Measuring bonding or attachment in the parent-infant-relationship: a systematic review of parent-report assessment measures, their psychometric properties and clinical utility

da Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL, UK.
b Greater Manchester Mental Health NHS Foundation Trust, Department of Clinical Psychology, Laureate House, Wythenshawe Hospital, Southmoor Road, Wythenshawe, Manchester M23 9LT, UK.
c MRC/CSO Social and Public Health Sciences Unit, Institute of Health and Wellbeing, University of Glasgow, 200 Renfield Street, Glasgow G2 3AX, UK.

Prepared for Clinical Psychology Review

*Corresponding author at: Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, 2nd Floor Zochonis Building, Brunswick Park, Manchester M13 9PL, UK.
Abstract

**Background:** Meaningful, valid and reliable self-report measures can facilitate the identification of important parent-infant-relationship factors, relevant intervention development and subsequent evaluation in community and clinical contexts. We aimed at identifying all available parent-report measures of the parent-infant-relationship or bond and to appraise their psychometric and clinimetric properties.

**Method:** A systematic review (PROSPERO: CRD42017078512) was conducted using the 2018 COSMIN criteria. Eight electronic databases were searched. Papers describing the development of self-report measures of the parent-infant-bond, attachment or relationship from pregnancy until two years postpartum or the assessment of their psychometric properties were included.

**Results:** Sixty-five articles evaluating 17 original measures and 13 modified versions were identified and reviewed. The studies’ methodological quality (risk of bias) varied between ‘very good’ and ‘inadequate’ depending on the measurement property assessed; however, scale development studies were mostly of ‘inadequate’ quality. Although most measures had good clinical utility, the psychometric evaluation of their properties was largely poor. The original or modified versions of the Postpartum Bonding Questionnaire collectively received the strongest psychometric evaluation ratings with high quality of evidence.

**Conclusions:** This novel review revealed that only a few antenatal and postnatal measures demonstrated adequate psychometric properties. Further studies are needed to determine the most robust perinatal measures for researchers and clinicians.

**Keywords:** Measurement; reliability; validity; mothers; fathers; quality assessment.
Highlights

• Parental perceptions of the parent-infant-bond are important in identifying any difficulties and strengths.

• This comprehensive review is the first to systematically assess the psychometric properties of 14 antenatal measures (eight original and six modified versions) and 18 postnatal measures (ten original measures and eight modified versions), of which the Pre- and Postnatal Bonding Scale can be used antenatally and postnatally.

• The Postpartum Bonding Questionnaire-25 (PBQ), together with its modified versions (PBQ-22, PBQ-14), was the only measure to show sufficient evidence for structural validity, internal consistency and reliability and was the most researched compared to other measures.

• The clinimetric and administrative properties were good for most measures, suggesting their feasibility, acceptability and attainability.

• Although several studies reported on validity and reliability of measures, most measures lacked adequate methodological quality according to risk of bias; thus, measurement developers and evaluators are advised to follow the COSMIN guidelines to report their findings adequately.
1. Introduction

A large body of research now confirms that the early stages of the parent and infant relationship exert an important influence over a child’s future development, psychological wellbeing and life chances (Ainsworth, 1979; Bowlby, 1982) and infancy is considered to be the most cost-effective time to intervene (Doyle, Harmon, Heckman, & Tremblay, 2009). Consequently, various organizations and government reports have been advocating the need to address the early stages of parenting with the intent of strengthening the early parent-infant-relationship (e.g., Allen, 2011; Ellyatt, 2017; Moullin, Waldfogel & Washbrook, 2014; NHS England, 2018; NICE, 2013; Public Health England, 2019; WHO, 2014; Wright et al., 2015).

A good early parent-infant-relationship, in which the parents are sensitive and responsive to their infant’s physical and emotional needs, lays the foundation for a child’s future self-esteem and resilience, their ability to regulate their emotions and their capacity to form close relationships (Bowlby, 1979, 1988; Thompson, 2000; Winston & Chicot, 2016; Wright et al., 2015). Conversely, poor early relationships place children at increased risk of poor cognitive, social and emotional outcomes (Leclère et al., 2014; van Ijzendoorn, Schuengel & Bakermans-Kranenburg, 1999; Winston & Chicot, 2016; Wright et al., 2015).

Given the importance of the early parent-child-relationship and emotional bond, it is paramount to identify how to support parents in strengthening or improving this relationship effectively when there are any difficulties. An important step in doing so is to be able to identify parents who may be struggling to bond with their developing fetus and/or baby in order to offer them an increased level of support (Royal College of Psychiatrists, 2018).

However, the prevalence of difficulties in the early parent-child-relationship can vary depending on how and what is being measured, with some researchers examining bonding via questionnaires (Condon & Corkindale, 1998; van Bussel, Spitz, & Demyttenaere, 2010a; Wittkowski, Wieck, & Mann, 2007) and this emotional bond or the reciprocal and interactive
relationship between parent and infant often referred to as attachment via observation others assessing attachment, and its reciprocal and interactive nature, via observation (Ainsworth, Blehar, Waters, & Wall, 1978; Crittenden, 2001; Bicking Kinsey, & Hupcey, 2013; Lotzin et al., 2015; Noorlander, Bergink, & van den Berg, 2008). Hereby, it is imperative to define the terms ‘bonding’ and ‘attachment’ because they are seen as different concepts but often used synonymously (e.g., Benoit, 2004; Bicking Kinsey & Hupcey, 2013; Redshaw & Martin, 2013). Bonding is described as the tie from the parent to the infant (Bicking Kinsey & Hupcey, 2013; Kennell & McGrath, 2005); it generally consists of feelings and emotions that parents experience towards their infant (Bicking Kinsey & Hupcey, 2013). Attachment is seen as the interplay and reciprocity between the parent and the child (Bicking Kinsey & Hupcey, 2013; Kennell & McGrath, 2005), which usually develops during pregnancy between the parent and the fetus (Condon & Corkindale, 1997). Attachment is part of the parent-child-relationship whereby the parent’s role is to ensure the safety, security and protection of the child (Bowlby, 1982). Since the concepts of ‘bonding’ and ‘attachment’ are closely related and have been widely researched, we have opted to include both within the term ‘parent-infant-relationship’.

The ‘gold standard’ for the assessment of parent-child-attachment, and as such the reciprocal aspect of the parent-child-relationship, is via the use of behavioural, observational measures used with parents or other caregivers and their children over 1 year old, such as the Strange Situation task (Ainsworth et al., 1978) and the Attachment Q-Set (Waters & Deane, 1985). Several observational assessment tools exist to evaluate attachment and interaction behaviours between parent and child (up to 30 months old) (e.g., for reviews see Gridley et al., 2019; Lotzin et al., 2015; Mesman & Emmen, 2013; Tryphonopoulos, Letourneau, & Ditommaso, 2014). However, these measures have two key limitations. Firstly, they are time- and resource-intensive and require extensive training to administer and interpret. This limits their use by practitioners, for example, in obstetric, pediatric or primary-care services which
mostly lack the time, facilities and training to administer these assessments (van Bussel et al., 2010a).

Whilst there may always be challenges in enabling parents to disclose any difficulties in bonding or forming emotional ties with their developing fetus or infant to healthcare professionals during routine appointments because parents fear being stigmatised (Morsbach & Prinz, 2006) and/or the involvement of social services and the potential loss of custody (NICE, 2014), developing reliable, valid and sensitive measures may be useful in assisting with the assessment of the early parent-child-relationship and the quest to endorse emotional experiences and beliefs in facilitating parental disclosure.

Furthermore, self-report measures allow us to gain insights into the factors parents perceive to influence their relationship with their child. Given the fact that attachment or bonding in the antenatal period is largely one-sided, consisting mainly of the subjective experiences reported by the parent with little observable behaviour (relative to the postnatal period) shown by the fetus, antenatal measures are usually self-reported. Although self-report measures are subject to social desirability bias which can skew interpretation (Arnold & Feldman, 1981; van de Mortel, 2008), they are less costly and labour-intensive to administer (Streiner, Norman, & Cairney, 2015). In addition, they allow us to gain an understanding into the parent’s subjective experience of their relationship with their child, which can be meaningful clinically and valuable for research (Condon & Corkindale, 1998; Scopesi et al., 2004). In clinical settings, valid and reliable measures, which are quick and easy to administer, can facilitate screening for difficulties in the parent-infant-relationship and they can also be used to assess change (Brockington et al., 2001). Moreover, the relative ease of administration means that these instruments can be more readily incorporated into large-scale studies and surveys, including those with multiple follow-ups, thereby facilitating research in this area (Pallant, Haines, Hildingsson, Cross, & Rubertsson, 2014). In order to have clinical and
research utility, self-report measures must meet criteria for validity and reliability (Crandall, 1976; Streiner et al., 2015) and ideally convergence or concurrent validity with a ‘gold standard’ observational measure. However, when choosing a measure, clinicians or researchers also need to know which measure is suitable for their population and which one accurately assesses change (Streiner et al., 2015), as evidence-based assessment is considered intrinsic to professional practice (e.g., Hunsley & Mash, 2008).

Several parent-report measures of the early parent-infant-relationship have been developed, which differ in terms of their focus, format, content, length, theoretical underpinnings, the purpose for which they were developed, and the extent to which information exists regarding their validity and reliability. Whilst recent reviews have explored the associations between pre-and postnatal bonding (McNamara, Townsend, & Herbert, 2019; Tichelman et al., 2019), only three reviews have explicitly assessed self-report measures of the parent-child-relationship and examined their psychometric properties (Perrelli, Zambaldi, Fantilino, & Sougey, 2014; Pritchett et al., 2011; Van den Bergh & Simons, 2009). These three reviews differed in their focus. Van den Bergh and Simons (2009) critically evaluated information of the construction and psychometric properties of three maternal-fetal attachment measures only: the *Prenatal Attachment Inventory* (PAI, Müller, 1993), the *Maternal-Fetal Attachment Scale* (MFAS, Cranley, 1981) and the *Maternal Antenatal Attachment Scale* (MAAS, Condon, 1993). Although the PAI and the MFAS appeared to have some robust psychometric properties, all three measures had weaknesses and required further psychometric validation. Pritchett et al. (2011) described the validity and reliability of measures of family functioning. However, the inclusion of 107 measures meant that no measures were reviewed in specific detail. Finally, Perrelli et al. (2014) undertook an integrative review of measures that could be used in pregnancy and in the first year postpartum. Their review identified 23 articles published after 2002 relating to 13 measures, of which 10 were measures completed by parents.
Whilst this review identified many of the important and widely used parent-report measures of the early parent-infant-relationship, only a relatively small number of research studies and measures were identified.

A further limitation of these reviews is that they did not use a formal quality appraisal tool which would have allowed for a detailed assessment of the papers’ methodological quality and an easier comparison between measures (Terwee et al., 2012). However, a standardized, evidence-based approach to reporting the psychometric properties is essential in order to ensure that the quality of measures used in clinical practice and service improvements is appropriately high (e.g., Kilbourne et al., 2018). Consequently, a review that can be considered a relevant and comprehensive guide for researchers and clinicians is required, especially given the focus on the expansion of perinatal mental health services (NHS England, 2018).

