0.242	0.252	-0.011	0.328	0.206
Flashbacks or intrusive thoughts/images from trauma	0.005	0.170	0.066	0.318	0.276
Bullying/threatening others	0.233	0.175	0.086	0.260	0.072
Lying	0.241	0.121	0.315	0.028	0.350
Sexual risk to others/inappropriate touch or comments	0.300	0.207	0.268	-0.028	0.325