Clinical Psychology Review 110 (2024) 102434

Contents lists available at ScienceDirect

Clinical Psychology Review

journal homepage: www.elsevier.com/locate/clinpsychrev

Review

Measurement tools for behaviours that challenge and behavioural function in people with intellectual disability: A systematic review and meta-analysis of internal consistency, inter-rater reliability, and test-retest reliability

Lauren Shelley [a,*], Chris Jones [b], Effie Pearson [a], Caroline Richards [b,d], Hayley Crawford [c,d], Arianna Paricos [a,1], Courtney Greenhill [a], Alixandra Woodhead [a,2], Joanne Tarver [a], Jane Waite [a,d]

[a] School of Psychology, College of Health and Life Sciences, Aston University, UK
[b] School of Psychology, University of Birmingham, Edgbaston, UK
[c] Mental Health and Wellbeing Unit, Warwick Medical School, University of Warwick, UK
[d] Cerebra Network for Neurodevelopmental Disorders, UK

ARTICLE INFO

Keywords:
Intellectual disability
Behaviours that challenge
Function
Measurement
Measurement properties

ABSTRACT

Behaviours that challenge (BtC) are common in people with intellectual disability (ID) and associated with negative long-term outcomes. Reliable characterisation of BtC and behavioural function is integral to person-centred interventions. This systematic review and meta-analytic study quantitatively synthesised the evidence-base for the internal consistency, inter-rater reliability, and test-retest reliability of measures of BtC and behavioural function in people with ID (PROSPERO: CRD42021239042). Web of Science, Embase, PsycINFO and MEDLINE were searched from inception to March 2024. Retrieved records (n = 3691) were screened independently to identify studies assessing eligible measurement properties in people with ID. Data extracted from 83 studies, across 29 measures, were synthesised in a series of random-effects meta-analyses. Subgroup analyses assessed the influence of methodological quality and study-level characteristics on pooled estimates. COSMIN criteria were used to evaluate the measurement properties of each measure. Pooled estimates ranged across measures: internal consistency (0.41–0.97), inter-rater reliability (0.29–0.93) and test-retest reliability (0.52–0.98). The quantity and quality of evidence varied substantially across measures; evidence was frequently unavailable or limited to a single study. Based on current evidence, candidate measures with the most evidence for internal consistency and reliability are discussed; however, continued assessment of measurement properties in ID populations is a key priority.

1. Introduction

Behaviours that challenge (BtC), such as self-injury and aggression, are frequently reported in people with intellectual disability (ID). Prevalence rates between 10% and 60% are reported; however, rates are shown to vary according to differences in definitions of BtC, methods of assessment, and the population studied (Deb, Thomas, & Bright, 2001; Emerson et al., 2001; Rojahn, Rick-Betancourt, Barnard-Brak, & Moore, 2017; Simo-Pinatella, Mumbardo-Adam, Alomar-Kurz, Sugai, & Simonsen, 2019). BtC persist overtime in ID and ID-associated genetic syndrome populations (Crawford, Karakatsani, Singla, & Oliver, 2019; Davies & Oliver, 2014; Emerson et al., 2001; Taylor, Oliver, & Murphy, 2011; Wilde et al., 2018). Persistence of BtC is concerning, given associations with detrimental long-term outcomes, such as impeded learning and development, increased social exclusion, and use of physical restraint and medication (Cooper et al., 2009; Emerson, 2001; Emerson,

* Corresponding author at: School of Psychology, College of Life and Health Sciences, Aston University, B4 7ET, UK.
E-mail addresses: shellel1@aston.ac.uk (L. Shelley), e.pearson1@aston.ac.uk (E. Pearson), c.r.richards@bham.ac.uk (C. Richards), hayley.crawford@warwick.ac.uk (H. Crawford), arianna.paricos@nottingham.ac.uk (A. Paricos), j.tarver@aston.ac.uk (J. Tarver), j.waite@aston.ac.uk (J. Waite).
[1] Present address: Division of Psychiatry and Applied Psychology, University of Nottingham, Nottingham, UK.
[2] Present address: Black Country Healthcare Foundation NHS Trust, Canalside, Walsall, UK.

https://doi.org/10.1016/j.cpr.2024.102434
Received 7 June 2023; Received in revised form 15 March 2024; Accepted 12 April 2024
Available online 16 April 2024

Einfeld, & Stancliffe, 2011; Holden & Gitlesen, 2006). Furthermore, BtC are associated with decreased caregiver quality of life and increased caregiver stress, fatigue, burnout, and social exclusion (Adams et al., 2018; Hastings, 2002; Lecavalier, Leone, & Wiltz, 2006). The prevalence, persistence, and consequences of BtC highlight the importance of early intervention to improve outcomes for people with ID (Davies & Oliver, 2016; Oliver & Richards, 2015). An essential first step towards intervention is valid and reliable assessment of a person's behavioural presentation.

Precise characterisation of BtC is integral for tailoring evidence-based interventions to a person's behavioural presentation. The UK National Institute for Health and Care Excellence (NICE) guidelines recommend the use of formal assessments to clearly describe BtC (National Collaborating Centre for Mental Health (NCCMH), 2015). In addition, a thorough assessment also includes collating information about why BtC has emerged and how it may be maintained over time (Lloyd & Kennedy, 2014). This includes assessment of person and environmental factors implicated in the presentation of BtC (Limbu, Unwin, & Deb, 2021; NCCMH, 2015). Such information is used by clinicians to develop formulations and interventions based on the operant model, whereby the consequences that follow behaviour might inadvertently increase the likelihood of behaviour reoccurring through operant reinforcement (Beavers, Iwata, & Lerman, 2013; Carr & Durand, 1985; Hanley, 2012; Healy, Brett, & Leader, 2013). In the operant model, reinforced behaviour is described as functional, in that it is repeated because it serves an inadvertent function for a person. Consequently, understanding behavioural function is often a focal part of assessment, along with consideration of gene-environment-behaviour interactions (Davies & Oliver, 2016; Waite et al., 2014).

Several methods of assessment are available to characterise a person's BtC and assess behavioural function. Indirect assessment methods include informant and self-report questionnaire and interview methodology, while direct assessment methods include naturalistic behavioural observation and functional assessment (Floyd, Phaneuf, & Wilczynski, 2005; Lloyd & Kennedy, 2014; Matson & Williams, 2014). Examples of indirect informant-report measures to assess BtC include the Behavior Problems Inventory-01 (Rojahn, Matson, Lott, Esbensen, & Smalls, 2001) and Challenging Behaviour Interview (Oliver et al., 2003) and examples of functional measures are the Questions About Behavioral Function Scale (Matson & Vollmer, 2007) and Motivation Assessment Scale (Durand & Crimmins, 1992). Direct assessment methods are often viewed as the 'gold standard' for assessment, but typically involve considerable time and expense, can be intrusive, and require well-trained observers (Madsen, Peck, & Valdovinos, 2016; Matson & Williams, 2014; Zarcone, Napolitano, & Valdovinos, 2008). Indirect assessment measures can overcome some of these limitations, such as enabling information to be gathered at reduced time and expensive, whilst meeting NICE objectives to precisely characterise the form, frequency, severity, and duration of BtC, and identify possible behavioural functions (Floyd et al., 2005; Madsen et al., 2016). This information can inform interventions that are tailored to a person's behavioural presentation and needs. In addition, informant-report measures can be used to monitor behavioural presentation overtime, including the success of interventions in improving behavioural outcomes, with the information gathered informing decisions about further intervention or service provision (Baker & Daynes, 2010; NCCMH, 2015; Zarcone et al., 2008). The soundness of such decisions is partially dependent on the quality of the informant-report measures being used. Therefore, to ensure information gathered is valid and reliable, it is important to understand the evidence base for the measurement properties of the methods of assessment being selected for use.

To date, several systematic reviews have been conducted to examine the measurement properties of measures to assess BtC and behaviour-related outcomes (Howell, Bradshaw, & Langdon, 2021; McConachie et al., 2015; Reyes-Martín, Simó-Pinatella, & Font-Roura, 2022; Turton, 2015). However, these reviews have typically focused on measures for specific populations, e.g., autistic children under 6 years (McConachie et al., 2015), or the use of measures with specific informants, e.g., special education teachers (Howell et al., 2021). Reyes-Martín et al. (2022) provide a recent synthesis of literature pertaining to the measurement properties of measures to assess BtC in people with intellectual and developmental disabilities. In addition, the authors explored variables related to BtC and evidence of interventions informed by measures characterising BtC. Whilst these reviews have made an important contribution to the literature regarding measures of BtC, to date, no systematic reviews have assessed the measurement properties of measures to assess behavioural function. Additionally, no systematic reviews have included a meta-analytic synthesis of the measurement properties of measures of BtC or behavioural function. Consequently, a systematic review and meta-analytic synthesis of measurement properties is warranted, to collate, synthesise, and appraise the measurement properties of informant-report measures used to assess BtC *and* functions of BtC in people with ID across all age groups. This is important as the choice of high-quality measures among clinicians and researchers is strongly determined by robust measurement properties (Maguire, Davison, McLaughlin, Simms, & Bunting, 2023).

Given the identified gaps in the literature, the current systematic review and meta-analytic study focuses on understanding current evidence for the internal consistency, inter-rater reliability, and test-retest reliability of informant-report measures of BtC and behavioural function in ID populations. The focus on internal consistency, inter-rater reliability and test-retest reliability is an important first step towards evaluating measurement properties and informing future recommendations for the use of measures in both research and clinical practice. However, it's important to note that measurement properties are not static or inherent within a measure; they can be influenced by various factors, such as the interaction between items within a measure and the specific population and context in which the measure is used (Swan et al., 2023). For instance, while multi-informant assessment can be a beneficial approach to gathering information across different contexts, measurement properties may differ due to differences in behavioural presentation and function across various situations, settings, and contexts (e.g., school settings vs home environments) (Alter, Conroy, Mancil, & Haydon, 2008; Chung et al., 2022). Furthermore, informant-related factors, such as the length of time an informant has known a person, time spent together, relationship quality, and experience of working with people with ID, can also influence measurement properties (Nicholson, Konstantinidi, & Furniss, 2006; Shrout, 1998). Consequently, the current review also aims to examine the impact of study methodological quality and study-level characteristics (e.g., level of ID, informant completing the measure, method of administration and recruitment setting) on measure internal consistency, inter-rater reliability and test-retest reliability estimates. This review focuses on measures enabling the assessment of three categories of observable and operationalisable BtC, self-injury, aggression, and destruction, commonly reported in ID populations (Arron, Oliver, Moss, Berg, & Burbidge, 2011; Emerson et al., 2001; Grey, Pollard, McClean, MacAuley, & Hastings, 2010; Rojahn et al., 2001). While several existing measures of BtC include subscales for a wider range of behavioural phenomena, such as hyperactivity and stereotyped and repetitive behaviour, this review does not extend to these subscales.

In summary, this systematic review and meta-analytic study aims to:

1. Conduct a preliminary search to identify standardised informant-report measures used to assess BtC and behavioural function in ID populations, including genetic syndromes associated with ID.
2. Meta-analytically synthesise published evidence of internal consistency, inter-rater reliability, and test-retest reliability for identified measures.
3. Conduct subgroup and meta-regression analyses to explore the impact of methodological bias and study level characteristics on reliability estimates.

4. Formally evaluate quality of evidence for the internal consistency, inter-rater reliability, and test-retest reliability of each identified measure using recommended guidelines.

2. Methods

A systematic review and meta-analysis was undertaken, reported in accordance with PRISMA guidelines (Page et al., 2021) and the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines (Prinsen et al., 2018). Methodology and analysis details were pre-registered in a PROSPERO protocol prior to completion of the review (CRD42021239042). This paper focuses on the internal consistency and inter-rater reliability and test-retest reliability portions of the PROSPERO pre-registration.

2.1. Preliminary search

A preliminary search was conducted to address the first aim, to identify measures of BtC and function that have been used in ID populations and inform measure search terms in the main search. The preliminary search yielded 50 measures (43 measures primarily assessing BtC and 7 measures primarily assessing behavioural function) that were taken forward to this review. Details of the preliminary search strategy and included/excluded measures are provided in Supplementary material 1.

2.2. Search strategy and selection criteria

Searches were conducted to identify papers examining the internal consistency (IC), inter-rater reliability (IRR), and test-retest reliability (TRTR) of the 50 measurement tools when used in ID populations. Electronic searches were conducted on 18 March 2022 and updated 1 March 2024.[3] Four databases were searched with no restriction on year of publication: Web of Science (Core Collection), Ovid PsycINFO, Ovid Embase, and Ovid Medline. Search terms consisted of three components, each containing synonyms, truncated where appropriate and combined using Boolean terms 'OR' and 'AND' (see Supplementary material 2 for the full search strategy for each database):

1. Measurement tools (names and acronyms for the 50 identified tools identified in the preliminary search)
2. ID and ID-associated genetic syndromes
3. Measurement properties
4. 1 AND 2 AND 3

Measurement property search terms were based on a search filter developed by the COSMIN group (Terwee, Jansma, Riphagen, & de Vet, 2009). ID-associated genetic syndrome search terms from the preliminary search were adjusted to include additional syndromes identified. Genetics Home Reference, an expert-reviewed online resource, and GeneReviews were consulted to identify terms for these syndromes.

Returned papers were assessed for inclusion with LS screening 100% (3691) of total papers. At stage one screening, papers were independently screened by review of titles and abstracts using predefined inclusion and exclusion criteria (see Table 1). At stage one, substantial agreement was established between LS and EP on 25% (923) of total papers (Kappa = 0.91). Where there were discrepancies regarding inclusion at stage one, an over-inclusive approach was adopted, and studies included to ensure relevant studies were not missed.

At stage two screening, additional criteria were employed to screen the full texts of papers (see Table 1). Substantial agreement was established between LS and EP (Kappa = 0.86) at stage two screening. To resolve discrepancies at stage two, a third reviewer was consulted, and consensus reached. Following screening, LS and AP completed forwards and backwards searches of the references lists and citations of included papers, resulting in the identification of an additional 3 papers (see Fig. 1).

Table 1
Inclusion and exclusion criteria for stage 1 and stage 2 screening of returned papers.

Stage one screening	
Inclusion criteria	Exclusion criteria
Studies employing an eligible measurement tool	Non-human studies
Studies including individuals with ID, IQ < 70, or diagnosis of an ID-associated genetic syndrome	Conference abstracts/papers, reviews, book chapters, patents, letters, editorial material, notes, brief reports, published protocols
Studies published in peer-reviewed journals	Studies employing a translated or non-English version of an eligible measurement tool
Studies published in English	

Stage two screening	
Inclusion criteria	Exclusion criteria
Studies evaluating one or more measurement properties of an eligible measurement tool or subscale (e.g., internal consistency, inter-rater reliability, test-retest reliability)	Qualitative study[a]
≥50% of participants with ID, IQ < 70, or diagnosis of an ID-associated genetic syndrome	Case studies or series
	Studies with sample overlap

[a] No qualitative studies were identified.