Taking into consideration the aforementioned gaps in the perinatal field, the main aim of the current systematic review is to provide an overview and evaluation of existing parent-measures of the parent-infant-relationship to assist researchers and clinicians in identifying the most suitable measure to use in their research, practice or service. Specifically, the current review addresses the limitations of previous reviews by bridging the gap between the depth and breadth of the included measures within the systematic review. We aimed to achieve this by a) appraising only measures that assess the parent-infant-relationship in terms of perceived bond or parent-reported attachment rather than broader or related concepts (e.g., maternal self-efficacy, maternal attitudes) in studies that specifically aimed to develop a measure or test its psychometric properties; b) utilising a systematic search strategy to increase confidence that the review included a comprehensive list of measures; c) applying the COnsensus-based Standards for the selection of health Measurement INstruMents (COSMIN; Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018), a comprehensive and systematic tool for appraising and comparing the quality of individual measures in terms of their psychometric...
properties and clinical utility, and d) identifying relevant administrative properties of the
identified measures (Bot et al., 2014).

2. Methods

This systematic review was conducted in line with the Preferred Reporting Items for
Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff,
Altman, & the PRISMA group, 2009) and registered with the PROSPERO database
(www.crd.york.ac.uk/prospero; registration number CRD42017078512).

2.1. Search strategy and paper inclusion

In keeping with the aims of the review, we conducted a systematic literature search in eight
electronic databases using four steps. In step 1 we designed a search strategy to retrieve peer-
reviewed papers relevant to the development, validation and implementation of self-report
measures of the early parent-infant-relationship, which was piloted by one reviewer (CG). The
aim of this pilot search was to increase specificity and sensitivity to capture the highest
possible proportion of relevant articles. This pilot search resulted in low specificity with too
many irrelevant articles being initially retrieved; thus, following further consultation with a
university librarian, we refined our search strategy by adding the ‘adjacency’ operator
(abbreviated as ‘ADJn’ whereby \( n \) refers to a number of words from each other in any order) to
our search terms. This strategy led to step 2 whereby another reviewer (SV) conducted the
search in three electronic platforms (Ovid, Clarivate and EBSCO) and their eight bibliographic
databases from their inception to the end of April 2019: MEDLINE, Embase, PsycINFO,
PsycARTICLES, Maternity and Infant Care, Health and Psychosocial Instruments, Web of
Science and CINAHL. The search was updated in August 2019.
Limitation for language or year of publication were not set because the exclusion of non-English studies could introduce risks of bias and, therefore, each non-English study should be evaluated case-by-case to maintain internal validity (Higgins & Green, 2011; Neumann Rasmussen & Montgomery, 2018).

In step 2, we searched the following terms in the title, abstract or keywords in those eight databases (see Appendix A for a sample search strategy): 1) (parent*or maternal or paternal or mother* or father*) adj7 (child or infant or newborn or foet* or fetus or fetal or baby or neonate); 2) (antenat* or prenat* or puerper* or postnat* or postpart* or peripartum or pregnan* or perinat*); 3) (measur* or scale$ or questionnaire$ or construct$ or tool$ or inventor* or instrument$ or test*) adj7 (attachment or relation* or bond* or orientation or synchrony or synchronicity or “emotional availability” or attitude* or belief* or responsiv* or feel* or interact*). Papers retrieved from this search were then screened for measures relevant to the aims of the review.

In step 3, further searches with the names of identified measures were conducted in a ninth database (i.e., PubMed) to identify the original development/validation paper(s) for that measure as well as papers reporting further validation work undertaken with any identified and included measures. In the final and fourth step, the reference lists of included articles were checked for additional relevant studies. When the initial development/validation work for a measure was unpublished, further information was sought from study authors. When this was not possible, we extracted relevant development and validation process information about this measure from papers by the original authors.

To verify inter-rater reliability of the screening, an independent research assistant (CS) independently double-screened 1% of all identified articles during the screening stage and 20% of potentially eligible articles to determine their inclusion or exclusion. The percentage of
inter-rater agreement and Cohen’s kappa were calculated on both types of screening to ensure the validity of the screening process.

2.2. Inclusion criteria of papers and article selection

Papers were included if they described the initial development and validation of a relevant measure. Papers were also included if they described an attempt to validate and/or to test the psychometric properties of an included measure, and this was the clearly stated aim of the paper. Decisions about the inclusion/exclusion of measures and papers were based on the initial judgment of two reviewers (AW and CG). Their decisions were verified by two other reviewers (SV and AM), and any disagreements were resolved through consultation with the fifth reviewer (MH).

2.3. Inclusion and exclusion criteria of measures

Measures were included if they were completed by the parent and assessed the parent’s perception of the parent-infant-relationship or bond during the antenatal period or the postnatal period up until an infant age of two years. Measures were excluded if they were not assessing the parent-infant-relationship *per se* but instead assessed a related concept (e.g., ‘parenting style’ or ‘attitudes to pregnancy’) or if they only assessed the parent-infant-relationship as part of a subscale in a longer inventory (e.g., the MAMA, Kumar et al., 1984). As the content of measures assessing related constructs (e.g., maternal self-efficacy, maternal attitudes, etc.) can be very similar to those of measures explicitly described by authors as measures of bonding or attachment, we based inclusion decisions on item content rather than author description (e.g., the *How I Feel About My Baby Now Scale*, FAB, Leifer, 1977; the *Mothers’ Object Relations Scales Short Form*, MORS-SF, Oates, Gervai, Danis, Lakatos, & Davies, 2018).
On several occasions, original measure authors or other researchers proposed shortened or alternative versions of measures which had already been identified and included in the current review. For example, we included the *Short Postpartum Bonding Questionnaire* (*SPBQ*; Kinsey, Baptiste-Roberts, Zhu, & Kjerulff, 2014), which was based on the original 25-item *Postpartum Bonding Questionnaire* (*PBQ*; Brockington et al., 2001) but shortened to address the need for a briefer instrument to measure parent-infant-bonding as part of a large scale telephone interview survey. Similarly, in several cases, researchers (but not the original authors) conducted psychometric testing on slightly different versions of measures (e.g., containing fewer items or having fewer Likert response categories). These alternative versions were also included in the current review and treated as separate independent measures.

2.4. Assessing the psychometric properties of included measures

We evaluated the measurement properties of the included studies and measures using: 1) the COSMIN criteria for evaluating the quality of the measure development studies and content validity studies (Terwee et al., 2018), 2) the COSMIN Risk of Bias checklist (Mokkink et al., 2018) to assess the methodological quality of the studies, 3) the COSMIN checklist to examine eight psychometric results, including structural validity, internal consistency, reliability, hypothesis testing for construct validity, cross-cultural validity/measurement invariance, measurement error, criterion validity and responsiveness (Mokkink et al., 2018; Prinsen et al., 2018), and 4) the modified Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to examine the quality of the evidence (Mokkink et al., 2018). All materials are available at www.cosmin.nl/index.html.
2.4.1. Step 1: Quality assessment of included studies

The first step in the process to assess the methodological quality of included studies is achieved via the application of the ‘COSMIN Risk of Bias checklist’ (Mokkink et al., 2018). This checklist consists of categories for appraising the quality of the outcome measure development studies as well as the quality of various psychometric measurement properties which are outlined above (see Table 1 for definitions of measurement properties, Mokkink et al., 2018).

Content validity was assessed in terms of relevance, comprehensiveness and comprehensibility of the measure’s items (Terwee et al., 2018).

During the pilot stage we identified that many researchers did not explicitly describe what they explored or evaluated in terms of content validity in their studies. Consequently, we expanded the COSMIN’s definitions of ‘relevance’ and ‘comprehensibility’: studies were considered to evidence ‘relevance’ when they evaluated the relevance, appropriateness, suitability and/or acceptability of each item in the target population. In terms of ‘comprehensibility’, studies were rated if they evaluated the understanding, coherence, clarity, meaning and/or ambiguity of the items and whether the response options, instructions and/or the recall period were clear and comprehensible. ‘Comprehensiveness’ was evaluated in accordance with the COSMIN guidelines whereby participants should have been explicitly asked about whether the items comprehensively covered the construct that the outcome measure (or the sub-scale) intended to measure or if the included domains together comprehensively covered the wider construct measured by the total score of the outcome measure (Terwee et al., 2018).

Each measurement property (including content validity) was rated across several items assessing different aspects of quality, using a four-point COSMIN Risk of Bias scale (i.e., 4 = ‘very good’, 3 = ‘adequate’, 2 = ‘doubtful’, 1 = ‘inadequate’). An overall score for the methodological quality of a study was determined for each measurement property separately by
taking the lowest rating of any of the items in a given category. When the developers of the
original version of a measure omitted to provide detailed information on one or more
psychometric properties, but there was sufficient information to assume that the study was
conducted adequately, we deviated from the stricter COSMIN guidance and opted to give an
‘adequate’ or ‘doubtful’ rating rather than an ‘inadequate’ rating. For example, if in the
measure development study it was unknown whether the qualitative data, collected for the
purposes of cognitive interview or pilot testing, were coded by one or two researchers
independently, we rated it as ‘doubtful’ rather than ‘inadequate’ due to lack of information.

Interpretability or the degree to which one can assign qualitative meaning to
quantitative scores (Mokkink et al., 2009, 2018) is not considered a measurement property but
an important characteristic of a measurement instrument (Mokkink et al., 2018). This means
that investigators should provide information about clinically meaningful differences in scores
between subgroups, floor and ceiling effects, and the minimal (clinically) important change
(Mokkink et al., 2009). However, since a limited number of studies reported aspects of
interpretability, we could not present this in our review.

2.4.2. Step 2: Assessment of study outcomes

The second step involved assessing the study results for each of the included measures,
according to the updated 2018 measurement for good measurement properties (Mokkink et al.,
2018; Prinsen et al., 2018). These criteria cover eight measurement properties, for each of
which the rater is required to assign ‘+’, ‘?’ or ‘−’. A ‘+’ is assigned when the study findings
provide good evidence of a measure exhibiting this property (i.e., ‘sufficient’ rating); a ‘?’ is
assigned when results are equivocal or appropriate tests have not been performed (i.e.,
‘indeterminate’ rating) and a ‘−’ is assigned when appropriate tests have been performed and
the result suggests that the measure does not exhibit this property as defined by the checklist criterion (i.e., ‘insufficient’ rating). This checklist and quality criteria are presented in Table 1.

The content validity of an outcome measure was evaluated according to the quality and results of the available studies and the outcome measure itself (Terwee et al., 2018). Although the COSMIN guidelines suggest not to rate a study if the quality of the study (according to the risk of bias assessment) was ‘inadequate’, we decided to rate all studies, including those with an ‘inadequate’ quality rating, in order to gain a comprehensive overview of a particular outcome measure. Content validity of each outcome measure was rated according to the development studies (scored as ‘+’, ‘?’ or ‘–’ for ‘sufficient’, ‘indeterminate’ and ‘insufficient’ ratings, respectively), available content validity studies (also scored as ‘+’, ‘?’ or ‘–’) and ratings given by two reviewers (AW and SV) (scored as ‘+’, ‘±’ or ‘–’ for ‘sufficient’, ‘inconsistent’ and ‘insufficient’ ratings, respectively). When no content validity studies were available or only content validity studies of inadequate quality were available, the overall ratings for content validity were determined according to the reviewers’ ratings as per COSMIN criteria.

In order to rate the structural validity of measures, we had to adapt the criteria as the current 2018 COSMIN criteria for good measurement properties do not include guidance for rating the results of Exploratory Factor Analysis (EFA). Consequently, EFAs were rated as ‘sufficient’ if $\geq 50\%$ of the variance was explained (as in previous versions of the COSMIN criteria; see Terwee et al., 2012). Such evidence was downgraded for methodological quality based on the risk of bias checklist (i.e., studies using EFA can only be rated as ‘adequate’ rather than ‘very good’). When the % of variance accounted for (in the case of EFA) or model
fit statistics (in the case of CFA) were not reported, an ‘indeterminate’ rating was assigned.