2.3. Data extraction

Data extraction was undertaken by LS and all data were independently checked by a second author (CG or AW) to ensure accurate data extraction. The following data were extracted from included studies: authors; year of publication; measurement properties evaluated: IC, IRR and/or TRTR; statistical tests utilised; time interval between measure administrations for studies assessing IRR and TRTR; measurement tool information: name, subscale(s), number of items, informant(s) who completed the measure and method of administration; sample size; study recruitment strategy; participant demographics: diagnosis, age, sex, and majority child or adult sample. A systematic process was applied for the inclusion of studies in the meta-analyses due to a small number of variations in methods used to report measure properties (see Supplementary material 3). Several identified papers reported overall measurement tool reliability estimates or reported estimates as a range across subscales. For these papers, study authors were contacted to obtain omitted statistical data required for inclusion in the meta-analysis. Omitted data were obtained from the authors of 3 studies and included data for 3 measures: the BPI-01 (Chan & Chien, 2017), C-SHARP (Farmer et al., 2015), and MAS (Kearney, Cook, Chapman, & Bensaheb, 2006).

2.4. Risk of bias assessment

The methodological quality of each study was assessed according to the COSMIN Risk of Bias Checklists for outcome measurements (Mokkink et al., 2018). Standards were assessed separately for each measurement property using a four-point rating system from 'inadequate' to 'very good'. The overall methodological quality of each study was based on the worst score counts principle (see Mokkink et al., 2018). LS

[3] A newly developed measure of BtC (Open Source-Challenging Behaviour Scale; Frazier et al., 2023) was identified within the updated search; however, as ID was reported in <50% of participants, this study did not meet criteria for inclusion in the current review.

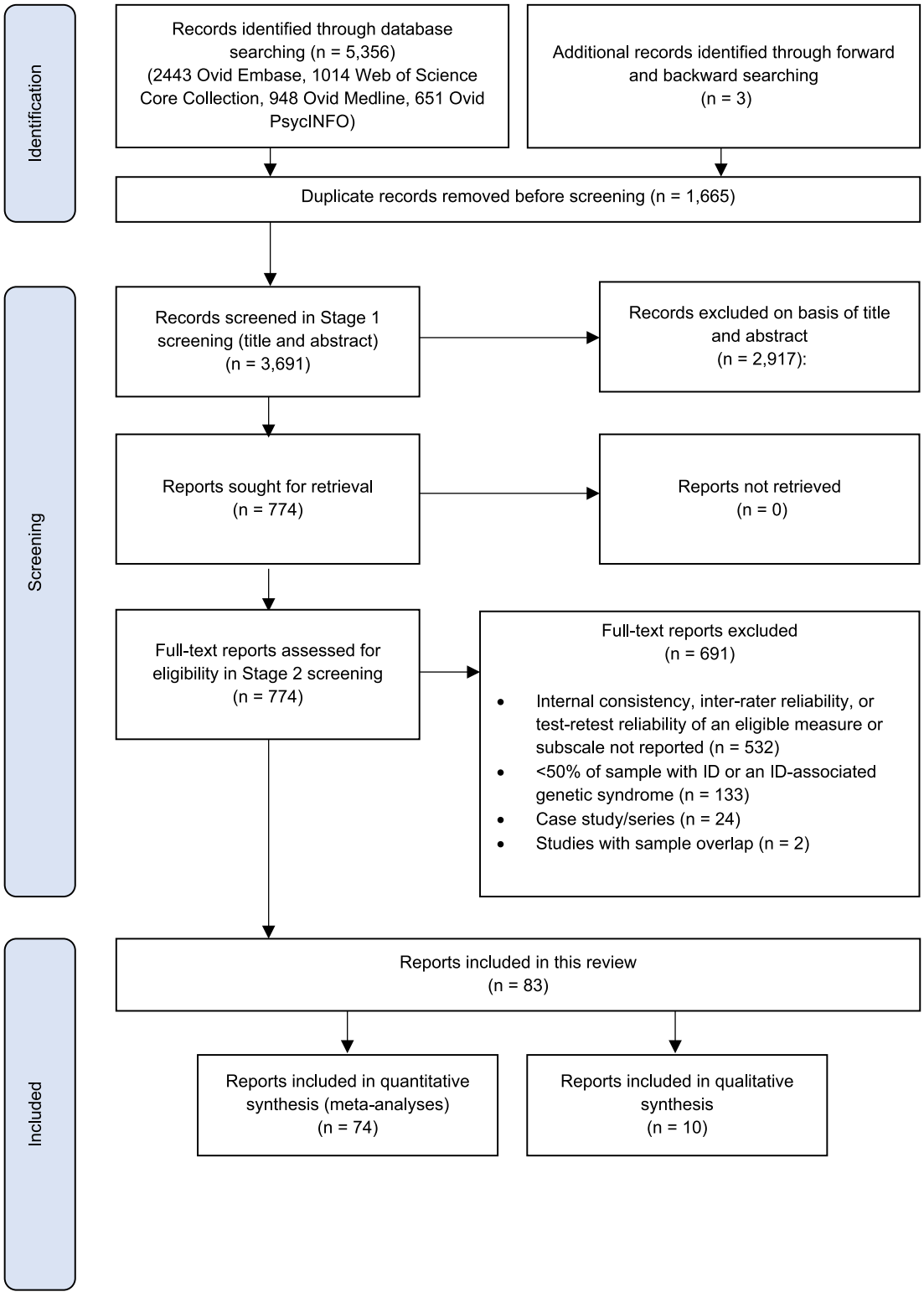

Fig. 1. PRISMA diagram detailing study selection.

completed Risk of Bias ratings for all included papers and EP independently completed ratings for 25% (k = 30) of papers; moderate to substantial IRR was established (Kappa = 0.70–1.00). Discrepancies were discussed and resolved by consensus. To ensure a comprehensive review of available measurement properties for each measurement tool, no minimum quality rating was required for study inclusion.

2.5. Data analysis

Data analysis was conducted using the 'Metafor' package for R, version 6.0. Extracted reliability data were used to generate pooled IC, IRR and TRTR estimates for each measurement tool. The genetic inverse method was used to generate pooled IC, IRR and TRTR estimates with a DerSimonian-Laird random-effects model. A random-effects model was

used over the fixed-model as it considers variation between studies and does not assume a common effect size (Hedges & Vevea, 1998; Tufanaru, Munn, Stephenson, & Aromataris, 2015). The appropriateness of the model was assessed by profiling of Quantile Quantile (QQ) plots to ensure model of distribution assumptions were held. Where meta-analyses included ≥10 effects, funnel plots were generated to assess the impact of publication bias and small-study effects.

Subgroup analyses were conducted to identify potential influence of methodological quality on the overall effects for the IC, IRR, and TRTR of measures of BtC and function. The Higgin's I^2 statistic was used to assess heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003). Potential sources of heterogeneity were further examined with exploratory subgroup analyses of study level characteristics, where there were sufficient data. Exploratory subgroup analyses examined the impact of recruitment strategy, method of administration, informant completing the measure, and the age of the participants on reliability estimates. Furthermore, meta-regression analyses were conducted to explore the impact of the time interval between test and retest on the temporal stability of TRTR estimates. A-priori analyses to explore the impact of level of ID on estimates were not possible due to a lack of data.

In line with guidelines for the interpretation of Cronbach's alpha, the minimal interpretable criteria for IC was set at 0.70; however, established measures should evidence estimates in excess of this criteria (Nunally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007). Minimal interpretable criteria for TRTR was also set at 0.70 (Nunally & Bernstein, 1994). Most studies measured IRR using Cohen's Kappa coefficient. Guidelines for the interpretation of Kappa suggest values >0.60 indicate substantial reliability, therefore, the minimal interpretable criteria for IRR was set at 0.60 (Landis & Koch, 1977).

2.6. Assessment of quality of evidence

Pooled IC, IRR and TRTR estimates for each measure were summarised using COSMIN updated criteria for good measurement properties; pooled estimates were rated as sufficient, insufficient, indeterminate, or inconsistent (Mokkink et al., 2018; Prinsen et al., 2018). The overall quality of evidence for the IC, IRR and TRTR of each measure was assessed using the COSMIN Grading of Recommendations Assessment, Development, and Evaluation Approach for systematic reviews; the level of evidence was rated as high, moderate, low, or very low depending on (1) risk of bias, (2) inconsistency, (3) imprecision, and/or (4) indirectness (Mokkink et al., 2018; Prinsen et al., 2018). The quality of evidence was not graded where the overall rating for measure IC, IRR or TRTR was indeterminate.

3. Results

5356 records were identified through database searching. Eighty-three studies were identified following screening, and 74 studies, assessing the measurement properties across 29 measures (23 measures of BtC and 6 measures of function), were included in the quantitative meta-analyses (see Fig. 1). Information on IC was available for 22 measures, IRR for 25 measures, and TRTR for 18 measures. Summary characteristics for each study are presented in Supplementary material 4, Tables D.1 and D.2. 10 studies (4 on IC, 7 on IRR and 3 on TRTR) were not included in the meta-analyses due to omission of vital statistical information (e.g., reliability reported for overall measures and not at subscale level) which could not be obtained by contacting the relevant authors. Reliability estimates from these studies are presented in Supplementary material 5.

3.1. Measures of BtC

3.1.1. Internal consistency

IC estimates were available for 17 measures of BtC from a total of 40 studies (see Table 2 for summary estimates for all measures). Overall weighted average estimates ranged from 0.64 (95% CI 0.24- ~ 1.00) for the ASD-BPA to 0.92 for the ABC-C *irritability* subscale (95% CI 0.91–0.93). Forest plots presenting the IC of all measures are shown in Supplementary material 6). Evidence of IC was limited to ≤2 studies for most measures. The ABC *irritability* subscale (k = 11), ABC-C *irritability* subscale (k = 6), BPI-01 (k = 22 effects from 7 studies), BPI-Short form (k = 12 effects from 3 studies), and NCBRF *self-injury/stereotypic* subscale (k = 4) had the largest number of studies assessing IC, with overall weighted average IC estimates also meeting the recommended threshold of α ≥0.70 for these measures. However, IC estimates were below the recommended threshold for the *Self-injurious behaviour frequency* and *Self-injurious behaviour severity* subscales from the BPI-01 (0.62 and 0.68, respectively), and *Self-injurious behaviour frequency* subscale from the BPI-Short form (0.68). A marked level of heterogeneity (Higgin's I^2 ≥75%) between reported IC estimates was identified for 7 measures, suggesting analyses of IC estimates were biased by uncontrolled or confounding factors (see Table 2).

3.1.1.1. The impact of methodological bias. Subgroup analyses were conducted to assess the overall impact of study level risk of bias on IC estimates for measures of BtC (see Table 3). It was not possible to assess the impact for COSMIN Risk of Bias IC criteria 1 (calculation of IC for separate subscales) and criteria 2–4 (statistical method used to assess IC), as all studies fell within the low-risk category. No differences in IC estimates were observed when considering flaws in the study design or statistical methods, such as the method used to account for missing data (criteria 5). See Supplementary material 4, Table D.1 for risk of bias ratings assigned to individual studies.

3.1.1.2. The impact of study level characteristics. Subgroup analyses were conducted to assess the impact of study level characteristics on IC estimates of BtC measures where sufficient data were available (see Table 4 for an overview of subgroup analyses conducted; see Supplementary material 7 for all subgroup forests plots). Across analyses, no significant differences in overall weighted average IC estimates were found across studies for the ABC *irritability* subscales and BPI-01, indicating consistency in IC estimates. Significant differences attributable to recruitment strategy were observed in the IC estimates of the NCBRF *self-injury/stereotypic* subscale, and significant differences attributable to the informant completing the measure were observed in the IC estimates of the CBCL. However, IC estimates for all informant subgroups were above the recommended α ≥0.70. No other significant differences in estimates were found for the CBCL and NCBRF.

3.1.2. Inter-rater reliability

IRR estimates were available for 19 measures of BtC from a total of 24 studies (See Table 2). Forest plots presenting the IRR of all measures are shown in Supplementary material 6. Overall weighted average IRR estimates ranged from 0.37 (95% CI 0.27–0.47) for the PDDBI-Parent *aggression* scale to 0.92 (95% CI 0.87–0.97) for the SIT. Evidence of IRR was limited to ≤2 studies for most measures. The ABC *irritability* subscale (k = 5) and BPI-01 (k = 10 effects from 4 studies) had the largest number of studies assessing IRR between informants. Overall and subscale level weighted average IRR estimates met the recommended IRR threshold (≥0.60) for the BPI-01. However, IRR estimates were below the threshold for the ABC (0.53). A marked level of heterogeneity (Higgin's I^2 ≥75%) between reported IRR estimates was identified for 5 measures, suggesting analyses of IRR estimates were biased by uncontrolled or confounding factors (see Table 2).

3.1.2.1. The impact of methodological bias. Subgroup analyses were conducted to assess the overall impact of study level risk of bias on IRR

Table 2
Overall weighted average internal consistency, inter-rater reliability and test-retest reliability estimates for measures of behaviours that challenge using random-effects models.