Finally, when higher quality evidence (e.g., CFA) was available for a given measure, lower quality evidence (e.g., EFA) was ignored.

In terms of hypothesis testing for construct validity, the decision was made to include any published measure as a comparator instruments that measured a similar construct (e.g., other attachment measures included in the current review or a subscale from a measure not included in the review, such as the attitudes toward pregnancy and the baby subscale of the MAMA scale, Kumar et al., 1984). To receive a ‘sufficient’ rating, 75% of the correlations tested had to meet the cut-off of \( r \geq .50 \) against a comparator instrument measuring a similar construct. Given the lack of an established self-report measure for the construct under study, caution is needed in interpreting the results for this measurement property.

We also adapted the criteria for rating reliability due to ambiguity in the COSMIN guidelines in a way that the studies that reported a correlation coefficient (i.e., Pearson’s or Spearman’s) but did not report the intraclass correlation coefficient (ICC) for (test-retest) reliability would still receive a ‘sufficient’ or ‘insufficient’ rating which would normally receive an ‘indeterminate’ rating if the ICC was not applied. Instead, we decided to reflect this in the quality of evidence (described in section 2.4.3) whereby studies that did not use the more robust method (i.e., the ICC) would get a lower rating even if the study received a ‘sufficient’ rating on reliability.

2.4.3. Step 3: Summary and quality grading of the evidence

As per COSMIN criteria, the psychometric findings reported in each of the included studies were summarized and graded for each measure. This process resulted in each measure being assigned two ratings: 1) an overall rating of ‘sufficient’ (‘+’), ‘insufficient’ (‘−’) or ‘indeterminate’ (‘?’) for the eight psychometric properties (except content validity), or an
overall rating of ‘sufficient’ (‘+’), ‘insufficient’ (‘–’) or ‘inconsistent’ (‘±’) for content validity, and 2) an overall rating of methodological quality for each measurement property (‘high’, ‘moderate’, ‘low’ or ‘very low’). The latter rating is achieved through following the modified GRADE approach, which involves consideration of several factors in rating methodological quality of the pooled results (e.g., the evidence for risk of bias, inconsistency of results and imprecision through small sample sizes). Importantly, this approach also takes into account the number of available studies and the methodological quality of each individual study. More detailed information on how GRADE was conducted can be found in the COSMIN manual (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). Definitions for each of the GRADE quality level ratings are shown in Table 2.

\[\text{Insert Table 2 about here}\]

2.4.4. Step 4: Assessment of practical administrative properties

Practical, administrative or clinimetric properties that would affect the ease with which each of the measures could be employed in a clinical or research context, but are not covered in the COSMIN or Terwee et al. (2018) checklists, were also assessed. The following properties were assessed:

1. **Time to administer:** As measure authors did not routinely report this property, this was assessed independently by two reviewers (SV and AM) who completed each measure and timed themselves. As per Bot et al.’s (2004) clinimetric checklist, a positive rating was given when the questionnaires could be completed within 10 minutes.

2. **Ease of scoring** refers to the extent to which the measure can be scored by a trained investigator or expert. In accordance with Bot et al.’s (2004) checklist, the scoring method was rated as easy when the items were simply summed, moderate when a visual analogue
scale or simple formula was used and difficult when either a visual analogue scale in combination with a formula or a complex formula was used.

3. **Readability and comprehension**: The Flesch Reading Ease (FRE; Flesch, 1948) method was used to assess readability and comprehension. The text is rated on a 100-point-scale in which 100 represents the easiest text and 0 the hardest. Measures scoring $\geq 90$ using the FRE were considered excellent for this property; measures scoring between 80 and 89 were considered good; measures scoring between 70 and 79 were considered fair and measures with scores below $< 69$ were considered poor.

4. **Availability and conditions of use** refers to the ease with which researchers/clinicians can obtain the questionnaire and whether it is free to use. If the measure was easily accessible on the internet or through e-mailing the first author, and it was also free to use, availability was classed as excellent. If a measure was difficult to obtain but was free of cost, the measure was classified as good. If a measure was easy to obtain but had a cost, the measure was classed as fair. Finally, if a measure was difficult to obtain and there was a cost for accessing or utilising the instrument, the measure was classified as ‘poor’.

2.5. **Inter-rater reliability**

Extraction of data and the assessment of the methodological quality (i.e., risk of bias) was performed by reviewers independently (AM, SV and CG). The assessment of all psychometric properties (except content validity) was performed by one reviewer (AM). The reliability of ratings was confirmed by having another reviewer (SV), who independently rated 20% of the papers. For the measure development studies and content validity studies, another reviewer (SV) completed risk of bias and rated content validity; 20% of those papers were independently rated by an independent research assistant (CS). Inter-rater reliability was met if Cohen’s kappa between the two reviewers was above .61, indicating ‘substantial’ agreement (McHugh, 2012),
on all psychometric ratings. When this was not achieved, the disagreements were discussed and
resolved through consultation with another reviewer (AW).

3. Results

3.1. Review process

The original search identified 15,924 papers. After removing duplicates, the titles and abstracts
of 12,081 papers were screened. The titles, abstracts and/or full texts of 220 papers were
examined against inclusion and exclusion criteria. In August 2019 the search was repeated
which resulted in 600 hits between January and August 2019 which were fully screened. Only
two studies were identified: a study describing the development of the MAAS-13 and PAAS-
13 (Göbel et al., 2019) which was included in the review and a non-English study describing a
Slovenian version of the PAI (presented in Appendix B alongside other non-English papers).
The agreement for the screening of 1% of all identified articles was 94.7% (kappa = .90) and
for the 20% of potentially eligible articles the agreement was 75% (kappa = .58). Any
discrepancies in the exclusion and inclusion of studies were resolved amongst all reviewers
through discussion.

After a detailed assessment, 65 papers evaluating 17 original measures in associated
development studies and 13 modified versions, derived from only four of the identified 17
measures, were included in the review (for the references of the included papers, please see
Appendix C). In total, 14 antenatal measures (eight original and six modified versions) and 18
postnatal measures (ten original measures plus eight modified version), of which one measure
(the Prenatal and Postnatal Bonding Scale, PPBS, Cuijlits et al., 2016) could be used
antenatally and postnatally, were reviewed. The majority of these measures were maternal, but
we also identified four paternal measures (three antenatal and one postnatal version). The
search process and outcome are illustrated in Figure 1.
3.2. Study characteristics and information on measure development

The publication dates of the studies describing the original 17 measures ranged from 1977 to 2018. The validation work undertaken for the original 17 measures included studies conducted in eight different countries, such as the USA (n=5), Australia (n=4), the UK (n=4), the Netherlands (n=1), Hungary (n=1), Korea (n=1), Sweden (n=1) and India (n=1); however, the sample in the MORS-SF (Oates et al., 2018) scale development comprised British and Hungarian mothers (see Table 3 for details). Study sample sizes reported by measure authors in their development studies ranged from 19 to 1050 women and 100 to 461 men. The majority of studies included non-clinical samples (n=15), with only two studies using a clinical sample of women with mental illness (Brockington et al., 2001; Hackney, Braithwaite & Radcliff, 1996).

3.3. Description of identified measures

A description of each of the 17 original measures is presented in Table 3. The majority of measures focused on the assessment of parent reported perceptions of bonding and attachment. However, one measure also focussed on the psychodynamic concept of object relations (i.e., the MORS-SF, Oates et al., 2018) and another measure also focussed on the assessment of maternal role attainment (i.e., the Mother-to-Infant Relations and Feelings Scale, MIRFS, Thorstensson, Hertfelt Wahn, Ekström, & Langius-Eklöf, 2012a). Although the authors did not set out to test the MIRFS, we included it in this review because Ekström and Nissen (2006) described the MIRFS’ development and evaluated its content validity.
Eight scales were measures of the parent-fetus-relationship, administered to expectant women and men in the antenatal period of pregnancy: the *How I Feel About My Baby Now Scale (FAB, Leifer, 1977)*, the *Maternal-Fetal Attachment Scale (MFAS, Cranley, 1981)*, the *Prenatal Attachment Inventory (PAI, Müller, 1993)*, the *Maternal Antenatal Attachment Scale (MAAS, Condon, 1993)*, the *Pre- and Postnatal Bonding Scale (PPBS, Cuijlits et al., 2016)*, the *Paternal-Fetal Attachment Scale (PFAS; Weaver & Cranley, 1983)*, the *Paternal Antenatal Attachment Scale (PAAS; Condon, 1993)* and the *Korean Paternal-Fetal Attachment Scale (K-PAFAS; Noh & Yeom, 2017)*.

As per COSMIN criteria, any modified versions of measures were reviewed separately even if they differed from the original scale by only one item. Three modified versions were identified for the original 24-item-MFAS offering different item totals: the *MFAS-23 (Müller & Ferketich, 1993; Müller, 1993)*, the *MFAS-20 (Busonera, Cataudella, Lampis, Tommasi, & Zavattini, 2016a)*, and the *MFAS-17 (Seimyr, Sjögren, Welles-Nystrom, & Nissen, 2009; Sjögren, Edman, Widstrom, Mathieson, & Uvnas-Moberg, 2004)*. The original 19-item-MAAS had also been shortened in modified versions referred to as the *MAAS-13 (Göbel et al., 2019)* and *MAAS-12 (Navarro-Aresti, Iraurgi, Iriarte, & Martinez-Pampliega, 2016)*.

Five original and five modified versions of these were measures of the mother-fetus-relationship: the *FAB*, the *MFAS-24*, the *MFAS-23*, the *MFAS-20*, the *MFAS-17*, the *PAI-21*, the *MAAS-19, the MAAS-12*, the *MAAS-12* and the *PPBS*. Only three measures assessed the father-fetus-relationship, namely the *PFAS*, the *PAAS* and the *K-PAFAS*. The *PAAS* has also been revised and shortened to the 13-item-*PAAS* (Göbel et al., 2019).

Ten original scales and eight modified versions were measures of the parent-infant-relationship. As can be seen in Table 3, the original scales were the *Maternal Attachment Inventory (MAI-26; Müller, 1994)*, the *Mother Infant Attachment Scale (MIAS; Bhakoo, Pershad, Mahajan, & Gambhir, 1994)*, the *Mother and Baby Interaction Scale (MABISC*;
Hackney, Braithwaite & Radcliffe, 1996), the Maternal Postnatal Attachment Scale (MPAS; Condon & Corkindale, 1998), the Postpartum Bonding Questionnaire (PBQ-25; Brockington et al., 2001), the Mother-to-Infant Bonding Scale (MIBS-8; Taylor et al., 2005), the Mother-to-Infant Relations and Feelings Scale (MIRFS; Thorstensson et al., 2012a), the PPBS (Cuijlits et al., 2016), the Mothers’ Object Relations Scales Short Form (MORS-SF; Oates et al., 2018) and the Paternal Postnatal Attachment Scale (PPAS; Condon, Corkindale, & Boyce, 2008). Only the PPAS was designed to assess the father-infant-relationship. Of these measures, only the PPBS could be used both antenatally and postnatally.

The MFAS (Cranley, 1981) and the PFAS (Weaver & Cranley, 1983) as well as the MAAS (Condon, 1993) and the PAAS (Condon, 1993) are maternal and paternal versions of the same measures, respectively, and can be completed by mothers and fathers in the same family. The MAI (Müller, 1994) is the postnatal version of the PAI (Müller, 1993). Condon and colleagues (Condon, 1993; Condon & Corkindale, 1997, 1998; Condon, Corkindale & Boyce, 2008; Condon, Corkindale, Boyce, & Gamble, 2013) have produced measures based on Condon’s (1993) model of human attachment with antenatal and postnatal measures for both mothers and fathers (i.e., MAAS, PAAS, MPAS and PPAS).