Measure	Subscale	Internal consistency				Inter-rater reliability				Test-retest reliability			
		k	ARAW	95% CI	I^{2a}	k	COR	95% CI	I^{2a}	k	COR	95% CI	I^{2a}
ABC	Irritability	11	**0.91**	0.90–0.92	59.00%	5	0.53	0.44–0.62	0.00%	3	**0.76**	0.47–1.05	90.00%
ABC-C	Irritability	6	**0.92**	0.91–0.93	55.00%	2	0.55	0.35–0.74	26.00%	4	**0.87**	0.78–0.96	78.00%
A-SHARP	Physical aggression problem scale	2	**0.87**	0.82–0.92	83.00%	1	**0.78**	0.66–0.90	–	–	–	–	–
	Physical aggression provocation scale	1	**0.86**	0.83–0.89		1	**0.78**	0.66–0.90	–	–	–	–	–
	Verbal aggression problem scale	2	**0.92**	0.91–0.93	0.00%	1	**0.70**	0.54–0.86	–	–	–	–	–
	Verbal aggression provocation scale	1	**0.82**	0.78–0.86		1	0.54	0.31–0.77	–	–	–	–	–
	Total measure weighted average	6	**0.88**	0.86–0.91	91.00%	4	**0.73**	0.64–0.82	26.00%	–	–	–	–
ASD-BPA	Self-injurious behaviour	1	0.43	0.28–0.58	–	1	0.46	0.34–0.58	–	1	0.66	0.42–0.90	–
	Aggression/destruction	**1**	**0.83**	0.79–0.87	–	1	0.47	0.35–0.59	–	1	0.65	0.41–0.89	–
	Total measure weighted average	2	0.64	0.24–1.03	96.00%	2	0.47	0.38–0.55	0.00%	2	0.66	0.49–0.82	0.00%
BPI-01	Aggressive/destructive behaviour frequency	7	**0.83**	0.79–0.86	93.00%	4	**0.75**	0.65–0.85	77.00%	6	**0.78**	0.70–0.87	87.00%
	Aggressive/destructive behaviour severity	4	**0.86**	0.82–0.90	92.00%	1	**0.77**	0.69–0.85	–	2	**0.75**	0.62–0.87	73.00%
	Self-injurious behaviour frequency	7	0.62	0.54–0.70	92.00%	4	**0.65**	0.45–0.84	91.00%	6	**0.77**	0.66–0.87	92.00%
	Self-injurious behaviour severity	4	0.68	0.58–0.79	95.00%	1	**0.63**	0.51–0.75	–	2	0.67	0.60–0.74	0.00%
	Total measure weighted average	22	**0.75**	0.71–0.78	98.00%	10	**0.71**	0.64–0.79	84.00%	16	**0.76**	0.71–0.81	89.00%
BPI-Short form	Aggressive/destructive behaviour frequency	3	**0.82**	0.81–0.89	98.00%	1	0.58	0.47–0.69	–	1	**0.77**	0.72–0.82	–
	Aggressive/destructive behaviour severity	3	**0.85**	0.81–0.89	94.00%	1	0.44	0.31–0.57	–	1	**0.71**	0.65–0.77	–
	Self-injurious behaviour frequency	3	0.68	0.65–0.71	40.00%	1	**0.71**	0.63–0.79	–	1	**0.87**	0.84–0.90	–
	Self-injurious behaviour severity	3	**0.71**	0.68–0.73	37.00%	1	**0.60**	0.50–0.70	–	1	**0.85**	0.82–0.88	–
	Total measure weighted average	12	**0.76**	0.72–0.81	98.00%	4	0.59	0.48–0.70	76.00%	4	**0.80**	0.74–0.87	90.90%
BISCUIT-Part 3	Aggressive/disruptive behaviour	2	**0.86**	0.83–0.89	78.00%	–	–	–	–	–	–	–	–
	Self-injurious behaviour	2	0.44	0.31–0.57	65.00%	–	–	–	–	–	–	–	–
	Total measure weighted average	4	0.68	0.57–0.79	98.00%	–	–	–	–	–	–	–	–
CBCL 1.5–5	Aggressive behaviour	2	**0.88**	0.87–0.90	0.00%	–	–	–	–	–	–	–	–
CBCL 6–18	Aggressive behaviour	2	**0.90**	0.88–0.93	0.00%	1	**0.65**	0.53–0.77	–	1	0.52	0.28–0.76	–
CBCL-TRF	Aggressive behaviour	1	**0.97**	0.96–0.98	–	–	–	–	–	2	**0.90**	0.86–0.95	0.00%
CBI	Disruption of environment severity	–	–	–	–	1	**0.77**	0.41–1.13	–	1	**0.77**	0.41–1.13	–
	Inappropriate vocalisations severity	–	–	–	–	1	0.02	−0.72-0.76	–	1	0.66	0.24–1.08	–
	Physical aggression severity	–	–	–	–	1	0.54	0.15–0.93	–	1	**0.76**	0.53–0.99	–
	Self-injury severity	–	–	–	–	1	**0.63**	0.24–1.02	–	1	**0.85**	0.67–1.03	–
	Verbal aggression severity	–	–	–	–	1	0.45	−0.10-1.00	–	1	**0.75**	0.45–1.05	–
	Total measure weighted average	–	–	–	–	5	0.58	0.39–0.78	0.00%	5	**0.79**	0.67–0.91	0.00%
CCB	Aggressive behaviour frequency	–	–	–	–	1	**0.73**	0.20–1.26	–	1	0.61	0.06–1.16	–
	Aggressive behaviour severity	–	–	–	–	1	0.57	−0.19-1.33	–	1	0.59	0.02–1.16	–
	Aggressive behaviour management difficulty	–	–	–	–	1	0.50	−0.35-1.35	–	1	0.53	−0.10-1.16	–
	Other challenging behaviour frequency	–	–	–	–	1	0.56	−0.22-1.34	–	1	0.63	0.10–1.16	–
	Other challenging behaviour management difficulty	–	–	–	–	1	0.53	−0.28-1.34	–	1	0.61	0.06–1.16	–
	Total measure weighted average	–	–	–	–	5	**0.61**	0.29–0.93	0.00%	5	0.60	0.35–0.85	0.00%
C-SHARP	Bullying problem scale	2	**0.89**	0.88–0.90	0.00%	1	**0.90**	0.82–0.98	–	–	–	–	–
	Bullying provocation scale	1	**0.81**	0.79–0.83		1	0.55	0.09–1.01	–	–	–	–	–
	Physical aggression problem scale	2	**0.75**	0.73–0.78	0.00%	1	**0.80**	0.65–0.95	–	–	–	–	–
	Physical aggression provocation scale	1	0.68	0.64–0.72		1	0.47	−0.21-1.15	–	–	–	–	–
	Verbal aggression problem scale	2	**0.91**	0.90–0.92	35.00%	1	**0.86**	0.75–0.97	–	–	–	–	–
	Verbal aggression provocation scale	1	**0.81**	0.79–0.83		1	**0.76**	0.42–1.10	–	–	–	–	–
	Total measure weighted average	9	**0.83**	0.79–0.87	98.00%	6	**0.86**	0.80–0.90	0.00%	–	–	–	–
IBR-MOAS	Physical aggression against objects	2	**0.80**	0.79–0.81	0.00%	1	**0.80**	0.66–0.94	–	1	**0.96**	0.92–1.00	–
	Physical aggression against others	2	**0.85**	0.78–0.93	82.00%	1	**0.70**	0.50–0.90	–	1	**0.92**	0.84–1.00	–
	Physical aggression against self	2	**0.85**	0.78–0.93	82.00%	1	**0.80**	0.66–0.94	–	1	**0.92**	0.84–1.00	–
	Verbal aggression towards others	2	**0.85**	0.84–0.86	0.00%	1	**0.83**	0.71–0.95	–	1	**0.87**	0.75–0.99	–
	Verbal aggression towards self	2	**0.78**	0.68–0.87	70.00%	1	**0.73**	0.54–0.92	–	1	**0.84**	0.69–0.99	–
	Total measure weighted average	10	**0.81**	0.78–0.85	97.00%	5	**0.80**	0.73–0.86	0.00%	5	**0.92**	0.87–0.97	31.80%
LDNAT	Challenging behaviour	1	**0.76**	0.74–0.78	–	–	–	–	–	1	**0.93**	0.88–0.98	–
MOAS	Verbal aggression	–	–	–	–	1	**0.90**	0.85–0.95	–	–	–	–	–
	Physical aggression against objects	–	–	–	–	1	0.56	0.38–0.74	–	–	–	–	–
	Physical aggression against self	–	–	–	–	1	0.49	0.30–0.68	–	–	–	–	–
	Physical aggression against other people	–	–	–	–	1	**0.90**	0.85–0.95	–	–	–	–	–
	Total measure weighted average	–	–	–	–	4	**0.76**	0.63–0.88	90.00%	–	–	–	–
NCBRF	Self- injury/stereotypic	4	**0.80**	0.77–0.83	67.00%	3	0.46	0.14–0.78	92.00%	1	**0.90**	0.82–0.98	–
OAS	Aggressive behaviour	–	–	–	–	1	**0.85**	0.64–1.06	–	–	–	–	–

(*continued on next page*)

Table 2 (*continued*)

Measure	Subscale	Internal consistency				Inter-rater reliability				Test-retest reliability			
		k	ARAW	95% CI	I^{2a}	k	COR	95% CI	I^{2a}	k	COR	95% CI	I^{2a}
PBCL	Challenging behaviour	–	–	–	–	1	**0.91**	0.85–0.97	–	–	–	–	–
PDDBI-Parent	Aggression	1	**0.89**	0.87–0.91	–	1	0.37	0.27–0.47	–	–	–	–	–
PDDBI-Teacher	Aggression	1	**0.88**	0.86–0.90	–	1	0.55	0.35–0.75	–	–	–	–	–
SIT	Location of self-injury	–	–	–	–	1	**0.99**	0.98–1.00	–	–	–	–	–
	Type of self-injury	–	–	–	–	1	**0.92**	0.88–0.96	–	–	–	–	–
	Number of self-injuries	–	–	–	–	1	**0.84**	0.76–0.92	–	–	–	–	–
	Severity of self-injury	–	–	–	–	1	**0.93**	0.89–0.97	–	–	–	–	–
	Number index	–	–	–	–	1	**0.88**	0.82–0.94	–	–	–	–	–
	Severity index	–	–	–	–	1	**0.90**	0.85–0.95	–	–	–	–	–
	Total measure weighted average	–	–	–	–	6	**0.92**	0.87–0.97	90.00%	–	–	–	–
SOAS-ID-R	Overall aggressive behaviour	–	–	–	–	1	**0.72**	0.52–0.92	–	–	–	–	–

Note. This table includes an overview of estimates for measures where data were available for at least one type of reliability and does not include all measures identified as eligible for the review in the preliminary search. The full list of measures of BtC identified in the preliminary search can be found in Supplementary material 1. Estimates meeting minimal interpretable criteria are highlighted in bold. k = number of effects in weighted average estimate.
ABC = Aberrant behaviour checklist, ABC-C = Aberrant behaviour checklist-Community version, A-SHARP = Adult Scale of Hostility and Aggression, ASD-BPA = Autism Spectrum Disorder-Behavior Problems for Adults, BPI-01 =Behavior Problems Inventory-01, BPI-Short Form =Behavior Problems Inventory-Short Form, BISCUIT-Part 3 = Baby and Infant Screen for Children with aUtism Traits-Part 3, CBCL 1.5–5 = Child Behavior Checklist 1.5–5, CBCL 6–18 = Child Behavior Checklist 6–18, CBCL-TRF=Child Checklist-Teacher Report Form, CBI=Challenging Behaviour Interview, CCB=Checklist of Challenging Behaviour, C-SHARP=Children's Scale of Hostility and Aggression, IBR-MOAS=Institute for Basic Research-Modified Overt Aggression Scale, LDNAT = Learning Disability Needs Assessment Tool, MOAS = Modified Overt Aggression Scale, NCBRF=Nisonger Child Behavior Rating Form, OAS=Overt Aggression Scale, PBCL = Problem Behavior Checklist, PDDBI-Parent = Pervasive Developmental Disorder Behavior Inventory-Parent Version, PDDBI-Teacher = Pervasive Developmental Disorder Behavior Inventory-Teacher Version, SIT = Self-injury Trauma Scale, SOAS-ID-R = Staff Observation Aggression Scale – Revised.
[a] = Higgin's I^2 not calculated when k = 1. - = no data were available.

estimates for measures of BtC (see Table 3). Significant differences in IRR estimates were observed for COSMIN Risk of Bias Reliability criteria 1, 2, 3, 4 and 8. Significantly higher IRR estimates were observed across measures when a shorter time interval (criteria 2) and similar test conditions (criteria 3) were used between informant ratings, while lower IRR estimates were observed across measures when there were flaws in the study design or statistical methods (criteria 8), such as ratings from informant pairs with poor IRR being excluded from the analysis (Rojahn & Helsel, 1991). Significant differences attributable to the stability of participant behaviour, e.g., not undergoing a behavioural intervention, between informant ratings (criteria 1) and statistic used (criteria 4) and were also observed. See Supplementary material 4, Table D.1 for risk of bias ratings assigned to individual studies.

3.1.2.2. The impact of study level characteristics. Subgroup analyses were conducted to assess the impact of study level characteristics on IRR estimates of BtC measures where sufficient data were available (see Table 4 for an overview of subgroup analyses conducted; see Supplementary material 7 for all subgroup forests plots). Across analyses, no significant differences in overall weighted average IRR estimates were found for the ABC *irritability* subscale, indicating consistency in weighted average IRR estimates across groups. Significant differences attributable to informants completing the measure were observed in the IRR estimates for the BPI-01, with educator-educator and professional-professional informant pairs (e.g., direct-care staff) evidencing higher IRR than educator-professional informant pairs (0.80 vs 0.30 respectively). Significant differences were also found for the NCBRF *self-injury/stereotypic* subscale, with educator-educator informant pairs evidencing higher IRR than parent-educator and educator-professional informant pairs (0.77 vs 0.54 and 0.03 respectively). No other significant differences in estimates were found for the BPI-01 and NCBRF.

3.1.3. Test-retest reliability

TRTR estimates were available for 12 measures of BtC from a total of 17 studies (see Table 2 for a summary of estimates for all measures). Forest plots presenting the IRR of all measures are shown in Supplementary material 6. Overall weighted average TRTR estimates ranged from 0.52 (95% CI 0.28–0.76) for the CBCL 6–18 *aggressive behaviour* scale to 0.93 (95% CI 0.88–0.98) for the LDNAT *challenging behaviour* scale. Evidence of TRTR in was limited to ≤2 studies for most measures. The ABC *irritability* subscale (k = 3), ABC-C *irritability* subscale (k = 4), and BPI-01 (k = 16 effects from 6 studies) had the largest number of studies assessing TRTR, with overall weighted average TRTR estimates also meeting the recommended threshold of ≥0.70 for these measures. However, the BPI-01 *Self-injurious behaviour severity* subscale did not reach the recommended TRTR reliability threshold (TRTR = 0.67). A marked level of heterogeneity (Higgin's I^2 ≥75%) between reported TRTR estimates was identified for 4 measures, suggesting analyses of TRTR estimates were biased by uncontrolled or confounding factors (see Table 2).

3.1.3.1. The impact of methodological bias. Subgroup analyses were conducted to assess the overall impact of study level risk of bias on TRTR estimates for measures of BtC. Analyses were conducted for each of the eight types of methodological bias according to the COSMIN Risk of Bias Reliability Checklist (see Table 3). Significant differences were observed when criteria 3 was violated, less similar conditions for test and retest completions of measures (e.g., type of administration, environment, and instructions) leading to lower TRTR estimates. Significant differences attributable to the statistic used to assess TRTR (criteria 4) were also observed; however, estimates were above the recommended TRTR threshold in all risk of bias categories. See Supplementary material 4, Table D.1 for risk of bias ratings assigned to individual studies.

3.1.3.2. The impact of study level characteristics. Subgroup analyses were conducted to assess the impact of study level characteristics on TRTR estimates of BtC measures where sufficient data were available (see Table 4 for an overview of all subgroup analyses conducted; see Supplementary material 7 for all subgroup forests plots). In addition, meta-regressions were conducted to assess the length of time between test and retest upon the temporal stability of measures of BtC.

Significant differences attributable to recruitment strategy were observed in the TRTR estimates of the ABC *irritability* subscale, with higher TRTR estimates observed with recruitment from healthcare and school settings, compared to recruitment from community-based settings (0.91 and 0.87 vs 0.59, respectively). Significant differences

Table 3
Subgroup analyses for the effect of study level risk of bias on reported internal consistency, inter-rater reliability, and test-retest reliability of measures of BtC.