Of the postnatal measures, the PBQ has received the most attention by other researchers who have produced shorter versions, including the PBQ-22 (Wittkowski, Williams, & Wieck, 2010), the PBQ-19 (Vengadavaradan, Bharadwaj, Sathynarayanan, Durairaj, & Rajaa, 2019), the PBQ-16 (Reck et al., 2006), the PBQ-16-J (Kaneko & Honjo, 2014), the PBQ-14 (Suetsugu Honjo, Ikeda, & Kamibeppu, 2015) and the Short PBQ with 10 items (Kinsey et al., 2014).

Although the PBQ has also been used with fathers at 2 months postpartum in a Swedish study (Edhborg, Matthiesen, Lundh, & Widström, 2005), the PBQ was not specifically developed to be used with fathers. As Edhborg et al. (2005) did not evaluate the psychometric properties of the PBQ in their male sample, this study was not rated in our review.
The only other postnatal measure with modified versions is the 8-item-MIBS, which had been reduced to seven items in the MIBS-J-7 (Ohara et al., 2016) and extended to 10 items in the MIBS-J-10 (Yoshida, Conroy, Marks, & Kumar, 2012).

3.3.1. Additional information on the measures’ items and target population

The majority of the measures comprise items that are worded as statements on a Likert-scale that typically enquire how the mother is feeling towards the developing fetus or the newborn. For example, the PAI includes items, such as “I stroke the baby through my tummy” or “I enjoy feeling the baby move”. The PBQ includes items, such as “I feel close to my baby” or “My baby winds me up”, and the MPAS includes items, such as “When I am not with the baby, I find myself thinking about the baby: …” or “Taking care of this baby is a heavy burden of responsibility. I believe this is: …”. Only one measure, the MIRFS, was a two-part measure in which seven items (worded as statements) evaluated the mothers’ perception about the relationship between the mother and her baby (e.g., “I talk a lot with my baby” and “I do not talk at all with my baby”) and seven items (worded as adjectives) explored the mothers’ current feelings towards the baby (e.g., “Difficult” and “Easy”). Although most studies described the population with whom the study was conducted, the majority of the studies did not specify the target population of parents by providing information about the gestation age or the infant’s age. Consequently, it was impossible to determine the measures’ applicability to parents of infants at different developmental ages and we had to assume that they targeted parents of children younger than two years old.

The number of items used in the original 17 measures ranged from five (e.g., the PPBS) to 26 items (e.g., the MAI). Of the original 17 measures, seven measures were unidimensional (FAB, PAI, MAI, PPBS, MABISC, MIBS-8 and MIAS), whereas ten measures included multiple sub-scales (MFAS, PFAS, MAAS, PAAS, K-PAFAS, MPAS, PPAS, PBQ, MORS-SF and
MIRFS), which ranged from two (e.g., MAAS, PAAS and MORS-SF) to five subscales (e.g., MFAS).

Most measures were designed for the assessment of parents within a non-clinical population who were asked to reflect on their feelings or thoughts in the present moment (e.g., the FAB), the past two weeks (e.g., the PAAS) or during the past month (e.g., the PPBS, the PAI-21, the MABISC). However, six measures did not state a specific recall time and three measures accepted a variable timeframe.

As our search did not exclude studies not written in English initially, our review identified that most of the 17 original and 13 modified measures are available in a total of 17 languages including English. Other language versions included measures in Chinese, Korean, Japanese, German, Italian, French, Portuguese, Spanish, Dutch, Swedish, Norwegian, Hungarian, Turkish, Persian, Tamil and Hindi. Four measures were only available in one language: the FAB in English only, the MIAS in Hindi only, the K-PAFAS in Korean only and there appears to be only a Swedish version of the MIRFS. In addition, the MORS-SF was validated on a mixed sample of British and Hungarian women.

Although it was our original intention to include measures not written in English, we were unable to a) apply the COSMIN criteria consistently across these studies ourselves and b) identify professional translators trained in the application of the COSMIN criteria for the foreign language papers we identified. For comprehensiveness, the foreign language papers are presented in Appendix B.

3.4. Measurement properties assessed

Sixty-five studies pertaining to the 17 measures and their 13 modified versions rated aspects of validity and reliability (see Table 4 and Table 5). Several of the included studies (e.g., Bienfait et al., 2011; Brockington et al., 2001, 2006) tested for diagnostic accuracy (i.e., sensitivity and
specificity of the measure in detecting bonding difficulties), but this property does not fall within the COSMIN taxonomy and consequently it was not rated.

3.4.1. Assessment of validity

3.4.1.1. Content validity

According to the COSMIN criteria for assessing content validity (Terwee et al., 2018), the relevance, comprehensiveness and comprehensibility of the 17 measures and their 13 modified versions were rated separately in a multi-step process. Firstly, the overall quality of the development (i.e., risk of bias checklist) of the 17 original outcome measures were evaluated: three original measure development studies (MABISC, MPAS and PPAS) had a ‘doubtful’ rating and 14 original measure development studies were rated as ‘inadequate’ (see detailed ratings in Table 4). In this step, the quality of the content validity studies was also evaluated according to whether participants in a content validity study had been asked about the relevance, comprehensiveness and comprehensibility of the measure items and whether professionals had been asked about the relevance and comprehensiveness of the measure items.

A total of 16 out of 65 studies (Table 4) evaluated content validity either among a participant group (i.e., mothers or fathers) or among a professional group (i.e., midwives, psychologists, psychiatrists, researchers, etc.). ‘Relevance’ and ‘comprehensibility’ was evaluated among participants in six and nine studies, respectively, and among professionals in 10 and five studies, respectively. However, none of the content validity studies evaluated ‘comprehensiveness’ among participants. From all content validity studies evaluating relevance, comprehensiveness and comprehensibility among participants and/or professionals, only two studies received an ‘adequate’ rating: Busonera et al. (2016a) for ‘relevance’ among
participants using the MFAS and Riera-Martin et al. (2018) for ‘relevance’ among professionals on the MPAS and the PPAS. The remaining studies received a ‘doubtful’ rating for ‘relevance’, ‘comprehensibility’ and/or ‘comprehensiveness’ when these properties were assessed in a particular study (see Table 4 for detailed ratings).

Once the quality of the development studies and content validity studies were rated for each measure, the content validity of an outcome measure was evaluated. With regards to the 17 original measures, the overall content validity was rated as ‘sufficient’ (‘+’) for the following measures: the FAB, the PAI and the MABISC. However, the remaining 14 measures were rated to have ‘inconsistent’ (‘±’) evidence for their overall psychometric properties (see Appendix D for detailed ratings). The reasons for ‘inconsistent’ ratings were explored and in the majority of the cases, the information on relevance, comprehensiveness and comprehensibility presented by the measure authors in the papers was poor. However, the independent evaluation by two reviewers of the scale alone indicated that the information given was ‘sufficient’ and met the COSMIN criterion of ‘+’.

In the third step, the quality of the evidence was rated using the GRADE approach. As the development study of the measure received ‘inadequate’ quality ratings, according to COSMIN criteria these studies have to receive a lowered rating in terms of the GRADE. Therefore, three scales (MABISC, MPAS and PPAS) were rated as ‘low’ and 14 measures were rated as ‘very low’ for relevance, comprehensiveness and comprehensibility according to the GRADE approach (Table 5).

3.4.1.2. Structural validity (part of construct validity)

In order to rate structural validity, we adapted the 2018 COSMIN criteria for good measurement properties as previously outlined, because many of the included studies
conducted Exploratory Factor Analysis (EFA); however, the COSMIN criteria does not provide guidance for rating the results of the EFA.

Structural validity was tested for the majority of the included measures (27 out of 29; 93%), but not for the FAB or the PBQ-16. Of the prenatal measures, only the MFAS-17 and the PPBS were assigned a ‘sufficient’ rating. The remaining prenatal measures were assigned ‘insufficient’ ratings. Of the postnatal measures, the MABISC, the PBQ-22, and the MIBS-J-7 were assigned ‘sufficient’ ratings, whereas the MPAS, the PBQ-25, the MIBS-J-10, the MORS-SF and the PPAS were assigned ‘insufficient’ ratings. The MAI-26 was assigned an ‘inconsistent’ rating because there was evidence for structural validity but for its two different factor structures. The remaining postnatal measures were assigned ‘indeterminate’ ratings (see Appendix E for detailed ratings). The quality of the evidence was graded from ‘very low’ to ‘high’ for this measurement property.

3.4.1.3. Hypothesis testing (part of construct validity)

As none of the studies included an observer-rated measure of parent-child attachment, correlations had to meet the agreed cut-off against a self-report instrument only.

Twenty-five of 30 measures (83.3%) had studies reporting information for this measurement property. Of the prenatal measures, the MFAS-23 and the PFAS were assigned ‘sufficient’ ratings, but the MFAS-24, MFAS-20, PAI-21, MAAS-19 and PAAS-19 were assigned ‘insufficient’ ratings, and the MFAS-17 and the PPBS were assigned an ‘indeterminate’ rating. Of the postnatal measures, the MABISC and MPAS were assigned ‘sufficient’ ratings, the MAI-26, PBQ-25, PBQ-14, MIBS-8 and PPAS were assigned ‘insufficient’ ratings, and the remaining measures were assigned ‘indeterminate’ ratings (i.e., when the hypothesis was not as defined by our review team). The quality of the evidence for all pre- and postnatal measures was graded ‘high’ for this measurement property.
3.4.1.4. Cross-cultural validity (part of construct validity)

Although many studies \((n = 38)\) aimed to adapt a given measure to different ethnic and language groups and some included back translation and other necessary procedures, none evaluated cross-cultural validity by comparing multiple groups by factor analysis and testing for differential item functioning (e.g., English- and Dutch-speaking), as stipulated in the COSMIN criteria. For this reason, this property is omitted from Table 5. See Appendix B for an overview of which measures have a version available in a different language.

3.4.1.5. Measurement invariance (part of construct validity)

Only two measures (6.7\%) could be rated for this measurement property (see Table 5). The MPAS and the PPAS were assigned ‘sufficient’ ratings of measurement invariance, demonstrating that these two measures appear to be measuring the same underlying construct (when tested with mothers and fathers, respectively). The quality of evidence for this property was ‘moderate’.

3.4.1.6. Criterion validity

None of the studies reported on assessment of criterion validity. Hence, this property was omitted from Table 5.

3.4.2. Assessment of reliability

3.4.2.1. Internal consistency

Twenty-eight of the 30 measures (93.3\%) had studies reporting on internal consistency. However, internal consistency could only be rated as ‘indeterminate’ for the majority of these measures because they did not demonstrate at least low evidence of ‘sufficient’ structural
validity (as per the COSMIN criteria). Of the prenatal measures, the PPBS was assigned a ‘sufficient’ rating, whereas the MFAS-17 was assigned an ‘insufficient’ rating; the remaining prenatal measures were assigned ‘indeterminate’ ratings. Of the postnatal measures, the PBQ-22 was the only measure to receive a ‘sufficient’ rating. The MIBS-J-7 was assigned an ‘insufficient’ rating, and the remaining measures were assigned ‘indeterminate’ ratings. The quality of the evidence was graded from ‘very low’ to ‘high’ for this measurement property (see Table 5).