COSMIN risk of bias box 4: Internal consistency criteria:	Rating [a]	Internal consistency				COSMIN risk of bias box 6: Reliability criteria:	Rating [a]	Inter-rater reliability				Test-retest reliability			
		k	ARAW	95% CI	X^2			k	COR	95% CI	X^2	k	COR	95% CI	X^2
1. Statistic calculated for separate subscales	Very good	94	0.82	0.80–0.83	-[c]	1. Participant stability between ratings	Very good	10	0.58	0.42–0.74	42.99**	0	-[b]	-[b]	3.27
	Adequate	–	–	–			Adequate	37	0.68	0.63–0.73		28	0.82	0.79–0.86	
	Doubtful	0	-[b]	-[b]			Doubtful	12	0.67	0.55–0.80		16	0.81	0.75–0.87	
	Inadequate	0	-[b]	-[b]			Inadequate	9	0.90	0.84–0.95		6	0.76	0.71–0.82	
	Not applicable	–	–	–			Not applicable	–	–	–		–	–	–	
2. Continuous scale statistic	Very good	92	0.80	0.77–0.84	-[c]	2. Time interval between ratings	Very good	37	0.75	0.70–0.80	0.42*	16	0.82	0.76–0.88	0.42
	Adequate	–	–	–			Adequate	–	–	–		–	–	–	
	Doubtful	0	-[b]	-[b]			Doubtful	31	0.65	0.58–0.71		28	0.80	0.76–0.85	
	Inadequate	0	-[b]	-[b]			Inadequate	0	-[b]	-[b]		6	0.84	0.82–0.86	
	Not applicable	2	-[b]	-[b]			Not applicable	–	–	–		–	–	–	
3. Dichotomous scale statistic	Very good	2	-[b]	-[b]	-[c]	3. Similarity of test conditions	Very good	22	0.78	0.73–0.84	24.46**	8	0.80	0.70–0.91	19.20**
	Adequate	–	–	–			Adequate	44	0.66	0.60–0.71		26	0.84	0.79–0.88	
	Doubtful	0	-[b]	-[b]			Doubtful	2	-[b]	-[b]		12	0.80	0.73–0.86	
	Inadequate	0	-[b]	-[b]			Inadequate	0	-[b]	-[b]		4	0.67	0.61–0.73	
	Not applicable	92	0.80	0.77–0.84			Not applicable	–	–	–		–	–	–	
4. IRT-based score statistic	Very good	0	-[b]	-[b]	-[c]	4. Continuous scale statistic	Very good	17	0.64	0.56–0.72	52.28**	4	0.73	0.67–0.80	9.13*
	Adequate	–	–	–			Adequate	17	0.73	0.65–0.82		15	0.85	0.81–0.90	
	Doubtful	–	–	–			Doubtful	20	0.65	0.58–0.73		29	0.79	0.75–0.84	
	Inadequate	0	-[b]	-[b]			Inadequate	11	0.91	0.86–0.96		0	-[b]	-[b]	
	Not applicable	94	0.82	0.80–0.83			Not applicable	3	-[b]	-[b]		2	-[b]	-[b]	
5. Other design or statistical flaws	Very good	80	0.87	0.87–0.88	0.59	5. Dichotomous, nominal, or ordinal scale statistic	Very good	3	0.62	0.28–0.95	0.23	2	-[b]	-[b]	-[c]
	Adequate	–	–	–			Adequate	–	–	–		–	–	–	
	Doubtful	14	0.81	0.77–0.85			Doubtful	–	–	–		–	–	–	
	Inadequate	0	-[b]	-[b]			Inadequate	0	-[b]	-[b]		0	-[b]	-[b]	
	Not applicable	–	–	–			Not applicable	65	0.70	0.65–0.75		48	0.81	0.77–0.84	
						6. Ordinal scale weighted kappa	Very good	1	-[b]	-[b]	-[c]	0	-[b]	-[b]	-[c]
							Adequate	–	–	–		–	–	–	
							Doubtful	0	-[b]	-[b]		0	-[b]	-[b]	
							Inadequate	–	–	–		–	–	–	
							Not applicable	67	0.69	0.64–0.74		50	0.80	0.77–0.84	
						7. Ordinal scale weighting scheme	Very good	0	-[b]	-[b]	-[c]	0	-[b]	-[b]	-[c]
							Adequate	1	-[b]	-[b]		0	-[b]	-[b]	
							Doubtful	–	–	–		–	–	–	
							Inadequate	–	–	–		–	–	–	
							Not applicable	67	0.69	0.64–0.74		50	0.80	0.77–0.84	
						8. Other design or statistical flaws	Very good	53	0.67	0.62–0.72	34.02**	28	0.78	0.73–0.83	5.66
							Adequate	–	–	–		–	–	–	
							Doubtful	11	0.85	0.80–0.91		16	0.85	0.81–0.89	
							Inadequate	4	0.55	0.44–0.67		6	0.79	0.62–0.97	
							Not applicable	–	–	–		–	–	–	

Note. ** $p < .001$.
[a] Not all rating levels are used for all COSMIN reliability/internal consistency criteria.
[b] ARAW/COR and 95% CI not reported where k < 4.
[c] Not interpretable.

Table 4
Subgroup analyses for differences in overall weighted average estimates of internal consistency, inter-rater reliability and test-retest reliability estimates for measures of behaviours that challenge, attributable to recruitment strategy, informant completing the measure, administration regarding children or adults, and method of administration.

Measure	Internal consistency				Inter-rater reliability				Test-retest reliability			
	Recruitment strategy	Informant	Administration regarding child or adult	Method of administration	Recruitment strategy	Informant	Administration regarding child or adult	Method of administration	Recruitment strategy	Informant	Administration regarding child or adult	Method of administration
ABC/ ABC-C[a]	o	o	o	o	o	o	o	–	+[g]	o	+[j]	–
A-SHARP	–	–	–	–	–	–	–	–	–	–	–	–
ASD-BPA	–	–	–	–	–	–	–	–	–	–	–	–
BPI-01	o	–	o	o	o	+[e]	o	–	+[h]	+[i]	o	–
BPI-Short form	–	o	–	–	–	–	–	–	–	–	–	–
BISCUIT-Part 3	–	–	–	–	–	–	–	–	–	–	–	–
CBCL[b]	o	+[d]	–	–	–	–	–	–	–	–	–	–
CBI	–	–	–	–	–	–	–	–	–	–	–	–
CCB	–	–	–	–	–	–	–	–	–	–	–	–
C-SHARP	–	–	–	–	–	–	–	–	–	–	–	–
IBR-MOAS	–	–	–	–	–	–	–	–	–	–	–	–
LDNAT	–	–	–	–	–	–	–	–	–	–	–	–
MOAS	–	–	–	–	–	–	–	–	–	–	–	–
NCBRF	+[c]	o	–	–	o	+[f]	–	–	–	–	–	–
OAS	–	–	–	–	–	–	–	–	–	–	–	–
PBCL	–	–	–	–	–	–	–	–	–	–	–	–
PDDBI-Parent	–	–	–	–	–	–	–	–	–	–	–	–
PDDBI-Teacher	–	–	–	–	–	–	–	–	–	–	–	–
SIT	–	–	–	–	–	–	–	–	–	–	–	–
SOAS-ID-R	–	–	–	–	–	–	–	–	–	–	–	–

Note. + = significant differences in estimates, o = no significant differences in estimates, − = subgroup analyses not conducted due to lack of available data.
See Supplementary material 7 for all subgroup analysis forest plots.

[a] estimates from the ABC and ABC-C were pooled for subgroup analyses due to the paucity of available data.

[b] Estimates from the CBCL 1.5–5, CBCL 6–18 and CBCL-TRF were pooled for subgroup analyses due to the paucity of available data.

[c] Higher internal consistency estimates evidenced with recruitment from healthcare (k = 2) compared to school (k = 1) settings; however, a ≥ 0.70 for all estimates.

[d] Higher internal consistency estimates evidenced when completed by educators (k = 1) compared to parents or caregivers (k = 4); however, a ≥ 0.70 for all estimates.

[e] Higher inter-rater reliability estimates evidenced when completed by educator-educator (k = 2, $r = 0.80$) and professional-professional rater pairs (k = 6, $r = 0.77$) compared to educator-professional rater pairs (k = 2, $r = 0.30$).

[f] Higher inter-rater reliability estimates evidenced when completed by educator-educator (k = 1, r = 0.77) rater pairs compared to parent-educator (k = 1, $r = 0.54$) and educator-professional (k = 1, $r = 0.03$) rater pairs.

[g] Higher test-retest reliability estimates evidenced with recruitment from healthcare (k = 4, $r = 0.91$) and school (k = 2, $r = 0.87$) settings compared to community-based settings (k = 1, $r = 0.59$).

[h] Higher test-retest reliability estimates evidenced with recruitment from healthcare (k = 2) and school (k = 2) settings compared to community-based settings (k = 8) and charity organisations (k = 4); however, all estimates are >0.70.

[i] Higher test-retest reliability estimates evidenced when completed by educators (k = 2) compared to parents or caregivers (k = 4) and professionals (k = 10); however, all estimates are >0.70.

[j] Higher test-retest reliability estimates evidenced when used to rate the behaviour of children (k = 2, r = 0.87) compared to adults (k = 1, r = 0.59).

attributable to administration regarding a majority child or adult sample were also observed for the ABC *irritability* subscale, with higher TRTR estimates observed when the irritability subscale was completed regarding the behaviour of children compared to adults (0.87 vs 0.59 respectively). There were no significant differences in TRTR estimates of the ABC *irritability* subscale according to the informant completing the measure. Meta-regression analyses assessing the impact of the length of time between test and retest on the temporal stability of the ABC revealed a non-significant reduction of TRTR overtime ($\beta = -0.0037$, $z = -0.91$, $p = .36$).

Significant differences attributable to recruitment strategy and informant completing the measure were observed in the TRTR estimates for the BPI-01; however, TRTR estimates were above the recommended threshold of ≥0.70 for all groups in both subgroup analyses. There were no significant differences in TRTR estimates of the BPI-01 according to administration regarding a majority child or adult sample. Meta-regression analyses assessing the impact of the length of time between test and retest on the temporal stability of the BPI-01 revealed a non-significant reduction of TRTR overtime ($\beta = -0.0003$, $z = -0.65$, $p = .51$).

3.1.4. The impact of publication and small study biases

Funnel plots were generated for studies assessing the IC, IRR and TRTR of BtC measures where ≥10 effects were available. It was not possible to assess the impact of publication bias and small study effects on IC, IRR and TRTR estimates for the majority of measures of BtC, as <10 effects were reported (see Table 2).

There was evidence of heterogeneity in studies assessing IC for the ABC and BPI-Short Form; however, minimal evidence of publication bias was observed as several small studies reported effects below the recommended threshold ($\alpha \geq 0.70$). Evidence of heterogeneity was observed in studies assessing the IC, IRR and TRTR of the BPI-01; however, several small studies reported effects below the recommended thresholds and there was no clear evidence of publication bias. Consequently, no simulation and adjustment for publication bias and small study effects was undertaken for these measures. Clear effects of heterogeneity were observed for studies assessing the IC of the IBR-MOAS; therefore, the Trim and Fill procedure was conducted to correct for the effects of publication bias (Duval & Tweedie, 2000a, 2000b). An imputed IC estimate of 0.79 (95% CI 0.76–0.83) was indicated, representing a negligible −2.39% decrease relative to the original omnibus analysis for the IBR-MOAS.

3.1.4.1. Overall quality of evidence for measures of BtC. Ratings of the quality and level of evidence for pooled IC, IRR and TRTR estimates for each measure are shown in Table 5 (Mokkink et al., 2018; Prinsen et al., 2018). Sufficient quality ratings for IC were available for 12 measures and inconsistent quality ratings for 5 measures. Across measures, the level of evidence for IC ranged from low to high. Moderate to high evidence of sufficient IC was available for the ABC *irritability* subscale, A-SHARP, CBCL 1.5–5 and 6–18 *aggressive behaviour* subscales, IBR-MOAS, LDNAT *challenging behaviour* scale, and NCBRF *self-injury/stereotypic* subscale. Moderate evidence of IC was available for the ASD-BPA and BPI-Short form; however, these measures received an inconsistent quality rating due to inconsistency in pooled IC estimates across subscales. Sufficient quality ratings for IRR were available for 8 measures, insufficient for 5 measures, and inconsistent for 6 measures. However, the level of evidence for IRR was low to very low for all measures except the ASD-BPA, which had a moderate level of evidence for inconsistent pooled IRR estimates between subscales. Sufficient quality ratings for TRTR were available for 7 measures, insufficient for 3 measures, and inconsistent for 2 measures. Levels of evidence for TRTR ranged from low to very low across all measures. Overall, levels of evidence for IC, IRR and TRTR were frequently downgraded due to risk of bias, inconsistency in pooled estimates between subscales or studies, and low numbers of participants (n ⟨100) (Mokkink et al., 2018; Prinsen et al., 2018).

3.2. Measures of behavioural function

3.2.1. Internal consistency

IC estimates were available for 6 measures of function from a total of 16 studies (see Table 6). Overall weighted average IC estimates ranged from 0.41 (95% CI 0.13–0.70) for the FAST to 0.94 (95% CI 0.93–0.95) for the FACT. Forest plots presenting the IC of all measures are shown in Supplementary material 8. Evidence of IC was limited to 1 study for the CAI, FAST and QABF-Short form. The FACT (k = 15 effects from 3 studies), MAS (k = 36 effects from 9 studies) and QABF (k = 30 effects from 6 studies) had the largest number of studies assessing IC, with overall and subscale level weighted average IC estimates also meeting the recommended threshold of $\alpha \geq 0.70$ for these measures. A marked level of heterogeneity (Higgin's $I^2 \geq 75\%$) between reported IC estimates was identified for 5 measures, suggesting analyses of IC estimates were biased by uncontrolled or confounding factors (see Table 6).

3.2.1.1. The impact of methodological bias. Subgroup analyses were planned to assess the overall impact of study level risk of bias on IC estimates for measures of function. However, it was not possible to assess the impact of bias, due to all studies falling under the low-risk category for each of the five types of methodological bias according to the COSMIN Risk of Bias IC Checklist (see Table 7). See Supplementary material 4, Table D.2 for risk of bias ratings assigned to individual studies.

3.2.1.2. The impact of study level characteristics. Subgroup analyses were conducted to assess the impact of study level characteristics on IC estimates of function measures where sufficient data were available (see Table 8 for an overview of subgroup analyses conducted; see Supplementary material 9 for all subgroup forests plots).

Significant differences attributable to recruitment strategy were observed in the IC estimates of the MAS, with higher IC estimates observed with recruitment from healthcare settings and community-based organisations, compared to recruitment from school settings (alphas of 0.77 and 0.80 vs 0.67 respectively). Significant differences attributable to the informant completing the measure were also observed for the MAS, with higher IC estimates observed when the MAS was completed by professionals and parents or caregivers, compared to educations (alphas of 0.80 and 0.77 vs 0.67 respectively).

Significant differences attributable to recruitment strategy, informant completing the measure, and administration regarding a majority child or adult sample were observed in the IC estimates for the QABF; however, in all subgroup analyses, IC estimates were above the recommended threshold of ≥0.70 for all groups. No other significant differences in estimates were found for the QABF.

3.2.2. Inter-rater reliability

IRR estimates were available for 6 measures of function from a total of 17 studies (see Table 6 for a summary of estimates for all measures). Overall weighted average IRR estimates ranged from 0.29 (95% CI 0.09–0.50) for the CAI to 0.93 (95% CI 0.91–0.96) for the QABF-Short form. Forest plots presenting the IRR of all measures are shown in Supplementary material 8. Evidence of IRR was limited to ≤2 studies for most measures. The MAS (k = 44 effects from 10 studies) and QABF (k = 15 effects from 3 studies) had the largest number of studies assessing IRR between informants. However, overall and subscale level weighted average IRR estimates for the MAS and QABF failed to meet the recommended threshold of ≥0.60. Overall and subscale level weighted average IRR estimates for the FACT and QABF-Short form met the

Table 5
Summary of quality ratings for the measurement properties of each measure of BtC according to COSMIN criteria for good measurement properties and COSMIN criteria for overall quality of evidence.