3.4.2.2. Reliability

Eleven of 30 measures (36.7%) had studies that reported test re-test reliability as defined by the COSMIN criteria. Of the maternal measures, the PAI-21, the MPAS, the PBQ-25, the PBQ-14, and the MORS-SF were all assigned ‘sufficient’ ratings, whereas the MAI-31, the MABISC, and the MIBS-J were assigned ‘insufficient’ ratings. Of the paternal measures, the PFAS was assigned a ‘sufficient’ rating whereas the PPAS was assigned an ‘insufficient’ rating. Evidence was graded from ‘very low’ to ‘moderate’ for this measurement property.

3.4.2.3. Measurement error and responsiveness

None of the included studies measured measurement error and responsiveness as defined by the COSMIN criteria; hence, these could not be rated and were omitted from Table 5.

3.5. Inter-rater reliability

The agreement between the two reviewers was 89.0% (kappa = .85) for the risk of bias ratings, 79.7% (kappa = .68) for the measurement properties and 84.5% (kappa = .73) for quality of evidence (i.e., GRADE).
3.6. Clinimetrics/clinical utility

The clinimetrics or clinical utility of the 17 measures, presented in Table 3, were assessed in terms of time of administration, ease of scoring, Flesch Reading Ease (FRE), availability and conditions of use.

3.6.1. Time to administer

Based on Bot et al.’s (2004) suggestion of a desirable completion time of less than 10 minutes, all 17 measures were assessed independently by two reviewers (SV and AM) and could be completed within 10 minutes. Most questionnaires (76.5%) took less than 2 minutes to complete. Four measures (the MAAS, PAAS, MPAS and PPAS) took longer than 2 minutes to administer because items covered multiple pages.

3.6.2. Ease of scoring

In terms of ease of scoring, most measures (n = 14, 82.4%) received an easy rating due to their use of the Likert-scale scoring system. Response options ranged from four to seven options (see Table 3). Only two prenatal measures (the MPAS and the PPAS) received a moderate rating due to a combination of Likert-scale and simple formula scoring. The MIRFS is a two-part-scale, administered postnatally, in which the first sub-scale is rated as a Likert-scale and the second as a semantic differential scale; thus, it was judged to be difficult to score. As indicated in Table 3, in the majority of measures (n = 11, 64.7%) higher scores indicated stronger bonding or attachment, but in four (25%) postnatal measures higher scores were indicative of greater difficulties in the parent-reported bond with their infant. One scale (the MORS-SF) consisted of two sub-scales, whereby higher scores indicated higher maternal perceived levels of warmth as well as invasiveness. The ease of scores for one measure (the
MIRFS) could not be reported because the scale development authors did not specify this in their paper (Thorstensson et al., 2012a).

3.6.3. Readability and comprehensiveness

Seventeen measures were assessed using the Flesch Reading Ease (FRE) test. As can be seen in Table 3, two measures (the K-PAFAS and the PPBS) received a poor rating. An excellent rating in terms of readability was given to one scale only (the MIAS). A fair rating was given to seven measures: FAB, PAI-21, MAAS-19, PAAS, MAI, MPAS, and PPAS, and seven measures, namely the MFAS, PFAS, MABISC, PBQ-25, MIBS-8, MORS-SF, and MIRFS, received a good rating. Of those measures, the MFAS and PFAS were the only antenatal measures.

3.6.4. Availability and conditions of use

The majority of scales (n = 14, 82.4%) were easily accessible on the internet and free of charge; thus, receiving an excellent rating. Three scales, namely the FAB, the PFAS and MIAS, were difficult to obtain but free of charge; hence, they received a good rating.

4. Discussion

In this review we systematically examined the literature to identify, describe and evaluate the psychometric and clinimetric properties of self-report questionnaires for measuring the mother’s or father’s perception of their bond or relationship with their child. Seventeen original measures and their 13 modified versions, described in 65 articles from seven countries, were included and their methodological quality was carefully evaluated. Of these, a few measures were antenatal and postnatal measures for both mothers (i.e., MAAS, MFAS, MPAS) and fathers (PAAS, PFAS, PPAS). The findings indicate that the evidence base for the robustness of self-
report questionnaires measuring the parent-infant-relationship or bond is rather limited; consequently, we can only advise that these measures are used with some caution.

4.1. Observations on the COSMIN guidelines

The current 2018 COSMIN criteria appear to be the most stringent and complex to apply due to the multi-step process whereby firstly the quality of the measure development studies and the content validity studies were evaluated, secondly the methodological quality (risk of bias) of all studies was rated, thirdly the psychometric measurement properties were assessed and finally the quality of the evidence was graded. In other reviews of measures, reviewers either did not choose to apply the COSMIN criteria and opted to use other guidelines for rating each psychometric property (e.g., Lotzin et al., 2015; Perrelli et al., 2014), or they used an older COSMIN version (Terwee et al., 2007, in Wittkowski, Garrett, Calam & Weisberg, 2017; Mokkink et al., 2010, in Bentley, Hartley & Bucci, 2019, or De Vet, Terwee, Mokkink, & Knol, 2011, in Jewell et al., 2019).

Despite using the older COSMIN criteria, reviewers such as Jewell et al. (2019) have highlighted the arbitrary nature of cut-off scores which determine if a measurement property is ‘adequate’ or ‘inadequate’ because in some cases the statistical values indicative of a negative rating were very close to values suggesting a positive one. Furthermore, Jewell et al. (2019) critiqued the use of the ‘worst case counts’ rule because a single flaw in the study would result in only a ‘fair’ or even a negative rating which means that the adequacy and sufficiency of measurement properties and the methodological quality of any evidence are not necessarily a true reflection and most likely an underestimation. This criticism also fits with our observations when applying the COSMIN 2018 guidance.

In the application of the latest COSMIN guidance, we also became aware of how much practice and reporting standards have changed over the course of the last few decades; for
example, the oldest measure our review identified was published in 1977 (e.g., the FAB). It was
frustrating to note that authors reported some relevant information but did so without
methodological consistency or rigor. For example, authors did not always report model fit
statistics for confirmatory factor analyses so that they can be rated appropriately. Moreover,
often authors only reported on correlation coefficients instead of reporting intraclass
correlation coefficient (ICC) or kappa scores for (test-retest) reliability. This information is
vital because the accurate assessment of a scale’s structural validity depends on it.

When following the COSMIN 2018 criteria, we evaluated any modified versions of
measures separately even if the total item count differed by only one item. However, in our
findings we observed that psychometric testing was conducted less rigorously on refined or
revised versions of measures compared to the originally developed measures. Thus, the risk of
bias ratings, which show the methodological quality of each measure, might have been
downgraded in line with the strict rules of COSMIN. This downgrading could be considered
unfair given that some development work (e.g., pilot assessment or cognitive interviewing)
might have been undertaken with the original scale.

Additionally, several of the included studies (e.g., Bienfait et al., 2011; Brockington et
al., 2001, 2006) tested for diagnostic accuracy (i.e., sensitivity and specificity of the measure in
detecting bonding difficulties) and discriminant (or divergent) validity but these properties do
not fall within the COSMIN taxonomy and consequently were not rated by us although we
believe these to be important aspects of psychometric testing. Furthermore, only a few studies
assessed aspects of interpretability (e.g., subgroup analyses, minimal important change, floor
and ceiling effects, etc.) and thus, we could not report on these properties in our review.
4.2. Content validity

A measure’s relevance, comprehensiveness and comprehensibility play a major deciding factor in why a measure may be chosen for clinical or research purposes; consequently, content validity may arguably be the most important psychometric property (Mokkink et al., 2018; Terwee et al., 2018). Based on COSMIN criteria, the methodological quality for the content validity of the original development measures identified in this review was ‘doubtful’ for the MABISC, MPAS and PPAS and ‘inadequate’ for the remaining 14 original measures. However, only 15 of the 65 included studies evaluated content validity at all, with the MFAS and PBQ having received the most attention, which was also noted in Tichelman et al.’s review (2019). Despite some studies evaluating ‘relevance’ and/or ‘comprehensibility’ among participants and/or professionals, none of the studies evaluated a measure’s ‘comprehensiveness’, highlighting the need for further research.

Following psychometric evaluation of all 17 measures, only three measures (FAB, PAI-21, and MABISC) were rated as ‘sufficient’ for overall content validity with the remaining measures receiving ‘inconsistent’ ratings. Nonetheless, the quality of the evidence regarding the FAB and the PAI-21 was ‘very low’ and for the MABISC ‘low’ which indicates some uncertainty regarding the trustworthiness of the overall ratings.

Despite the increased trend towards paying more attention to evaluating content validity in terms of relevance, comprehensiveness and comprehensibility, particularly since 2010, there was a high percentage of studies being rated as ‘inadequate’ or ‘inconsistent’ with a ‘very low’ quality evidence. To mitigate against the very strict COSMIN criteria, we applied a more flexible approach by including studies that described other relevant aspects of content validity, such as appropriateness, suitability, acceptability, understandability, coherence, ambiguity and clarity of the items or overall measure. We consider these to be important aspects of content validity and they are potentially worthy of consideration as an expansion of content validity in
the COSMIN guidelines. Although study authors described their evaluation of content validity, it often remained unclear what they had actually explored and how they had conducted the evaluation, because these studies were not developed or conducted in accordance with the high reporting standards of the COSMIN criteria. Hence, applying the COSMIN 2018 criteria resulted in most studies being rated as ‘doubtful’, ‘inadequate’, ‘inconsistent’ and/or ‘indeterminate’.

Furthermore, rating content validity according to the COSMIN criteria is a complex and multi-step process whereby the overall rating depends on the ratings of the development study, content validity study (or studies) and reviewers’ ratings. On occasions when there were no content validity studies and the development studies received an ‘indeterminate’ rating, the overall ratings of the study were determined according to the reviewers’ ratings which lead to increased subjectivity and to a higher likelihood of giving positive or ‘sufficient’ ratings. When content validation studies had been conducted, the measures tended to receive a lower overall score due to a lower quality of the content validity study. Thus, to minimise the bias and ambiguity, we encourage the reader to refer to individual ratings of the development study, content validity study (or studies) and reviewers’ ratings which would give a more accurate overview of the measure’s content validity.

4.3. Structural validity

The risk of bias for most studies assessing antenatal and postnatal measures was rated to be ‘adequate’ or ‘very good’. However, only two measures, namely the PBQ-22 and the MIBS-J-7 which were adapted versions of original measures, showed ‘sufficient’ evidence for structural validity and at least ‘moderate’ quality of evidence. The fact that many widely used measures had ‘insufficient’ evidence for structural validity is problematic and this issue needs to be explored further in future studies. In addition, most papers identified did not report the CFA
estimation method used (e.g., Maximum Likelihood or Weighted Least Squares Mean and Variance), the appropriateness of which depends on several factors (see Rhemtulla, Brosseau-Liard, & Savalei, 2012 for recommendations). We suggest that future studies report this information to increase transparency and facilitate quality ratings.

Separately, we note that if a scale does not demonstrate good structural validity, COSMIN would advise against the consideration of internal consistency as the next step, because one could not be confident in the existence of any subscales or dimensions.

Unlike Chiarotto, Ostelo, Boers and Terwee (2018), who followed the COSMIN guidance strictly by only reporting the content validity and structural validity of the studies, we did report on other psychometric properties of the identified measures.

4.4. Construct validity

Construct validity comprises hypothesis testing, cross-cultural validity/measurement invariance and criterion validity; however, none of the included studies evaluated criterion validity. The risk of bias for hypothesis testing for antenatal and postnatal measures was mostly ‘very good’ with all studies consistently showing ‘high’ quality of evidence. Nevertheless, despite these promising results, only three antenatal measures (MFAS-23, PFAS and K-PAFAS) and two postnatal measures (MABISC and MPAS) offered the best evidence of any hypothesis testing undertaken. It is also important to note that construct validity was not assessed against any ‘gold standard’ self-report measure, since none has yet been identified. Furthermore, none of the studies included in this review used a ‘gold standard’ observer-rated measure of parent-child attachment to assess their measure’s construct validity. This is clearly an area of further investigation.
Measurement invariance was rarely assessed in the identified measures. Based on ‘adequate’ methodological quality (i.e., risk of bias ratings), the MPAS and the paternal equivalent, the PPAS, were assigned ‘sufficient’ ratings with ‘moderate’ quality of evidence. Due to little evidence of construct validity, more research needs to be undertaken.

4.5. Reliability

Except for three antenatal versions (FAB, MAAS-I3 and PAAS-I3) and three postnatal versions (MABISC, PBQ-I6 and MIAS), the quality of evidence for the internal consistency of ten antenatal versions and 13 postnatal versions was considered to be ‘high’. However, most studies did not provide enough information about the internal consistency of the antenatal and postnatal measures assessed and only the PBQ showed ‘sufficient’ evidence for internal consistency. This is because internal consistency could only be rated if the studies demonstrated at least low evidence of ‘sufficient’ structural validity (as per the 2018 COSMIN criteria); if a measure does not demonstrate good structural validity, there is no confidence that those subscales exist. This explains why most of the measures received ‘indeterminate’ ratings. In addition, all of the studies reported internal consistency as Cronbach’s Alpha values. Although this approach is the most popular and widely applied statistic of internal consistency (Dunn, Baguley, & Brunsden, 2013), it has been criticised for having several flaws. For example, it is considered an inappropriate statistic to estimate a scale’s reliability (Peters, 2014) and homogeneity of unidimensionality (Green, Lissitz, & Mulaik, 1977; Schmitt, 1996; Sijtsma, 2009). Thus, researchers evaluating the internal consistency of measures are encouraged to use alternatives, such as McDonald’s omega in the future (Dunn et al., 2013; Peters, 2014).

The reliability of measures could be assessed for two antenatal measures only because information for the remaining antenatal measures was not available. Whilst the reliability for
the PAI-21 was ‘sufficient’ with ‘moderate’ quality of evidence, the reliability for the K-PAFAS was ‘insufficient’ with ‘low’ quality of evidence. Thus, based on the available evidence, the PAI-21 appears to be the most robust antenatal measures to be used with pregnant women.

In relation to the 17 versions of postnatal measures, seven versions (i.e., MAI-26, MPAS, PBQ-25, PBQ-14, MIBS-J-10, MORS-SF and PPAS) evaluated the methodological quality (i.e., risk of bias ratings) of reliability in ten studies which varied between ‘adequate’, ‘doubtful’ and ‘inadequate’ ratings. However, conclusive summaries regarding the methodological quality cannot be made due to these measures presenting with mixed evidence, depending on the country, language and sample size of each conducted study. Of all postnatal measures, only two measures were considered to have ‘sufficient’ evidence for their reliability with ‘moderate’ quality of evidence. Hence, the MPAS and the PBQ-25 appear to be the most reliable postnatal measures.

However, it is important to note that none of the identified studies reported measurement error and responsiveness as part of their assessment of a scale’s additional reliability. Clearly, more work is needed to fully establish the reliability of these scales.

4.6. Clinimetric properties

The assessment of administrative properties of a measure in addition to its psychometric properties has been recommended (e.g., Thornicroft & Slade, 2000; Wittkowski et al., 2017). As it is assumed that their modified and often simplified versions would achieve similar ratings, the clinimetric properties were assessed for the original 17 measures only. Except for three measures (e.g., MPAS, PPAS and MIRFS), the scoring of all measures was rated to be ‘easy’. In addition, none of the measures had an excessive number of items. Hence their completion time was rated as ‘excellent’. With item numbers ranging from seven (e.g., the
MIBS-7) to 26 (e.g., the MAI), all measures could be completed under five minutes, which suggests that they are acceptable and feasible measures, suitable for the use in routine outcome assessments.

In terms of readability and comprehension, all measures (except for the K-PAFAS) obtained ratings in the ‘fair’ to ‘excellent’ range. Of the antenatal measures, the MFAS obtained the best (i.e., ‘good’) score in readability, whereas six of the nine postnatal measures were rated to be ‘good’ or ‘excellent’ and hence easy to understand (e.g., MABISC, PBQ-25, MIAS). Hereby it is noteworthy that the item exploring whether the women or their partners felt that the woman’s body was ‘ugly’ (reverse scored) in the MFAS (and the paternal equivalent, the PFAS) is occasionally removed from the scale because it does not refer to maternal feelings (Müller & Ferketich, 1993; van den Berg & Simons, 2008). With the Flesch readability scores ranging from 55.6 (K-PAFAS) to 96.0 (MIAS), all of the measures appeared to be acceptable.

4.7. Strengths and limitations of the review

A clear strength of this review is its comprehensiveness evidenced by the fact that eight databases were searched and more than 12,000 records were screened, in all languages and publication years. Any measure and study inclusion criteria were determined in advance and registered. In addition, data extraction and evaluation processes were verified by an independent rater and these showed good to excellent inter-rater agreement and reliability. Furthermore, by reporting information pertaining to the clinimetric properties of the identified measures, we assist clinicians and researchers in their assessment of how ‘user-friendly’ a measure is. Furthermore, the strict application of the latest COSMIN criteria (2018) ensured that a rigorous assessment was undertaken of the validity and reliability evidence of the identified measures. Compared to previous reviews in this area (e.g., Perrelli et al., 2014), the
use of the most recent and stringent COSMIN criteria adds substantial credibility to this
detailed assessment of measures. Finally, although the methodological quality of studies
varied, we chose to report those variations rather than exclude those studies.

In terms of limitations, it should be acknowledged that the search was restricted to peer-
reviewed studies only, which introduces a publication bias (Rothstein, 2014). However,
although unpublished measures with good psychometric properties may exist, without being
published appropriately their impact in the field may be minimal. For this reason, Lotzin et al.
(2015) searched all available literature in their review of observational tools for measuring
parent-infant-interaction, but then decided to exclude tools that were only used in one or no
peer-reviewed journal articles.

Although we included different language versions of identified measures in our
evidence synthesis (e.g., K-PAFAS, MIRFS and MIAS), it proved impossible for us to apply the
COSMIN criteria to assess the psychometric properties of measures in identified studies if they
were not written in English. However, for comprehensiveness we offer further information on
those studies, including their sample size and the psychometric properties tested.

In addition, when rating the content validity of studies, we chose to deviate from the
stricter 2018 COSMIN guidelines on four occasions. Firstly, we opted to rate the psychometric
properties of all studies evaluating content validity even if the study’s methodological quality
(i.e., risk of bias) was ‘inadequate’ as this resulted in a detailed and thorough overview of all
included measures. Secondly, we adapted the criteria of rating the risk of bias of the measure
development studies slightly in cases where the authors had presented ‘adequate’ evidence
regarding the study’s conduct; however, this could have resulted in assigning higher risk of
bias ratings in some cases. Thirdly, we modified the criteria for structural validity since many
studies in this review had undertaken EFA but the COSMIN criteria do not include guidance
on how to rate the results of an EFA. Fourthly, we deviated from the guidelines when rating
hypothesis testing and the results of this measurement property should be interpreted with caution.

Another limitation that should be acknowledged is the fact that we excluded measures in which the parent-infant-relationship was examined but only alongside other and arguably less relevant aspects, such as exploring the woman’s body or diet. For example, despite containing seven items that could be said to reflect the mother’s attitude towards the developing fetus, the 60-item Maternal Adjustment and Maternal Attitude (MAMA) questionnaire (Kumar, Robson & Smith, 1984) was excluded because it contained many other items relating to the mother’s perceptions of her body or items of somatic symptoms, the marital relationship, attitudes to sex and attitudes towards pregnancy. The postnatal version of the MAMA was also excluded, although it was judged to contain slightly more relevant items (n = 9). Only 13 of the 26-item What Being The Parent of a Baby is Like (WBTPBL, Pridham & Chang, 1985) scale related directly to the parent-infant-relationship with the other items asking about the new parent’s adaptation to parenthood, relationships with family members and the stress of being a new parent. Finally, although the Maternal Infant Responsiveness Instrument (MIRI; Amankwaa, Younger, Best & Pickler, 2002) partly met our inclusion criteria, the scale mostly examines parental perceptions of baby responsiveness and was therefore excluded.

4.8. Implications for research and practice

The fact that we identified a lack of evidence for robust psychometric properties across a wide variety of antenatal and postnatal parent-report measures is problematic because any conclusions based on these measures will have inherent limitations.

Although some of these measures may have been extensively used (e.g., the PBQ) or their use (e.g., the PBQ and the MORS-SF) may have been recommended (Royal College of Psychiatrists, 2018), it is advisable that clinicians and researchers alike scrutinise each measure
in order to determine if it fits their purpose. For example, all of these measures were validated using predominantly non-clinical populations (with the exception of the PBQ-25 and the MABISC) and this means that clinicians and researchers need to consider a measure’s relevance when applied to their intended population or purpose. Besides, in line with the review’s aims, we only included certain studies but studies may exist that report on a measure’s use with clinical populations (e.g., see Wittkowski, Williams & Wieck, 2010) and possibly on correlations with observer-rated measures as well.

Given the recent proliferation of measures being adapted for use in other countries and in languages other than English, we believe that there is a need for appropriate and more stringent testing for cross-cultural validity. For example, studies with different cultural or ethnic groups should conduct factor analyses for multiple groups (e.g., in English and in Dutch) and complete measurements of invariance or differential item functioning (DIF) to provide information on whether the measures are comparable when used in differing cultural contexts. This could be one of the future directions when testing psychometric properties of the measures.

We also believe that future studies conducting content validity evaluation should describe more explicitly how they evaluated content validity and what aspects they did evaluate and to consult and follow COSMIN criteria when developing the method of a new measure or assessing the method of an already existing measure. This may include conducting a qualitative study (i.e., a focus group or interviews), using appropriate data collection and analysis methods and ideally exploring the relevance, comprehensiveness and comprehensibility of the measure among a sufficient sample of participants and professionals, which would lead to a higher quality and more credible evidence of the measure’s content validity.
5. Conclusion

This is the first systematic review to provide a synthesis of robust validity and reliability evidence for available self-report measures of the parent-infant-relationship. A total of 17 measures and 13 modified versions were identified and evaluated, of which the majority lacked adequate methodological quality despite being widely used and with some being recommended measures. Only the *Postpartum Bonding Questionnaire* (PBQ), and some of its modified versions, were found to demonstrate sufficient evidence for structural validity, internal consistency and reliability with high quality of evidence. The PBQ was also the most frequently adapted tool which is indicative of its perceived relevance and popularity in this field. However, due to the inadequate methodological quality and insufficient psychometric measurement evaluation of most measures, in addition to the lack of comprehensive psychometric evaluation of many measures, firm conclusions regarding the most valid and reliable parent-infant-relationship measure(s) cannot be drawn.

The current review is important and timely given the increasing importance of routine self-report outcome monitoring within a range of perinatal services and within research studies (e.g., NHS England, 2018). Despite the wealth of antenatal and postnatal measures, the psychometric properties of these tools remain poor and understudied. It is advisable that future researchers developing new or modified measures follow the current COSMIN guidelines and that research into evaluation of psychometric properties would continue in order to bring measures to the industry standard and facilitate the selection of the most robust antenatal and postnatal measures by researchers and clinicians.
Author Disclosure Policy

Role of funding sources
This review was undertaken as part of the THRIVE project funded by the National Institute for Health Research (NIHR) Public Health Research Programme (PHR Project: 11/3002/01), led by MH. As part of this grant, CG, AM and SV were employed as researchers and AW was their supervisor as well as Manchester lead for the THRIVE project. The funder had no direct involvement in the conduct of this review. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. MH was supported by the MRC/CSO Quinquennial funding of the Relationships and Health Improvement Programme, which is part of the Social and Public Health Sciences Unit, based at University of Glasgow, MC_UU_12017-11 and SPHSU11.