Measure	Measurement property	Number of papers (reference number) [1]	Participants (*n*)	Pooled measurement property quality rating [2]	Quality of evidence rating (GRADE) [3]
ABC *irritability*	IC	11 (3, 5, 7, 19, 32, 50, 59, 63, 66, 71, 77)	3058	Sufficient	Moderate
	IRR	**5** (4, 5, 59, 69, 73)	245	Insufficient	Very Low
	TRTR	3 (4, 5, 69)	242	Sufficient	Low
ABC-C *irritability*	IC	6 (2, 10, 24, 36, 48) [b]	2029	Sufficient	Low
	IRR	2 (24, 28)	102	Insufficient	Low
	TRTR	4 (24, 48) [c]	125	Sufficient	Low
A-SHARP	IC	2 (38, 62)	667	Sufficient	High
	IRR	1 (39)	39	Inconsistent [a]	Low
	TRTR	0	0	Indeterminate	N/A
ASD-BPA	IC	1 (40)	171	Inconsistent [a]	Moderate
	IRR	1 (40)	171	Insufficient	Moderate
	TRTR	1 (40)	23	Insufficient	Very Low
BPI-01	IC	**7** (23, 60, 61, 64, 65, 66, 81)	2695	Inconsistent [a]	Low
	IRR	4 (25, 64, 81) [d]	270	Sufficient	Low
	TRTR	6 (12, 25, 61, 64, 66, 81)	703	Inconsistent [a]	Very Low
BPI-Short form	IC	3 (9, 37, 65)	2063	Inconsistent [a]	Moderate
	IRR	1 (37)	147	Inconsistent [a]	Very Low
	TRTR	1 (37)	147	Sufficient	Low
BISCUIT-Part 3	IC	2 (42, 43)	914	Inconsistent [a]	Low
	IRR	0	0	Indeterminate	N/A
	TRTR	0	0	Indeterminate	N/A
CBCL 1.5–5 *aggressive behaviour*	IC	2 (49, 58)	280	Sufficient	Moderate
	IRR	0	0	Indeterminate	N/A
	TRTR	0	0	Indeterminate	N/A
CBCL 6–18 *aggressive behaviour*	IC	2 (18, 48)	132	Sufficient	High
	IRR	1 (18)	88	Sufficient	Low
	TRTR	1 (48)	36	Insufficient	Very Low
CBCL-TRF *aggressive behaviour*	IC	1 (48)	47	Sufficient	Low
	IRR	0	0	Indeterminate	N/A
	TRTR	2 (48) [e]	70	Sufficient	Low
CBI	IC	0	0	Indeterminate	N/A
	IRR	1 (54)	6–14 [f]	Inconsistent [a]	Very Low
	TRTR	1 (54)	6–14 [f]	Inconsistent [a]	Very Low
CCB	IC	0	0	Indeterminate	N/A
	IRR	1 (27)	4	Inconsistent [a]	Very Low
	TRTR	1 (27)	6	Insufficient	Very Low
C-SHARP	IC	2 (20, 22)	380–384 [g]	Inconsistent [a]	Low
	IRR	1 (21)	6–22 [h]	Inconsistent [a]	Very Low
	TRTR	0	0	Indeterminate	N/A
IBR-MOAS	IC	2 (14) [i]	3572	Sufficient	High
	IRR	1 (14)	25	Sufficient	Very Low
	TRTR	1 (14)	16	Sufficient	Very Low
LDNAT *challenging behaviour*	IC	1 (57)	1692	Sufficient	High
	IRR	0	0	Indeterminate	N/A
	TRTR	1 (57)	27	Sufficient	Low
MOAS	IC	0	0	Indeterminate	N/A
	IRR	1 (55)	60	Inconsistent [a]	Very Low
	TRTR	0	0	Indeterminate	N/A
NCBRF *self-injury/stereotypic*	IC	4 (6, 53, 64) [j]	1212	Sufficient	High
	IRR	3 (6, 64) [k]	281	Insufficient	Low
	TRTR	1 (64)	24	Sufficient	Very Low
OAS *aggressive behaviour*	IC	0	0	Indeterminate	N/A
	IRR	1 (28)	8	Sufficient	Very Low
	TRTR	0	0	Indeterminate	N/A
PBCL *challenging behaviour*	IC	0	0	Indeterminate	N/A
	IRR	1 (79)	38	Sufficient	Very Low
	TRTR	0	0	Indeterminate	N/A
PDDBI-Parent *aggression*	IC	1 (13)	311	Sufficient	Low
	IRR	1 (13)	271	Sufficient	Very Low
	TRTR	0	0	Indeterminate	N/A
PDDBI-Teacher *aggression*	IC	1 (13)	298	Sufficient	Low
	IRR	1 (13)	49	Insufficient	Very Low
	TRTR	0	0	Indeterminate	N/A
SIT	IC	0	0	Indeterminate	N/A
	IRR	1 (31)	50	Sufficient	Very Low
	TRTR	0	0	Indeterminate	N/A
SOAS-ID-R *overall aggressive behaviour*	IC	0	0	Indeterminate	N/A
	IRR	1 (80)	23	Sufficient	Very Low
	TRTR	0	0	Indeterminate	N/A

Note. This table includes an overview of evidence for measures where data were available for at least one type of reliability and does not include all measures identified as eligible for the review in the preliminary search. The full list of measures of BtC identified in the preliminary search can be found in Supplementary material 1.

[1] Reference numbers align with numbers of included papers listed in Supplementary material 10.
Reference numbers align with numbers of included papers listed in Supplementary material 10.
[2] Ratings based on COSMIN criteria for good measurement properties (Mokkink et al., 2018); "sufficient" = IC ≥0.70, IRR ≥0.60 or TRTR ≥0.70, "insufficient" = IC <0.70, IRR <0.60 or TRTR <0.70, "inconsistent" = pooled estimates range below to above 0.70 (IC or TRTR) or 0.60 (IRR) across measure subscales, "indeterminate" = No data for IC, IRR or TRTR were available.
[3] Ratings based on COSMIN quality of evidence criteria using the GRADE approach (Mokkink et al., 2018; Prinsen et al., 2018); "high" = very confident that the true measurement property lies close to that of the pooled estimate, "moderate" = moderately confident that the true measurement property is likely to be close to the pooled estimate, but there is a possibility that it is substantially different, "low" = limited confidence in the pooled estimate; the true measurement property may be substantially different from the pooled estimate, "very low" = very little confidence in the pooled estimate; the true measurement property may be substantially different from the pooled estimate.
[a] Inconsistent evidence as subscale pooled estimates for IRR range from <0.60 to ≥0.60, *or* subscale estimates for IC *or* TRTR range from <0.70 to ≥0.70 (see Table 2 for pooled estimates for all measures).
[b] IC data was derived from two analyses in paper 24.
[c] IRR data was derived from two analyses in paper 24 and two groups in paper 48.
[d] TRTR data was derived from two analyses in paper 48.
[e] IRR data was derived from two analyses in paper 64.
[f] Number of participants varied between CBI scales, $n = 14$ for physical aggression severity, $n = 10$ for self-injury severity, $n = 9$ for verbal aggression severity, $n = 8$ for inappropriate vocalisation severity, $n = 6$ for disruption of the environment severity.
[g] Number of participants from paper 20 varied for C-SHARP subscales; $n = 12$ for all problem scales, $n = 8$ for all provocation scales.
[h] Number of participants from paper 21 varied for C-SHARP subscales; $n = 22$ for all problem scales, $n = 7$ for the verbal aggression provocation scale, $n = 10$ for the bullying provocation scale and $n = 6$ for the physical aggression provocation scale.
[i] IC data was derived from two analyses in paper 14.
[j] IC data was derived from two analyses in paper 6.
[k] IRR data was derived from two analyses in paper 64.

recommended threshold of ≥0.60. A marked level of heterogeneity (Higgin's I^2 ≥75%) between reported IC estimates was identified for 3 measures, suggesting analyses of IRR estimates were biased by uncontrolled or confounding factors (see Table 6).

3.2.2.1. The impact of methodological bias. Subgroup analyses were conducted to assess the overall impact of study level risk of bias on IRR estimates for measures of function (see Table 7). Analyses were conducted for each of the eight types of methodological bias according to the COSMIN Risk of Bias Reliability Checklist. Significant differences in IRR estimates were observed for reliability criteria 1, 2, 3, 4 and 8. Significantly higher IRR estimates were observed across measures when participants were stable between informant ratings (criteria 1), when a shorter time interval (criteria 2) and similar test conditions (criteria 3) were used between informant ratings, and when there were flaws in the study design or statistical methods (criteria 8); however, IRR estimates were below the recommended ≥0.60 threshold for all risk of bias categories in each analysis. Significant differences attributable to the statistic used (criteria 4) were also observed, with IRR estimates exceeding the recommended ≥0.60 threshold when ICC were used to assess IRR compared to Pearson's or Spearman's correlations. See Supplementary material 4, Table D.2 for risk of bias ratings assigned to individual studies.

3.2.2.2. The impact of study level characteristics. Subgroup analyses were conducted to assess the impact of study level characteristics on IRR estimates of function measures where sufficient data were available (see Table 8 for an overview of subgroup analyses conducted; see Supplementary material 9 for all subgroup forests plots). Significant differences attributable to the informants completing the measure were observed in the IRR estimates for the MAS, with professional-professional informant pairs and teacher-professional informant pairs evidencing higher IRR than educator-educator pairs; however, the weighted average IRR estimates of all informant groups was below the recommended threshold of 0.60. No other significant differences in estimates were found.

3.2.3. Test-retest reliability

TRTR estimates were available for 6 measures of function from a total of 5 studies (see Table 6 for a summary of estimates for all measures). Forest plots presenting the IRR of all measures are shown in Supplementary material 8. Overall weighted average TRTR estimates ranged from 0.59 (95% CI 0.43–0.75) for the MAS to 0.98 (95% CI 0.96–0.99) for the QABF-Short form. Evidence of TRTR was limited to 1 study for all measures except the QABF (k = 15 effects from 3 studies). Overall weighted average TRTR estimates of the QABF met the recommended TRTR threshold (≥0.70); however, the *non-social* subscale failed to meet the threshold. A marked level of heterogeneity (Higgin's I^2 ≥75%) between reported TRTR estimates was identified for the FAST, suggesting analyses of TRTR estimates were biased by uncontrolled or confounding factors (see Table 6).

3.2.3.1. The impact of methodological bias. Subgroup analyses were conducted to assess the overall impact of study level risk of bias on TRTR estimates for measures of function. Analyses were conducted for each of the eight types of methodological bias according to the COSMIN Risk of Bias Reliability Checklist (see Table 7) Significant differences were observed when participants were stable between informant ratings (criteria 1), when a shorter time interval was used between test and retest (criteria 2). Significantly lower TRTR estimates were also observed across measures when there were flaws in the study design or statistical methods (criteria 8), such as differences in the behaviour being rated between test and retest (McAtee, Carr, & Schulte, 2004). See Supplementary material 4, Table D.2 for risk of bias ratings assigned to individual studies.

3.2.3.2. The impact of study level characteristics. Subgroup and meta-regression analyses were planned to assess the impact of study level characteristics on TRTR estimates; however, these analyses could not be conducted to identify causes of heterogeneity within TRTR estimates, due to the paucity of data.

3.2.4. The impact of publication and small study biases

Funnel plots were generated for studies assessing the IC, IRR and TRTR of measures of function where ≥10 effects were available. Accordingly, plots were generated for studies assessing the IC of the FACT, MAS and QABF, and IRR and TRTR of the QABF. There was evidence of heterogeneity in studies assessing IC for the MAS and QABF, and IRR and TRTR for the QABF; however, there was no clear evidence of publication bias as several small studies reported effects below the recommended thresholds for interpretability. Consequently, no simulation and adjustment for publication bias and small study effects was undertaken for these measures. Clear effects of heterogeneity were observed for studies assessing the IC of the FACT; therefore, the Trim and Fill procedure was conducted to correct for the effects of publication

Table 6
Overall weighted average internal consistency, inter-rater reliability and test-retest reliability estimates for measures of function using random-effects models.

Measure	Subscale	Internal consistency				Inter-rater reliability				Test-retest reliability			
		k	ARAW	95% CI	I^{2a}	k	COR	95% CI	I^{2a}	k	COR	95% CI	I^{2a}
CAI	Social/cultural	1	**0.91**	0.85–0.97	–	1	0.28	−0.13-0.69	–	1	0.61	0.33–0.89	–
	Task/activity	1	**0.91**	0.85–0.97	–	1	0.22	−0.21-0.65	–	1	**0.71**	0.49–0.93	–
	Physical	1	**0.78**	0.63–0.93	–	1	0.34	−0.06-0.74	–	1	0.57	0.27–0.87	–
	Biological	1	0.57	0.28–0.86	–	1	0.32	−0.08-0.72	–	1	0.67	0.42–0.92	–
	Total measure weighted average	4	**0.87**	0.79–0.94	60.00%	4	0.29	0.09–0.50	0.00%	4	0.65	0.52–0.78	0.00%
FACT	Attention	3	**0.93**	0.92–0.95	52.00%	1	**0.74**	0.66–0.82	–	1	**0.89**	0.85–0.93	–
	Escape	3	**0.93**	0.91–0.94	63.00%	1	**0.70**	0.61–0.79	–	1	**0.78**	0.78–0.88	–
	Physical	3	**0.95**	0.93–0.96	77.00%	1	**0.65**	0.55–0.75	–	1	**0.83**	0.83–0.91	–
	Sensory	3	**0.93**	0.90–0.96	94.00%	1	**0.79**	0.73–0.85	–	1	**0.84**	0.84–0.92	–
	Tangible	3	**0.95**	0.94–0.95	0.00%	1	**0.73**	0.65–0.81	–	1	**0.76**	0.76–0.88	–
	Total measure weighted average	15	**0.94**	0.93–0.95	79.00%	5	**0.73**	0.69–0.78	35.00%	5	**0.86**	0.84–0.89	0.00%
FAST	Social attention	1	0.05	−0.22-0.32	–	2	0.58	0.39–0.76	23.60%	1	0.57	0.45–0.69	–
	Social escape	1	0.12	−0.13-0.37	–	2	0.49	0.22–0.75	24.70%	1	**0.73**	0.65–0.81	–
	Automatic sensory	1	0.60	0.49–0.71	–	2	0.37	0.11–0.64	25.30%	1	**0.72**	0.64–0.80	–
	Automatic pain	1	**0.77**	0.70–0.84	–	2	0.46	0.35–0.57	26.40%	1	**0.82**	0.76–0.88	–
	Total measure weighted average	4	0.41	0.13–0.70	94.00%	8	0.48	0.39–0.57	80.00%	4	**0.72**	0.63–0.81	81.00%
MAS	Attention	9	**0.80**	0.74–0.87	92.00%	11	0.42	0.23–0.61	92.00%	1	0.52	0.19–0.85	–
	Escape	9	**0.75**	0.70–0.80	62.00%	11	0.43	0.27–0.59	80.00%	1	0.35	−0.04-0.74	–
	Sensory	9	**0.73**	0.69–0.77	40.00%	11	0.50	0.34–0.66	84.00%	1	**0.73**	0.52–0.94	–
	Tangible	9	**0.84**	0.81–0.87	64.00%	11	0.57	0.44–0.70	80.00%	1	0.53	0.21–0.85	–
	Total measure weighted average	36	**0.78**	0.75–0.81	85.00%	44	0.49	0.41–0.56	86.00%	4	0.59	0.43–0.75	12.30%
QABF	Attention	6	**0.89**	0.86–0.92	74.00%	7	0.51	0.33–0.69	90.00%	3	**0.80**	0.69–0.90	77.00%
	Escape	6	**0.87**	0.84–0.91	73.00%	7	0.55	0.42–0.68	80.00%	3	**0.74**	0.68–0.79	0.00%
	Non-social	6	**0.84**	0.80–0.88	73.00%	7	0.55	0.39–0.72	90.00%	3	**0.82**	0.72–0.92	81.00%
	Physical	6	**0.93**	0.90–0.95	86.00%	7	0.46	0.35–0.57	60.00%	3	0.67	0.54–0.81	67.00%
	Tangible	6	**0.89**	0.88–0.91	0.00%	7	0.59	0.41–0.77	94.00%	3	**0.83**	0.76–0.91	72.00%
	Total measure weighted average	30	**0.88**	0.87–0.90	87.00%	35	0.53	0.46–0.61	90.00%	15	**0.78**	0.74–0.82	74.00%
QABF-Short form	Attention	1	**0.92**	0.89–0.95	–	1	**0.96**	0.93–0.99	–	1	**0.98**	0.97–0.99	–
	Escape	1	**0.91**	0.87–0.95	–	1	**0.93**	0.89–0.97	–	1	**0.98**	0.97–0.99	–
	Non-social	1	**0.84**	0.78–0.90	–	1	**0.93**	0.89–0.97	–	1	**0.99**	0.98–1.00	–
	Physical	1	**0.94**	0.92–0.96	–	1	**0.82**	0.71–0.93	–	1	**0.95**	0.91–0.99	–
	Tangible	1	**0.79**	0.71–0.87	–	1	**0.93**	0.89–0.97	–	1	**0.84**	0.73–0.95	–
	Total measure weighted average	5	**0.89**	0.86–0.93	78.00%	5	**0.93**	0.91–0.96	50.00%	5	**0.98**	0.96–0.99	69.40%

Note. This table includes an overview of estimates for measures where data were available for at least one type of reliability and does not include all measures identified as eligible for the review in the preliminary search. The full list of measures of function identified in the preliminary search can be found in Supplementary material 1. Estimates meeting minimal interpretable criteria highlighted in bold. k = number of effects in weighted average estimate.
CAI=Contextual Assessment Inventory, FACT = Functional Assessment for Multiple CausaliTy, FAST = Functional Assessment Screening Tool, MAS = Motivation Assessment Scale, QABF = Questions About Behavioural Function Scale, QABF-Short Form = Questions About Behavioural Function-Short Form.
[a] = Higgin's I^2 not calculated when k = 1.

bias (Duval & Tweedie, 2000a, 2000b). An imputed IC estimate of 0.94 (95% CI 0.93–0.94) was indicated, representing a negligible −0.45% decrease relative to the original omnibus analysis.

3.2.5. *Overall quality of evidence for measures of function*

Ratings of the quality and level of evidence for pooled IC, IRR and TRTR estimates for each measure are shown in Table 9 (Mokkink et al., 2018; Prinsen et al., 2018). Moderate to high evidence of sufficient IC was available for the FACT, MAS, QABF and QABF-Short form. Very low evidence of inconsistent IC estimates was available for the CAI and FAST, due to inconsistency in IC estimates between subscales. Pooled IRR estimates were sufficient for the FACT and QABF-Short form, and insufficient for the CAI, FAST, MAS and QABF. Pooled TRTR estimates were sufficient for the FACT and QABF-Short form, and inconsistent for the CAI, FACT, MAS and QABF, due to inconsistency in estimates across subscales. The quality of evidence for IRR ranged from low to very low across measures, while the quality of evidence for TRTR was very low for all measures. Overall, levels of evidence for IC, IRR and TRTR were frequently downgraded due to risk of bias, inconsistency in pooled estimates across subscales or between studies, and low numbers of participants ($n < 100$) (Mokkink et al., 2018; Prinsen et al., 2018)).

4. Discussion

This is the first systematic review and meta-analytic study to quantitatively synthesise current evidence for the IC, IRR and TRTR of informant-report measures of BtC and behavioural function in ID populations. A total of 50 measures were identified for inclusion in the review through a rigorous and systematic preliminary search. Despite the large number of identified measures, the main search strategy revealed evidence for IC, IRR *and* TRTR was limited to 14 (28%) measures. No published evidence of IC, IRR *or* TRTR was identified for 20 (40%)

Table 7
Subgroup analyses for the effect of study level risk of bias on reported internal consistency, inter-rater reliability, and test-retest reliability of measures of function.

COSMIN risk of bias box 4: Internal consistency criteria:	Internal consistency					COSMIN risk of bias box 6: Reliability criteria:	Inter-rater reliability					Test-retest reliability			
	Rating [a]	k	ARAW	95% CI	X^2		Rating [a]	k	COR	95% CI	X^2	k	COR	95% CI	X^2
1. Statistic calculated for separate subscales	Very good	94	0.86	0.85–0.87	-[c]	1. Participant stability between ratings	Very good	0	-[b]	-[b]	13.77**	0	-[b]	-[b]	40.87**
	Adequate	–	–	–			Adequate	81	0.58	0.53–0.63		19	0.89	0.86–0.93	
	Doubtful	0	-[b]	-[b]			Doubtful	20	0.43	0.36–0.49		14	0.81	0.78–0.84	
	Inadequate	0	-[b]	-[b]			Inadequate	0	-[b]	-[b]		5	0.69	0.64–0.75	
	Not applicable	–	–	–			Not applicable	–	–	–		–	–	–	
2. Continuous scale statistic	Very good	75	0.93	0.91–0.94	-[c]	2. Time interval between ratings	Very good	8	0.43	0.38–0.49	12.57*	14	0.94	0.92–0.96	72.94**
	Adequate	–	–	–			Adequate	–	–	–		–	–	–	
	Doubtful	0	-[b]	-[b]			Doubtful	88	0.56	0.51–0.61		5	0.65	0.53–0.77	
	Inadequate	0	-[b]	-[b]			Inadequate	5	0.54	0.47–0.61		19	0.78	0.75–0.82	
	Not applicable	19	0.84	0.82–0.85			Not applicable	–	–	–		–	–	–	
3. Dichotomous scale statistic	Very good	19	0.84	0.82–0.85	-[c]	3. Similarity of test conditions	Very good	58	0.53	0.47–0.59	8.76*	8	0.80	0.70–0.91	1.41
	Adequate	–	–	–			Adequate	30	0.59	0.52–0.67		26	0.84	0.79–0.88	
	Doubtful	0	-[b]	-[b]			Doubtful	8	0.39	0.28–0.50		12	0.80	0.73–0.86	
	Inadequate	0	-[b]	-[b]			Inadequate	5	0.54	0.47–0.61		4	0.67	0.61–0.73	
	Not applicable	75	0.93	0.91–0.94			Not applicable	–	–	–		–	–	–	
4. IRT-based score statistic	Very good	0	-[b]	-[b]	-[c]	4. Continuous scale statistic	Very good	23	0.74	0.68–0.80	55.97**	0	-[b]	-[b]	20.24
	Adequate	–	–	–			Adequate	17	0.52	0.45–0.59		23	0.82	0.79–0.85	
	Doubtful	–	–	–			Doubtful	57	0.46	0.39–0.53		15	0.83	0.79–0.87	
	Inadequate	0	-[b]	-[b]			Inadequate	4	0.44	0.37–0.51		0	-[b]	-[b]	
	Not applicable	94	0.86	0.85–0.87			Not applicable	0	-[b]	-[b]		0	-[b]	-[b]	
5. Other design or statistical flaws	Very good	94	0.86	0.85–0.87	-[c]	5. Dichotomous, nominal, or ordinal scale statistic	Very good	0	-[b]	-[b]	-[c]	0	-[b]	-[b]	-[c]
	Adequate	–	–	–			Adequate	–	–	–		–	–	–	
	Doubtful	0	-[b]	-[b]			Doubtful	–	–	–		–	–	–	
	Inadequate	0	-[b]	-[b]			Inadequate	0	-[b]	-[b]		0	-[b]	-[b]	
	Not applicable	–	–	–			Not applicable	101	0.54	0.49–0.59		38	0.81	0.78–0.84	
						6. Ordinal scale weighted kappa	Very good	0	-[b]	-[b]	-[c]	0	-[b]	-[b]	-[c]
							Adequate	–	–	–		–	–	–	
							Doubtful	0	-[b]	-[b]		0	-[b]	-[b]	
							Inadequate	–	–	–		–	–	–	
							Not applicable	101	0.54	0.49–0.59		38	0.81	0.78–0.84	
						7. Ordinal scale weighting scheme	Very good	0	-[b]	-[b]	-[c]	0	-[b]	-[b]	-[c]
							Adequate	0	-[b]	-[b]		0	-[b]	-[b]	
							Doubtful	–	–	–		–	–	–	
							Inadequate	–	–	–		–	–	–	
							Not applicable	101	0.54	0.49–0.59		38	0.81	0.78–0.84	
						8. Other design or statistical flaws	Very good	97	0.55	0.50–0.59	5.63*	33	0.82	0.79–0.52	7.70*
							Adequate	–	–	–		–	–	–	
							Doubtful	4	0.29	0.09–0.50		5	0.65	0.53–0.77	
							Inadequate	0	-[b]	-[b]		0	-[b]	-[b]	
							Not applicable	–	–	–		–	–	–	

Note. * $p < .05$, ** $p < .001$.
[a] Not all rating levels are used for all COSMIN reliability/internal consistency criteria.
[b] ARAW/COR and 95% CI not reported where k < 4.
[c] [c] Not interpretable.

Table 8
Subgroup analyses for differences in overall weighted average estimates of internal consistency, inter-rater reliability and test-retest reliability estimates for measures of function, attributable to recruitment strategy, informant completing the measure, administration regarding children or adults, and method of administration.

Measure	Internal consistency				Inter-rater reliability				Test-retest reliability			
	Recruitment strategy	Informant	Administration regarding child or adult	Method of administration	Recruitment strategy	Informant	Administration regarding child or adult	Method of administration	Recruitment strategy	Informant	Administration regarding child or adult	Method of administration
CAI	–	–	–	–	–	–	–	–	–	–	–	–
FACT	–	–	–	–	–	–	–	–	–	–	–	–
FAST	–	–	–	–	–	–	–	–	–	–	–	–
MAS	+[a]	+[c]	–	–	o	+[f]	o	o	–	–	–	–
QABF	+[b]	+[d]	+[e]	o	o	–	–	–	–	–	–	–
QABF-Short form	–	–	–	–	–	–	–	–	–	–	–	–

Note. + = significant differences in estimates, o = no significant differences in estimates, – = subgroup analyses not conducted due to lack of available data.
See Supplementary material 9 for all subgroup analysis forest plots.
[a] Higher internal consistency estimates evidenced with recruitment from healthcare (k = 4, a = 0.77) and community-based (k = 24, a = 0.80) settings compared to school settings (k = 4, a = 0.67).
[b] Higher internal consistency estimates evidenced with recruitment from school (k = 5) and community-based (k = 20) settings compared to healthcare settings (k = 5); however, a ≥ 0.70 for all estimates.
[c] Higher internal consistency estimates when completed by professionals (k = 20, a = 0.80) or parents or caregivers (k = 4, a = 0.77) compared to educators (k = 4, a = 0.67).
[d] Higher internal consistency estimates when completed by professionals (k = 20) compared to parents or caregivers (k = 5); however, a ≥ 0.70 for all estimates.
[e] Higher internal consistency estimates when used to rate the behaviour of adults (k = 20) compared to children (k = 5); however, a ≥ 0.70 for all estimates.
[f] Higher inter-rater reliability estimates evidenced when completed by professional-professional (k = 24, r = 0.48) and educator-professional (k = 4, r = 0.50) rater pairs compared to educator-educator (k = 4, r = 0.14) rater pairs; however, all estimates are <0.60.

measures, and evidence for one or more types of reliability was limited to a single study for 14 (28%) measures.

4.1. Measures of BtC

Evidence of IC, IRR and TRTR was highly variable across measures of BtC. Data on at least one type of reliability were available for 23 (53%) measures of BtC. Where IC data were available, many measures of BtC met the minimum interpretable criteria for IC; however, estimates for some measures exceeded the maximum recommended alpha value (Tavakol & Dennick, 2011). Conversely, studies assessing the IRR and TRTR of BtC measures were not available or limited to a single study for the majority of measures, impacting the robustness of estimates. Based on the current available evidence, candidate BtC measures with the most evidence of reliability in children and adults with ID are the ABC/ABC-C *irritability* subscale, BPI-01, and BPI-Short Form. Meta-analytic syntheses of several studies revealed pooled estimates which exceeded recommended thresholds for the IC, IRR *and* TRTR of the BPI-01. The ABC/ABC-C *irritability* subscales and BPI-Short Form also evidenced pooled IC and TRTR estimates which exceeded recommended thresholds, and although pooled IRR estimates were below the recommended threshold for IRR, moderate IRR was evidenced (Koo & Li, 2016; Landis & Koch, 1977).

While these measures are shown to be reliable, it is of note that they vary in the breadth of BtC assessed. The ABC/ABC-C *irritability* subscale is arguably a broad scale which assesses a range of behaviours within a single scale, whereas the BPI-01 and BPI-Short Form are more focused, with several subscales to assess specific forms of BtC (e.g., topographies of self-injury and aggression/destruction) (Aman, Singh, Stewart, & Field, 1985; Rojahn et al., 2012; Rojahn et al., 2001). While broad measures of behaviour may be beneficial in capturing the breadth of behaviour and for screening and monitoring purposes, more focused measures might have higher value in precisely characterising behavioural presentations to inform person-centred intervention (Beavers et al., 2013; Oliver et al., 2003).

4.2. Measures of behavioural function

Fewer measures of behavioural function were identified compared to measures of BtC. Despite this, there was more published evidence for the IC, IRR and TRTR of measures of function; evidence for IC, IRR *and* TRTR was available for 6 (86%) of function measures. Based on the current available evidence, the FACT, QABF and QABF-Short Form are candidate measures of behavioural function with the most evidence of reliability in people with ID. Pooled estimates for the IC, IRR *and* TRTR of the FACT and QABF-Short Form exceeded recommended thresholds. However, IRR and TRTR estimates for these measures were based on a single study where the measures were completed about the behaviour of adults with ID. Consequently, these estimates should be interpreted cautiously and information on the reliability of these measures in children with ID is needed. The QABF evidenced pooled IC and TRTR estimates which exceeded recommended thresholds, and although pooled IRR estimates were below the recommended threshold for IRR, moderate IRR was evidenced (Koo & Li, 2016; Landis & Koch, 1977).

Measures of behavioural function typically included subscales for attention, escape, sensory/non-social and tangible functions. While these categories provide a good baseline for understanding behaviour, BtC is often a result of multifaceted person and environmental interactions, and consideration of wider contextual factors, as well as person characteristics that may underpin functions for BtC, e.g., anxiety and executive functioning, is important (Davies & Oliver, 2016; Waite et al., 2014). The Comprehensive Assessment of Triggers for Behaviours of Concern Scale (CATS) is a recently developed measure which assesses a broader range of contextual variables and antecedents for BtC; however, information on the measurement properties of the CATS is currently unavailable (Limbu et al., 2021). Overall, the reliability

Table 9
Summary of quality ratings for the measurement properties of each measure of function according to COSMIN criteria for good measurement properties and COSMIN criteria for overall quality of evidence.