Contributions
AW devised the idea for this review and oversaw the whole review process. CG and AW wrote the initial protocol which was later updated with SV and AM’s assistance. SV, CG and AM conducted the literature searches. AM and SV rated the included studies, under AW’s supervision. Any disagreements were resolved through consultation with the fifth reviewer, MH. All authors contributed to drafts and approved the final manuscript.

Conflict of interest
All authors declare that they have no conflict of interest.

Acknowledgements
The authors wish to thank Catherine Shone for assisting with double-screening articles and rating content validity as an independent second rater.
References


Table 1: Definitions and criteria for good measurement properties*

<table>
<thead>
<tr>
<th>Measurement property</th>
<th>Definition</th>
<th>Rating</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validity</strong> (the degree to which a participant-reported outcome measure (PROM) measures the construct(s) it purports to measure)**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Content validity** (includes relevance, comprehensiveness and comprehensibility)
- The degree to which the content of a PROM is an adequate reflection of the construct to be measured. Important aspects are whether all items are relevant for the construct, aim and target population and if no important items are missing (comprehensiveness).
  - + Above 85% of the items of the PROM or subscale are relevant for the construct of interest AND are relevant for the target population of interest AND are relevant for the context of use of interest AND have appropriate response options OR have appropriate recall period AND include all key concepts AND together comprehensively reflect the construct to be measured.
  - ? No or not enough information available or quality of the study inadequate
  - - Less than 85% of the items of the PROM or subscale fulfil the criterion

**Structural validity**
- Part of construct validity alongside hypothesis testing and cross-cultural validity, structural validity is the degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured.
  - + **Classical Test Theory (CTT):** Confirmatory Factor Analysis (CFA): comparative fit index (CFI) or Tucker-Lewis index (TLI) or comparable measure > 0.95 AND Root Mean Square Error of Approximation (RMSEA) < 0.06 OR Standardized Root Mean Residuals (SRMR) < 0.08 **[factor structures should be equal across studies]** (Schermelleh-Engel, Moosbrugger, & Müller, 2003) OR **IRT/Rasch:** No violation of unidimensionality: CFI or TLI or comparable measure > 0.95 (Schermelleh-Engel et al., 2003) OR RMSEA < 0.06 (Schermelleh-Engel et al., 2003) OR SRMR < 0.08 (Schermelleh-Engel et al., 2003) **AND** no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3’s < 0.37 **AND** no violation of monotonicity: adequate looking graphs.
| Hypotheses testing for construct validity | OR item scalability > 0.30 *(Stochl, Jones, & Croudace, 2012)*  
AND adequate model fit: Item Response Theory (IRT): \( \chi^2 > 0.01 \) *(Maydeu-Olivares, 2013)*  
Rasch: infit and outfit mean squares \( \geq 0.5 \) and \( \leq 1.5 \) *(Linacre, 2002)*  
OR Z-standardized values \( > 2 \) and \( < 2 \)  
? | CTT: Not all information for ‘+’ reported  
IRT/Rasch: Model fit not reported  
- Criteria for ‘+’ not met |
|---|---|
| Part of construct validity alongside structural validity and cross-cultural validity, hypothesis testing is the degree to which the scores of a PROM are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PROM validly measures the construct to be measured. | + | At least 75% of the result is in accordance with the hypothesis  
? | No hypothesis defined (by the review team)  
- | Less than 75% of the result is in accordance with the hypothesis |
| Cross-cultural validity/ measurement invariance | Part of construct validity alongside structural validity and hypothesis testing, cross-cultural validity is the degree to which the performance of the items on a translated or culturally adapted PROM are an adequate reflection of the performance of the items of the original version of the PROM. | + | No important differences found between group factors (such as age, gender, language) in multiple group factor analysis  
OR no important differential item functioning (DIF) for group factors (McFadden's R2 < 0.02)  
? | No multiple group factor analysis  
OR DIF analysis performed  
- | Important differences between group factors  
OR DIF was found |
| Criterion validity | The degree to which the scores of a PROM are an adequate reflection of a ‘gold standard’. | + | Correlation with gold standard \( \geq 0.70 \)  
OR area under the curve (AUC) \( \geq 0.70 \)  
? | Not all information for ‘+’ reported  
- | Correlation with gold standard < 0.70  
OR AUC < 0.70 |
### Reliability (the degree to which the measurement is free from measurement error)

<table>
<thead>
<tr>
<th>Property</th>
<th>Definition</th>
<th>+ Criteria</th>
<th>- Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal consistency</strong></td>
<td>The degree of the interrelatedness among the items.</td>
<td>At least low evidence (as per GRADE) for sufficient structural validity AND Cronbach’s alpha(s) ( \geq 0.70 ) for each unidimensional scale or subscale.</td>
<td>Criteria for “At least low evidence (as per GRADE) for sufficient structural validity” not met.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>At least low evidence (as per GRADE) for sufficient structural validity AND Cronbach’s alpha(s) &lt; 0.70 for each unidimensional scale or subscale.</td>
<td></td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>The proportion of the total variance in the measurements which is due to ‘true’ differences between patients.</td>
<td>Intraclass correlation coefficient (ICC), weighted Kappa or Pearson or Spearman correlation coefficient ( \geq 0.70 )</td>
<td>ICC, weighted Kappa, Pearson or Spearman correlation coefficient not reported.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICC, weighted Kappa, Pearson or Spearman correlation coefficient &lt; 0.70.</td>
<td></td>
</tr>
<tr>
<td><strong>Measurement error</strong></td>
<td>The systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured.</td>
<td>Smallest Detectable Change (SDC) or Limits of Agreement (LoA) &lt; Minimal Important Change (MIC)</td>
<td>MIC not defined.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SDC or LoA &gt; MIC.</td>
<td></td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>The ability of a PROM to detect change over time in the construct to be measured.</td>
<td>At least 75% of the result is in accordance with the hypothesis OR AUC ( \geq 0.70 )</td>
<td>No hypothesis defined (by the review team).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Less than 75% of the result is in accordance with the hypothesis OR AUC &lt; 0.70.</td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations.** PROM – patient- or participant-reported outcome measure.

**Note.*** - adapted from Mokkink et al. (2009, 2018) & Prinsen et al. (2018). The measurement properties in italics were not considered in the present review, because none of the included studies assessed these properties as outlined in the COSMIN criteria.
Table 2: Definitions of quality levels using the GRADE approach

<table>
<thead>
<tr>
<th>Quality level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong></td>
<td>We are very confident that the true measurement property lies close to that of the estimate of the measurement property.</td>
</tr>
<tr>
<td><strong>Moderate</strong></td>
<td>We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property but there is a possibility that it is substantially different.</td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property.</td>
</tr>
<tr>
<td><strong>Very low</strong></td>
<td>We have very little confidence in the measurement property: the true measurement property is likely to be substantially different from the estimate of the measurement property.</td>
</tr>
</tbody>
</table>

Notes. Estimate of the measurement property refers to the pooled or summarized result of the measurement property of a given measure. These definitions were adapted from the GRADE approach.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Descriptive information</th>
<th>Clinimetric information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANTENATAL MEASURES</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Measure</td>
<td>Related measures</td>
</tr>
<tr>
<td><strong>Maternal measures</strong></td>
<td>The How I Feel About My Baby Now Scale (FAB, Leifer, 1977)</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>The Maternal-Fetal Attachment Scale (MFAS-24, Cranley, 1981)</td>
<td>PFAS</td>
</tr>
<tr>
<td></td>
<td>The Prenatal Attachment Inventory (PAI-21, Müller, 1993)</td>
<td>MAI</td>
</tr>
<tr>
<td></td>
<td>The Maternal Antenatal Attachment Scale (MAAS-19, Condon, 1993)</td>
<td>PAAS</td>
</tr>
<tr>
<td>Scale Name</td>
<td>Language</td>
<td>Country</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>-----------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>The Prenatal and the Postnatal Scale (PPBS, Cuijlits, 2016)</td>
<td>None</td>
<td>Dutch/ English (1050)</td>
</tr>
<tr>
<td>The Paternal-Fetal Attachment Scale (PFAS, Weaver &amp; Cranley, 1983)</td>
<td>None</td>
<td>English/ English (100)</td>
</tr>
<tr>
<td>The Paternal Antenatal Attachment Scale (PAAS, Condon, 1993)</td>
<td>MAAS</td>
<td>English/ 112</td>
</tr>
<tr>
<td>The Korean Paternal-Fetal Attachment Scale (K-PAFAS, Noh &amp; Yeom, 2017)</td>
<td>None</td>
<td>Korean/ 200</td>
</tr>
<tr>
<td>Scale</td>
<td>Country</td>
<td>Language(s)</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>The Maternal Attachment Inventory (MAI, Müller, 1994)</strong></td>
<td>Healthy USA mothers with healthy babies of 4 months &amp; 8 months</td>
<td>English</td>
</tr>
<tr>
<td><strong>The Mother Infant Attachment Scale (MIAS, Bhakoo et al., 1994)</strong></td>
<td>Healthy women with healthy babies of ≤ 8 months</td>
<td>None</td>
</tr>
<tr>
<td><strong>The Mother and Baby Interaction Scale (MABISC, Hackney et al, 1996)</strong></td>
<td>Healthy women and women with postnatal depression</td>
<td>English</td>
</tr>
<tr>
<td>Scale</td>
<td>Short Form</td>
<td>Subscales</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>The Maternal Postnatal Attachment Scale (MPAS, Condon &amp; Corkindale, 1998)</td>
<td>MAAS, PAAS, PPAS</td>
<td>None</td>
</tr>
<tr>
<td>The Postpartum Bonding Questionnaire (PBQ-25, Brockington et al., 2001)</td>
<td>Short PBQ</td>
<td>PBQ-22; PBQ-16; PBQ-14</td>
</tr>
<tr>
<td>The Mother-to-Infant Bonding Scale (MIBS-8, Taylor et al., 2005)</td>
<td>None</td>
<td>MIBS-10</td>
</tr>
<tr>
<td>Instrument</td>
<td>Language</td>
<td>Country</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>The Mother-to-Infant Relations and Feelings Scale (MIRFS, Thorstensson et al., 2012a)</td>
<td>Swedish/ English</td>
<td>Sweden</td>
</tr>
<tr>
<td>The Mothers’ Object Relations Scales-Short Form (MORS-SF, Oates &amp; Gervai, 2018)</td>
<td>English</td>
<td>UK &amp; Hungary</td>
</tr>
<tr>
<td>The Paternal Postnatal Attachment Scale (PPAS, Condon, Corkindale, &amp; Boyce, 2008)</td>
<td>English</td>
<td>Australia</td>
</tr>
</tbody>
</table>
Table 4: Quality of the measure development (n = 17 measures) and content validity (n = 16 measures)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Design</th>
<th>Cognitive interview (CI) study</th>
<th>Content validity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General design requirements</td>
<td>General design requirements</td>
<td>TOTAL DEVELOPMENT</td>
</tr>
<tr>
<td></td>
<td>Clear construct</td>
<td>Clear origin of construct</td>
<td>Clear target population</td>
</tr>
<tr>
<td></td>
<td>Concept elicitation</td>
<td>CI study performed in sample representing the target population</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td><strong>FAB</strong> (Leifer, 1977)</td>
<td>V</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td><strong>MFAS</strong> (Cranley, 1981)</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td><strong>MFAS</strong> (Busonera et al., 2016a)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td><strong>MFAS</strong> (Lingeswaram &amp; Bindu, 2012)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td><strong>PAI</strong> (Müller, 1993)</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td><strong>PAI</strong> (Samani et al., 2016)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>4</td>
<td><strong>MAAS</strong> (Condon, 1993)</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td><strong>MAAS</strong> (Golbasi et al., 2015)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>5</td>
<td><strong>PPBS</strong> (Cuijlits et al., 2016)</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>6</td>
<td><strong>PFAS</strong> (Weaver &amp; Cranley, 1983)</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>7</td>
<td><strong>PAAS</strong> (Condon, 1993)</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td><strong>PAAS</strong> (Della Vedova et al., 2017)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>8</td>
<td><strong>K-PAFAS</strong> (Noh &amp; Yeom, 2017)</td>
<td>V</td>
<td>V</td>
</tr>
</tbody>
</table>