Measure	Measurement property	Number of papers (reference number) [1]	Participants (*n*)	Pooled measurement property quality rating [2]	Quality of evidence [3]
CAI	IC	1 (46)	20	Indeterminate [a]	Very Low
	IRR	1 (46)	20	Insufficient	Very Low
	TRTR	1 (46)	20	Indeterminate [a]	Very Low
FACT	IC	3 (44, 82) [b]	494	Sufficient	High
	IRR	1 (82)	130	Sufficient	Low
	TRTR	1 (82)	130	Sufficient	Very Low
FAST	IC	1 (82)	130	Indeterminate [a]	Very Low
	IRR	2 (30, 82)	326	Insufficient	Very Low
	TRTR	1 (82)	130	Indeterminate [a]	Very Low
MAS	IC	9 (1, 8, 16, 23, 34, 35, 51, 70, 75)	888	Sufficient	Moderate
	IRR	11 (1, 16, 33, 35, 51, 70, 72, 75, 78, 83) [c]	456	Insufficient	Low
	TRTR	1 (70)	20	Indeterminate [a]	Very Low
QABF	IC	6 (23, 35, 52, 67, 70, 82)	535	Sufficient	Moderate
	IRR	7 (35, 41, 47, 52, 70, 82) [d]	475	Insufficient	Very Low
	TRTR	3 (47, 70, 82)	265	Indeterminate [a]	Very Low
QABF-Short form	IC	1 (74)	75	Sufficient	Moderate
	IRR	1 (74)	38	Sufficient	Very Low
	TRTR	1 (74)	29	Sufficient	Very Low

Note. This table includes an overview of evidence for measures where data were available for at least one type of reliability and does not include all measures identified as eligible for the review in the preliminary search. The full list of measures of function identified in the preliminary search can be found in Supplementary material 1.

[1] Reference numbers align with numbers of included papers listed in Supplementary material 10.
Reference numbers align with numbers of included papers listed in Supplementary material 10.

[2] Ratings based on COSMIN criteria for good measurement properties (Mokkink et al., 2018); "sufficient" = IC ≥0.70, IRR ≥0.60 or TRTR ≥0.70, "insufficient" = IC <0.70, IRR <0.60 or TRTR <0.70, "inconsistent" = pooled estimates range below to above 0.70 (IC or TRTR) or 0.60 (IRR) across measure subscales, "indeterminate" = No data for IC, IRR or TRTR were available.

[3] Ratings based on COSMIN quality of evidence criteria using the GRADE approach (Mokkink et al., 2018); "high" = very confident that the true measurement property lies close to that of the pooled estimate, "moderate" = moderately confident that the true measurement property is likely to be close to the pooled estimate, but there is a possibility that it is substantially different, "low" = limited confidence in the pooled estimate; the true measurement property may be substantially different from the pooled estimate, "very low" = very little confidence in the pooled estimate; the true measurement property may be substantially different from the pooled estimate.

[a] Inconsistent evidence as subscale pooled estimates for IRR range from <0.60 to ≥0.60, or subscale estimates for IC or TRTR range from <0.70 to ≥0.70 (see Table 6 for pooled estimates for all measures).

[b] IC data was derived from two analyses in paper 44.

[c] IRR data was derived from two analyses in paper 83.

[d] IRR data was derived from two analyses in paper 41.

estimates for measures of behavioural function may be influenced by behavioural presentation, the frequency of behaviour (Matson & Wilkins, 2009), or the extent of operationalisation to ensure the same behaviour is being rated between informants or test and retest administrations (McAtee et al., 2004). More frequent BtC may provide informants with more information to identify contingencies with increased reliability when compared to less frequent behaviour; therefore, lower reliability estimates obtained for some measures may be attributable to lower frequencies of behaviour topographies or ambiguity of subscales (Durand & Crimmins, 1988; Matson & Wilkins, 2009; Zarcone, Rodgers, Iwata, Rourke, & Dorsey, 1991).

4.3. The impact of study level characteristics

Given measurement properties are influenced by the population and context in which a measure is being used (Swan et al., 2023), the current meta-analytic study endeavoured to conduct subgroup and meta-regression analyses to examine the impact of study methodological quality and study level characteristics on measure IC, IRR and TRTR estimates. Studies assessing the measurement properties of measures often reported limited participant (e.g., age, sex, level of ID) and procedural information (e.g., informant/s, method of administration, time interval between measure administrations). Further, participant characteristics were frequently described at an overall participant level, and characteristics of specific subgroups of participants involved in smaller IRR and TRTR analyses were often unreported. This precluded the ability to conduct subgroup comparisons for many measures and, where subgroup analyses were conducted, they were typically limited to commonly used measures where adequate data were available (e.g., ABC/ABC-C, BPI-01 and QABF). However, some measures were excluded from some subgroup analyses based on their characteristics, such as being designed for specific informants (e.g., teacher or parent report) or populations (e.g., the A-SHARP is designed for use with adults and the C-SHARP is designed for use with children) (Farmer & Aman, 2009; Matlock & Aman, 2011).

Subgroup analyses based on the informant completing the measure revealed interesting differences in reliability estimates. For example, higher IRR estimates were found between educator-educator informant pairs for the BPI-01 and NCBRF *self-injury/stereotypic* subscale, compared to educator-parent or educator-professional informant pairs. These differences may be attributable to how well an informant knows a person and the settings in which a person's behaviour is observed. For example, educators (e.g., teachers and teaching assistants) typically observe a person's behaviour within a school setting, whereas parents or caregivers are more likely to observe more variable behaviour across a range of settings and environments. Despite this, it is important to note that weighted average reliability estimates in subgroup analyses were typically based on few studies and so the findings of such analyses should be interpreted cautiously.

4.4. Limitations

The current systematic review and meta-analytic study has several limitations. While the scale of the selection process enabled a comprehensive synthesis of published evidence for the reliability of measures in ID populations, evidence was not evenly distributed between measures,

and there was a large proportion of measures where no evidence for reliability was discovered or where available evidence was limited to a single study. Grey literature, including dissertations and tool manuals, were excluded due to the lack of rigorous peer-review. As such, it is possible that some properties reported within tool manuals were missed. Moreover, the review process disadvantages newly developed measures where measurement properties may accumulate in the future; however, if measures are to continue to be used it is important to ensure the evidence base underpinning them is robust. The heterogeneity of methods was enhanced where evidence of a measures IC, IRR or TRTR was limited to a single study, which may reduce confidence in the accuracy and robustness of estimates. Furthermore, overall measure reliability estimates were obtained by pooling multiple effects from subscales within a measure from a single study. The repetition of sample sizes in meta-analyses of measures with multiple subscales means that confidence intervals for overall reliability estimates may have been artificially deflated; therefore, overall reliability estimates should be interpreted cautiously.

The use of COSMIN criteria to assess the impact of methodological bias on measurement properties is a strength of the review, however, for inclusion in the meta-analysis, studies were required to assess measure IC using Cronbach's alpha or KR-20 values at measure subscale level. This meant risk of bias ratings were biased towards the 'low risk' category, limiting the ability for subgroup analysis to detect nuances in the impact of study level risk of bias on overall IC estimates. However, it should be noted that Cronbach's alpha has faced criticism as a standalone coefficient of IC (Cortina, 1993; Cortina et al., 2020; McNeish, 2018; Sijtsma, 2009; Sijtsma & Pfadt, 2021). For instance, large alpha coefficients are often misinterpreted to imply the structural validity of a measure, when structural validity may not have been thoroughly investigated (Sijtsma & Pfadt, 2021). COSMIN guidelines differentiate IC from structural validity, defining IC as "the degree of interrelatedness among the items" (Mokkink et al., 2010). The guidelines further emphasise that evidence of structural validity, such as factor analyses, is necessary for the clear interpretation of IC statistics (Mokkink et al., 2018). Therefore, IC estimates within this review should be interpreted with consideration of wider evidence of the structural validity of each measure.

In addition, this review focused on measures in English, and translated versions of measures were not included, given language and cultural differences can impact measure performance (Wild et al., 2005). Furthermore, while the current study focused on properties of reliability, consideration of the content, construct, and criterion validity of measures and their responsiveness to change is also important if measures are to be recommended for clinical and research purposes (Swan et al., 2023). Although the scope of the current review and meta-analysis did not extend to validity and responsiveness to change, a systematic review of the criterion validity of some measures of BtC has been conducted (Turton, 2015), and several systematic reviews in specific populations or informants have included validity (Howell et al., 2021; McConachie et al., 2015; Reyes-Martín et al., 2022). Future systematic reviews and meta-analytic studies to assess the validity and responsiveness to change of measures of BtC and function would be informative.

4.5. Future directions

Significant variability in the evidence base underpinning the reliability of measures of BtC and behavioural function is highlighted by the current review. For many measures, the quality of evidence for IC, IRR and TRTR was deemed 'low' or 'very low' according to COSMIN criteria (Mokkink et al., 2018; Prinsen et al., 2018). Consequently, further evaluation of the IC, IRR and TRTR of existing measures is a priority for future research. Given the large number of identified measures relative to studies assessing their measurement properties, it could be argued that efforts might be best served in refining and understanding the measurement properties of existing measures, as an alternative to developing new measures that assess similar domains.

Generating data that contribute to understanding the measurement properties of existing measures in ID populations will enable future meta-analyses to include a larger number of studies. This may involve assessing measurement properties within studies, such as incorporating additional components to assess measure IC, IRR and/or TRTR into studies using measures of BtC or behavioural function. In addition, multi-institutional collaborations would generate opportunities for researchers and clinicians to work together to refine measures and better understand their measurement properties. For instance, clinicians may support studies aiming to evaluate measurement properties by administering measures within clinical settings and services. Such collaborations might also facilitate sharing of clinical expertise and insights, informing researchers with ways in which to improve and refine existing measures to ensure their relevance, practicality, and sensitivity to the needs of specific ID groups. Studies assessing the measurement properties of measures should strive to adhere to reporting guidelines (e.g., COSMIN guidelines) to increase transparency, methodological quality, and the interpretability of findings (Gagnier, Lai, Mokkink, & Terwee, 2021). Thorough reporting of study level characteristics, such as participant and procedural information, will enable the impact on measurement properties to be more thoroughly assessed. This would facilitate future recommendations on suitable and robust measures for different settings and contexts, e.g., for use with children or adults with ID, specific informants, or according to level of ID.

While a thorough assessment of the measurement properties of each measure would facilitate future recommendations on the suitability of each measure, it should be noted that the relevance of certain measurement properties may vary based on the setting and context of a measure's planned use. Therefore, careful consideration of a measure's measurement properties and their relevance to the intended setting and context is necessary to ensure differences in reported BtC and behavioural functions are meaningful and not solely attributed to measurement error. For example, IRR may be more relevant in clinical settings where comprehensive behavioural assessments involve a measure being completed by multiple informants, however, may have less relevance in contexts where a measure is completed by the same informant over multiple occasions. Conversely, TRTR may have greater relevance when the same informant completes a measure on multiple occasions to monitor behaviour overtime. As such, TRTR may be particularly relevant in clinical or community settings where an assessment of the effectiveness of interventions or person-centred support requires an evaluation of whether there are meaningful changes in behaviour over time.

5. Conclusions

In conclusion, the current systematic review and meta-analytic study provides a synthesis of the IC, IRR and TRTR of measures of BtC and function, specifically in ID populations. The findings provide guidance on the quality of the current evidence base underpinning the reliability of measures which may be used to characterise and monitor BtC and behavioural function, as recommended by NICE clinical guidelines (NCCMH, 2015). However, the lack of evidence for many identified measures is striking. While this review is an important first step towards quantifying measurement properties to inform future recommendations for measures to assess BtC and behavioural function in people with ID in research and clinical practice, a key priority is for future research to continue to evaluate the IC, IRR and TRTR of existing measures. Systematic reviews and meta-analytic studies to examine other measurement properties, such as validity and sensitivity to change, are also required for measure recommendations to be made. Based on current available evidence, the following measures of BtC - ABC/ABC-C *irritability* subscale, BPI-01, and BPI-Short Form – and the following measures of behavioural function - FACT, QABF and QABF-Short Form - hold the

most evidence of IC, IRR and TRTR in people with ID.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cpr.2024.102434.

Role of funding sources

This work was supported by The Baily Thomas Charitable Fund, grant number 5682–8803. The funders had no role in the study design, data collection, analysis and interpretation of data, preparation of the manuscript or the decision to submit the paper for publication.

Contributors

LS, JW, JT, HC, and CR conceived and designed the study. LS conducted electronic database searching. LS, EP and AP selected papers for inclusion in the study. LS, AP, AW, and CG extracted data from included studies. LS and EP conducted ratings of study methodological bias. LS and CJ completed the statistical analyses. LS wrote the drafts of the manuscript, CJ, JT, HC, CR and JW provided feedback on revisions of the draft manuscript.

Declaration of competing interest

All authors declare that they have no conflicts of interest.

Data availability

Data availability is not applicable to this article as no new data were created or analysed in this study.

Acknowledgements

We are extremely grateful to The Baily Thomas Charitable Fund for funding the study. We wish to thank Poonam Virdee, Esther Smith and Megan Bird who assisted with the preliminary searches.

References

Adams, D., Clarke, S., Griffith, G., Howlin, P., Moss, J., Petty, J., … Oliver, C. (2018). Mental health and well-being in mothers of children with rare genetic syndromes showing chronic challenging behavior: A cross-sectional and longitudinal study. *American Journal on Intellectual and Developmental Disabilities, 123*(3), 241–253. https://doi.org/10.1352/1944-7558-123.3.241

Alter, P. J., Conroy, M. A., Mancil, G. R., & Haydon, T. (2008). A comparison of functional behavior assessment methodologies with young children: Descriptive methods and functional analysis. *Journal of Behavioral Education, 17*(2), 200–219. https://doi.org/10.1007/s10864-008-9064-3, 2008/06/01.

Aman, M. G., Singh, N. N., Stewart, A. W., & Field, C. (1985). The aberrant behavior checklist: a behavior rating scale for the assessment of treatment effects. *American journal of mental deficiency, 89*(5), 485–491.

Arron, K., Oliver, C., Moss, J., Berg, K., & Burbidge, C. (2011). The prevalence and phenomenology of self-injurious and aggressive behaviour in genetic syndromes. *Journal of Intellectual Disability Research, 55*(2), 109–120. https://doi.org/10.1111/j.1365-2788.2010.01337.x

Baker, P., & Daynes, S. (2010). Outcome measurement for people with intellectual disability who present challenging behaviour. *Advances in Mental Health and Intellectual Disabilities, 4*(2), 13–19. https://doi.org/10.5042/amhid.2010.0314

Beavers, G. A., Iwata, B. A., & Lerman, D. C. (2013). Thirty years of research on the functional analysis of problem behavior. *Journal of Applied Behavior Analysis, 46*(1), 1–21. https://doi.org/10.1002/jaba.30

Carr, E. G., & Durand, V. M. (1985). Reducing behavior problems through functional communication training. *Journal of Applied Behavior Analysis, 18*(2), 111–126. https://doi.org/10.1901/jaba.1985.18-111

Chan, J. S. L., & Chien, W. T. (2017). A randomised controlled trial on evaluation of the clinical efficacy of massage therapy in a multisensory environment for residents with severe and profound intellectual disabilities: A pilot study. *Journal of Intellectual Disability Research, 61*(6), 532–548. https://doi.org/10.1111/jir.12377. , Jun.

Chung, J. C. Y., Lowenthal, R., Mevorach, C., Paula, C. S., Teixeira, M. C. T. V., & Woodcock, K. A. (2022). Cross-cultural comparison of the contexts associated with emotional outbursts. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-022-05708-7, 2022/08/19.

Cooper, S. A., Smiley, E., Allan, L. M., Jackson, A., Finlayson, J., Mantry, D., & Morrison, J. (2009, Mar). Adults with intellectual disabilities: Prevalence, incidence and remission of self-injurious behaviour, and related factors. *Journal of Intellectual Disability Research, 53*, 200–216. https://doi.org/10.1111/j.1365-2788.2008.01060.x

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., … Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the *Journal of Applied Psychology*. *Journal of Applied Psychology, 105*(12), 1351–1381. https://doi.org/10.1037/apl0000815

Crawford, H., Karakatsani, E., Singla, G., & Oliver, C. (2019, Jul). The persistence of self-injurious and aggressive behavior in males with fragile X syndrome over 8 years: A longitudinal study of prevalence and predictive risk markers. *Journal of Autism and Developmental Disorders, 49*(7), 2913–2922. https://doi.org/10.1007/s10803-019-04002-3

Davies, L. E., & Oliver, C. (2014, Sep). The purported association between depression, aggression, and self-injury in people with intellectual disability: A critical review of the literature. *Ajidd-American Journal on Intellectual and Developmental Disabilities, 119*(5), 452–471. https://doi.org/10.1352/1944-7558-119.5.452

Davies, L. E., & Oliver, C. (2016, Feb-Mar). Self-injury, aggression and destruction in children with severe intellectual disability: Incidence, persistence and novel, predictive behavioural risk markers. *Research in Developmental Disabilities, 49-50*, 291–301. https://doi.org/10.1016/j.ridd.2015.12.003

Deb, S., Thomas, M., & Bright, C. (2001). Mental disorder in adults with intellectual disability. 2: The rate of behaviour disorders among a community-based population aged between 16 and 64 years. *Journal of Intellectual Disability Research, 45*(6), 506–514. https://doi.org/10.1046/j.1365-2788.2001.00373.x

Durand, M. V., & Crimmins, D. B. (1988). Identifying the variables maintaining self-injurious behavior. *Journal of Autism and Developmental Disorders, 18*(1), 99–117. https://doi.org/10.1007/BF02211821, 1988/03/01.

Durand, V. M., & Crimmins, D. B. (1992). *The motivation assessment scale (MAS) administration guide*. Monaco and Associates.

Duval, S., & Tweedie, R. (2000a). A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–98. https://doi.org/10.1080/01621459.2000.10473905, 2000/03/01.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Emerson, E. (2001). *Challenging behaviour: Analysis and intervention in people with severe intellectual disabilities*. Cambridge University Press.

Emerson, E., Einfeld, S., & Stancliffe, R. J. (2011, Nov). Predictors of the persistence of conduct difficulties in children with cognitive delay. *Journal of Child Psychology and Psychiatry, 52*(11), 1184–1194. https://doi.org/10.1111/j.1469-7610.2011.02413.x

Emerson, E., Kiernan, C., Alborz, A., Reeves, D., Mason, H., Swarbrick, R., Mason, L., & Hatton, C. (2001). Predicting the persistence of severe self-injurious behavior. *Research in Developmental Disabilities, 22*(1), 67–75. https://doi.org/10.1016/s0891-4222(00)00062-7

Farmer, C., Butter, E., Mazurek, M. O., Cowan, C., Lainhart, J., Cook, E. H., … Aman, M. (2015, Apr). Aggression in children with autism spectrum disorders and a clinic-referred comparison group. *Autism, 19*(3), 281–291. https://doi.org/10.1177/1362361313518995

Farmer, C. A., & Aman, M. G. (2009). Development of the children's scale of hostility and aggression: Reactive/proactive (C-SHARP). *Research in Developmental Disabilities, 30*(6), 1155–1167. https://doi.org/10.1016/j.ridd.2009.03.001

Floyd, R. G., Phaneuf, R. L., & Wilczynski, S. M. (2005). Measurement properties of indirect assessment methods for functional behavioral assessment: A review of research. *School Psychology Review, 34*(1), 58–73. ://WOS:000228068400006.

Frazier, T. W., Khaliq, I., Scullin, K., Uljarevic, M., Shih, A., & Karpur, A. (2023). Development and psychometric evaluation of the Open-Source Challenging Behavior Scale (OS-CBS). *Journal of Autism and Developmental Disorders, 53*(12), 4655–4670. https://doi.org/10.1007/s10803-022-05750-5

Gagnier, J. J., Lai, J., Mokkink, L. B., & Terwee, C. B. (2021). COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research, 30*(8), 2197–2218. https://doi.org/10.1007/s11136-021-02822-4, 2021/08/01.

Grey, I., Pollard, J., McClean, B., MacAuley, N., & Hastings, R. (2010, Nov). Prevalence of psychiatric diagnoses and challenging behaviors in a community-based population of adults with intellectual disability. *Journal of Mental Health Research in Intellectual Disabilities, 3*(4), 210–222. https://doi.org/10.1080/19315864.2010.527035

Hanley, G. P. (2012). Functional assessment of problem behavior: Dispelling myths, overcoming implementation obstacles, and developing new lore. *Behavior Analysis in Practice, 5*(1), 54–72. ://MEDLINE:23326630.

Hastings, R. P. (2002). Parental stress and behaviour problems of children with developmental disability. *Journal of Intellectual & Developmental Disability, 27*(3), 149–160. https://doi.org/10.1080/1366825021000008657, 2002/01/01.

Healy, O., Brett, D., & Leader, G. (2013). A comparison of experimental functional analysis and the Questions About Behavioral Function (QABF) in the assessment of challenging behavior of individuals with autism. *Research in Autism Spectrum Disorders, 7*(1), 66–81. https://doi.org/10.1016/j.rasd.2012.05.006, 2013/01/01/.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Holden, B., & Gitlesen, J. P. (2006, Jul-Aug). A total population study of challenging behaviour in the county of Hedmark, Norway: Prevalence, and risk markers.

Research in Developmental Disabilities, 27(4), 456–465. https://doi.org/10.1016/j.ridd.2005.06.001

Howell, M., Bradshaw, J., & Langdon, P. E. (2021, Mar). A systematic review of behaviour-related outcome assessments for children on the autism spectrum with intellectual disabilities in education settings. *Review Journal of Autism and Developmental Disorders, 8*(1), 67–91. https://doi.org/10.1007/s40489-020-00205-y

Kearney, C. A., Cook, L. C., Chapman, G., & Bensaheb, A. (2006). Exploratory and confirmatory factor analyses of the motivation assessment scale and resident choice assessment scale. *Journal of Developmental and Physical Disabilities, 18*(1), 1–11. https://doi.org/10.1007/s10882-006-9000-1, 2006/03/01.

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. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lecavalier, L., Leone, S., & Wiltz, J. (2006, Mar). The impact of behaviour problems on caregiver stress in young people with autism spectrum disorders. *Journal of Intellectual Disability Research, 50*, 172–183. https://doi.org/10.1111/j.1365-2788.2005.00732.x

Limbu, B., Unwin, G., & Deb, S. (2021). Comprehensive Assessment of Triggers for Behaviours of Concern Scale (CATS): Initial development. *International Journal of Environmental Research and Public Health, 18*(20), 10674. https://www.mdpi.com/1660-4601/18/20/10674.

Lloyd, B. P., & Kennedy, C. H. (2014). Assessment and treatment of challenging behaviour for individuals with intellectual disability: A research review. *Journal of Applied Research in Intellectual Disabilities, 27*(3), 187–199. https://doi.org/10.1111/jar.12089

Madsen, E. K., Peck, J. A., & Valdovinos, M. G. (2016, Sep). A review of research on direct-care staff data collection regarding the severity and function of challenging behavior in individuals with intellectual and developmental disabilities. *Journal of Intellectual Disabilities, 20*(3), 296–306. https://doi.org/10.1177/1744629515612328

Maguire, S., Davison, J., McLaughlin, M., Simms, V., & Bunting, B. (2023). Exploring the psychometric properties of self-report instruments used to measure health-related quality of life and subjective wellbeing of adolescents with intellectual disabilities: A Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) systematic review. *Journal of Applied Research in Intellectual Disabilities, 17*(1). https://doi.org/10.1111/jar.13110

Matlock, S. T., & Aman, M. G. (2011). Development of the adult scale of hostility and aggression: Reactive–proactive (A-SHARP). *American Journal on Intellectual and Developmental Disabilities, 116*(2), 130–141. https://doi.org/10.1352/1944-7558-116.2.130

Matson, J. L., & Vollmer, T. (2007). *Questions about behavioral function: QABF*. Disability Consultants, LLC.

Matson, J. L., & Wilkins, J. (2009, Mar). Factors associated with the questions about behavior function for functional assessment of low and high rate challenging behaviors in adults with intellectual disability. *Behavior Modification, 33*(2), 207–219. https://doi.org/10.1177/0145445508320342

Matson, J. L., & Williams, L. W. (2014). Functional assessment of challenging behavior. *Current Developmental Disorders Reports, 1*(2), 58–66. https://doi.org/10.1007/s40474-013-0006-y, 2014/06/01.

McAtee, M., Carr, E. G., & Schulte, C. (2004). A contextual assessment inventory for problem behavior: Initial development. *Journal of Positive Behavior Interventions, 6*(3), 148–165. Sum https://doi.org/10.1177/10983007040060030301.

McConachie, H., Parr, J. R., Glod, M., Hanratty, J., Livingstone, N., Oono, I. P., … Williams, K. (2015, Jun). Systematic review of tools to measure outcomes for young children with autism spectrum disorder. *Health Technology Assessment, 19*(41), 1. https://doi.org/10.3310/hta19410

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

National Collaborating Centre for Mental Health UK (NCCMH). (2015). *Challenging behaviour and learning disabilities: Prevention and interventions for people with learning disabilities whose behaviour challenges (Vol. 8)*. NICE Guideline. No. 11. Available from: https://www.ncbi.nlm.nih.gov/books/NBK355385/.

Nicholson, J., Konstantinidi, E., & Furniss, F. (2006, May-Jun). On some psychometric properties of the questions about behavioral function (QABF) scale. *Research in Developmental Disabilities, 27*(3), 337–352. https://doi.org/10.1016/j.ridd.2005.04.001

Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3d ed.). New York: McGraw-Hill Book Company.

Oliver, C., McClintock, K., Hall, S., Smith, M., Dagnan, D., & Stenfert-Kroese, B. (2003, Mar). Assessing the severity of challenging behaviour: Psychometric properties of the challenging behaviour interview. *Journal of Applied Research in Intellectual Disabilities, 16*(1), 53–61. https://doi.org/10.1046/j.1468-3148.2003.00145.x

Oliver, C., & Richards, C. (2015). Practitioner review: Self-injurious behaviour in children with developmental delay. *Journal of Child Psychology and Psychiatry, 56*(10), 1042–1054. https://doi.org/10.1111/jcpp.12425

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews* (p. 372). Bmj.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007, Dec). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills, 105*(3 Pt 1), 997–1014. https://doi.org/10.2466/pms.105.3.997-1014

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3, 2018/05/01.

Reyes-Martín, J., Simó-Pinatella, D., & Font-Roura, J. (2022). Assessment of challenging behavior exhibited by people with intellectual and developmental disabilities: A systematic review. *International Journal of Environmental Research and Public Health, 19*(14), 8701. https://www.mdpi.com/1660-4601/19/14/8701.

Rojahn, J., & Helsel, W. J. (1991). The aberrant behavior checklist with children and adolescents with dual diagnosis. *Journal of Autism and Developmental Disorders, 21*(1), 17–28. https://doi.org/10.1007/BF02206994

Rojahn, J., Matson, J. L., Lott, D., Esbensen, A. J., & Smalls, Y. (2001, Dec). The Behavior Problems Inventory: An instrument for the assessment of self-injury, stereotyped behavior, and aggression/destruction in individuals with developmental disabilities. *Journal of Autism and Developmental Disorders, 31*(6), 577–588. https://doi.org/10.1023/a:1013299028321

Rojahn, J., Rick-Betancourt, B., Barnard-Brak, L., & Moore, L. (2017). An independent investigation into the psychometric properties of the Adult Scale of Hostility and Aggression (A-SHARP). *Journal of Mental Health Research in Intellectual Disabilities, 10*(4), 253–266. https://doi.org/10.1080/19315864.2017.1299266

Rojahn, J., Rowe, E., Sharber, A., Hastings, R., Matson, J., Didden, R., Kroes, D., & Dumont, E. (2012, May). The behavior problems inventory-short form for individuals with intellectual disabilities: Part I: Development and provisional clinical reference data [Empirical Study; Quantitative Study]. *Journal of Intellectual Disability Research, 56*(5), 527–545. https://doi.org/10.1111/j.1365-2788.2011.01507.x

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*(3), 301–317. https://doi.org/10.1177/096228029800700306

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika, 86*(4), 843–860. https://doi.org/10.1007/s11336-021-09789-8

Simo-Pinatella, D., Mumbardo-Adam, C., Alomar-Kurz, E., Sugai, G., & Simonsen, B. (2019, Sep). Prevalence of challenging behaviors exhibited by children with disabilities: Mapping the literature. *Journal of Behavioral Education, 28*(3), 323–343. https://doi.org/10.1007/s10864-019-09326-9

Swan, K., Speyer, R., Scharitzer, M., Farneti, D., Brown, T., Woisard, V., & Cordier, R. (2023). Measuring what matters in healthcare: A practical guide to psychometric principles and instrument development. *Frontiers in Psychology, 14*, 1225850. https://doi.org/10.3389/fpsyg.2023.1225850

Tavakol, M., & Dennick, R. (2011, Jun 27). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, L., Oliver, C., & Murphy, G. (2011, Mar). The chronicity of self-injurious behaviour: A long-term follow-up of a total population study. *Journal of Applied Research in Intellectual Disabilities, 24*(2), 105–117. https://doi.org/10.1111/j.1468-3148.2010.00579.x

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research, 18*(8), 1115–1123. https://doi.org/10.1007/s11136-009-9528-5, 2009/10/01.

Tufanaru, C., Munn, Z., Stephenson, M., & Aromataris, E. (2015). Fixed or random effects meta-analysis? Common methodological issues in systematic reviews of effectiveness. *JBI Evidence Implementation, 13*(3), 196–207. https://doi.org/10.1097/xeb.0000000000000065

Turton, R. W. (2015, Mar). Criterion-related validity of challenging behaviour scales: A review of evidence in the literature. *Journal of Applied Research in Intellectual Disabilities, 28*(2), 81–97. https://doi.org/10.1111/jar.12098

Waite, J., Heald, M., Wilde, L., Woodcock, K., Welham, A., Adams, D., & Oliver, C. (2014). The importance of understanding the behavioural phenotypes of genetic syndromes associated with intellectual disability. *Paediatrics and Child Health, 24*(10), 468–472. https://doi.org/10.1016/j.paed.2014.05.002v

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for Patient-Reported Outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health, 8*(2), 94–104. https://doi.org/10.1111/j.1524-4733.2005.04054.x, 2005/03/01/.

Wilde, L., Wade, K., Eden, K., Moss, J., de Vries, P. J., & Oliver, C. (2018, Dec). Persistence of self-injury, aggression and property destruction in children and adults with tuberous sclerosis complex. *Journal of Intellectual Disability Research, 62*(12), 1058–1071. https://doi.org/10.1111/jir.12472

Zarcone, J., Napolitano, D., & Valdovinos, M. (2008, Dec). Measurement of problem behaviour during medication evaluations. *Journal of Intellectual Disability Research, 52*, 1015–1028. https://doi.org/10.1111/j.1365-2788.2008.01109.x

Zarcone, J. R., Rodgers, T. A., Iwata, B. A., Rourke, D. A., & Dorsey, M. F. (1991). Reliability analysis of the motivation assessment scale: A failure to replicate. *Research in Developmental Disabilities, 12*(4), 349–360. https://doi.org/10.1016/0891-4222(91)90031-M, 1991/01/01/.