**ANTENATAL MEASURES**

**Maternal measures**

**Paternal measures**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>MAI (Müller, 1994)</th>
<th>V</th>
<th>V</th>
<th>V</th>
<th>V</th>
<th>A</th>
<th>D</th>
<th>I</th>
<th>—</th>
<th>—</th>
<th>—</th>
<th>I</th>
<th>I</th>
<th>I</th>
<th>D</th>
<th>D</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td></td>
<td>MAI (Shin &amp; Kim, 2007)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAI (Chen et al., 2013)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>MIAS (Bhakoo et al., 1994)</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>D</td>
<td>D</td>
<td>V</td>
<td>I</td>
<td>D*</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>MABISC (Hackney et al., 1996)</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>A</td>
<td>D</td>
<td>V</td>
<td>D</td>
<td>D*</td>
<td>D</td>
<td>D</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MPAS (Riera-Martin et al., 2018)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>PBQ (Brockington et al., 2001)</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>I</td>
<td>—</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBQ (Vengadavaradan et al., 2019)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>I</td>
<td>I</td>
<td>D</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBQ (Siu et al., 2010)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>D</td>
<td>I</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBQ (Baldisserotto et al., 2018)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>D</td>
<td>I</td>
<td>D</td>
<td>D</td>
<td>I</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>MIBS (Taylor et al., 2005)</td>
<td>I</td>
<td>D</td>
<td>V</td>
<td>V</td>
<td>I</td>
<td>—</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>MIRFS (Ekström &amp; Nissen, 2006)</td>
<td>I</td>
<td>D</td>
<td>I</td>
<td>D</td>
<td>A</td>
<td>D</td>
<td>V</td>
<td>I</td>
<td>D*</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MIRFS (Thorstenson, Hertfelt et al., 2012a)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>D</td>
<td>I</td>
<td>D</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MIRFS (Thorstenson, Nissen, &amp; Ekström, 2012b)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>D</td>
<td>I</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>MORS-SF (Oates &amp; Gervai, 2018)</td>
<td>V</td>
<td>D</td>
<td>V</td>
<td>D</td>
<td>I</td>
<td>—</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>PPAS (Condon et al., 2008)</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>D</td>
<td>V</td>
<td>D</td>
<td>D*</td>
<td>D</td>
<td>D</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPAS (Riera-Martin et al., 2018)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>A</td>
</tr>
</tbody>
</table>

Notes. **very good**: adequate; doubtful; inadequate; not reported by the study authors. * - not clear or not assessed by the studies were rated as ‘doubtful’.
Table 5: Synthesis of psychometric properties and quality of evidence (using GRADE)*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Content validity</th>
<th>Structural validity</th>
<th>Internal consistency</th>
<th>Hypothesis testing</th>
<th>Measurement invariance</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relevance</td>
<td>Comprehensiveness</td>
<td>Rating of results</td>
<td>Quality of evidence</td>
<td>Rating of results</td>
<td>Quality of evidence</td>
</tr>
<tr>
<td>FAB</td>
<td>+</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>–</td>
<td>HIGH</td>
</tr>
<tr>
<td>MFAS-24</td>
<td>+</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MFAS-23</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MFAS-20</td>
<td>+</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MFAS-17</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PAI-21</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MAAS-19</td>
<td>(+)</td>
<td>VERY LOW</td>
<td>[+</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MAAS-13</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MAAS-12</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PPBS</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+</td>
<td>HIGH</td>
</tr>
<tr>
<td>PFAS</td>
<td>+</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PAAS-19</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PAAS-13</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>K-PAFAS</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+</td>
<td>HIGH</td>
</tr>
<tr>
<td>MAI-26</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MIAS</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MACBSC</td>
<td>[+]</td>
<td>LOW</td>
<td>[+]</td>
<td>LOW</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MPAS</td>
<td>[+]</td>
<td>LOW</td>
<td>[+]</td>
<td>LOW</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PBQ-25</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PBQ-16</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>PBQ-22</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PBQ-16-J</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PBQ-14</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PBQ-19</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>PBQ</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>S-PBQ</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MIBS-8</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MIBS-10</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MIBS-J-7</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MIRFS</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>MAAS-4F</td>
<td>[+]</td>
<td>VERY LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
<tr>
<td>Fat</td>
<td>[+]</td>
<td>LOW</td>
<td>[+]</td>
<td>HIGH</td>
<td>[+]</td>
<td>HIGH</td>
</tr>
</tbody>
</table>

Notes: [+]=sufficient; [±]=insufficient; [?] = indeterminate; [±]=inconsistent; - not reported by the study authors. * - some studies also tested for diagnostic accuracy (i.e. sensitivity and specificity of the measure in detecting bonding difficulties) but it is not included within the COSMIN taxonomy and thus not rated. Structural validity ratings were based on the best fitting model presented in the paper (this was not necessarily the factor structure proposed by the original authors). As per the COSMIN criteria, internal consistency could only be rated as sufficient if there was at least low evidence of sufficient structural validity (otherwise an indeterminate rating was assigned). PBQ-J-16 contains different items from the PBQ-16. When results from two or more studies are in conflict, the results of higher quality studies superseded those of lower quality studies (unless there was a clear possible reason for conflicting results, in which case they would be treated separately).
Figure 1: Flowchart of paper selection based on PRISMA guidance

Total records identified (n = 15924)

Records identified through Ovid (n = 6557)
Records identified through CINAHL (n = 7612)
Records identified through Web of Science (n = 1755)

Records after duplicates removed (n = 12081)

Records screened (n = 12081)

Records excluded by title and abstract (n = 11861)

Articles excluded (n = 163):
- Not a developmental or validation study (n = 34)
- Not primary research (n = 16)
- Not peer-reviewed (n = 6)
- Article not in English (n = 10)
- Not measuring the parent-infant relationship (n = 57)
- Not relating to antenatal or postnatal period (n = 1)
- Not completed by parents (n = 2)
- Not self-report, e.g. observational tool or interview (n = 27)
- Items not relevant in the measure (n = 10)
- Measure not in English (n = 1)

220 articles assessed for eligibility using full text

17 included original measures and 13 modified version in 65 eligible and included articles
Appendix A: Example search strategy in OVID

1. ((parent* or maternal or paternal or mother* or father*) adj7 (child or infant or newborn or foet* or fetus or fetal or baby or neonate)).mp
2. (antenat* or prenat* or puerper* or postnat* or postpart* or peripartum or pregnan* or perinat*).mp
3. ((measure* or scale$ or questionnaire$ or construct$ or tool$ or inventor* or instrument$ or test*) adj7 (attachment or relation* or bond* or orientation or synchrony or synchronicity or “emotional availability” or attitude* or belief* or responsive* or feel* or interact*)).mp
4. 1 and 2 and 3
5. limit 4 to all journals
6. limit 5 to (female or humans or male)
7. limit 6 to peer reviewed journal
8. limit 7 to original articles
9. limit 8 to (2200 psychometrics & statistics & methodology or 2220 tests & testing or 2222 developmental scales & schedules or 2223 personality scales & inventories or 2224 clinical psychological testing or 2225 neuropsychological assessment or 2226 health psychology testing or 2240 statistics & mathematics or 2260 research methods & experimental design or 2520 neuropsychology & neurology or 2600 psychology & the humanities or 2840 psychosocial & personality development or 3000 social psychology or 3040 social perception & cognition or 3360 health psychology & medicine)
Appendix B: Overview and reference list of non-English studies identified in the systematic search

<table>
<thead>
<tr>
<th>Measure(s) evaluated</th>
<th>Authors</th>
<th>Language</th>
<th>Sample size</th>
<th>Psychometric properties tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFAS</td>
<td>Lauriola et al. (2010)</td>
<td>Italian</td>
<td>254 women</td>
<td>Internal reliability; construct validity</td>
</tr>
<tr>
<td>MFAS</td>
<td>Andrek et al (2016)</td>
<td>Hungarian</td>
<td>114 women</td>
<td>Internal reliability</td>
</tr>
<tr>
<td>PAI</td>
<td>Jurgens et al. (2009)</td>
<td>French</td>
<td>112 women</td>
<td>Internal reliability; construct validity</td>
</tr>
<tr>
<td>PAI</td>
<td>Lauriola et al. (2010)</td>
<td>Italian</td>
<td>254 women</td>
<td>Internal reliability; construct validity</td>
</tr>
<tr>
<td>PAI</td>
<td>Pavše et al. (2019)</td>
<td>Slovenian</td>
<td>619 women</td>
<td>Internal reliability; structural validity</td>
</tr>
<tr>
<td>MAAS</td>
<td>Camarneiro &amp; Justo (2010)</td>
<td>Portuguese</td>
<td>212 couples</td>
<td>Internal reliability; structural validity</td>
</tr>
<tr>
<td>MAAS</td>
<td>Lauriola et al. (2010)</td>
<td>Italian</td>
<td>254 women</td>
<td>Internal reliability; construct validity</td>
</tr>
<tr>
<td>MAAS</td>
<td>Denis et al (2013)</td>
<td>French</td>
<td>117 women</td>
<td>Internal reliability; construct validity; concurrent validity; divergent validity</td>
</tr>
<tr>
<td>MAAS</td>
<td>Nie &amp; Fan (2017)</td>
<td>Chinese</td>
<td>545 women</td>
<td>Internal reliability; convergent validity</td>
</tr>
<tr>
<td>PAAS</td>
<td>Camarneiro &amp; Justo (2010)</td>
<td>Portuguese</td>
<td>212 couples</td>
<td>Internal reliability; structural validity</td>
</tr>
<tr>
<td>MAI</td>
<td>Kavlak &amp; Sirin, 2009</td>
<td>Turkish</td>
<td>165 women</td>
<td>Internal reliability; content validity</td>
</tr>
<tr>
<td>PBQ</td>
<td>Yalcin et al. (2014)</td>
<td>Turkish</td>
<td>189 women</td>
<td>Internal reliability</td>
</tr>
<tr>
<td>MIBS</td>
<td>Figueiredo et al. (2005)</td>
<td>Portuguese</td>
<td>456 parents</td>
<td>Internal reliability; test-retest reliability</td>
</tr>
<tr>
<td>MIBS</td>
<td>Yalcin et al. (2014)</td>
<td>Turkish</td>
<td>189 women</td>
<td>Internal reliability</td>
</tr>
<tr>
<td>MORS-SF</td>
<td>Danis et al. (2012)</td>
<td>Hungarian</td>
<td>1164 parents</td>
<td>Internal reliability</td>
</tr>
</tbody>
</table>


Appendix C: Reference list of included studies (n = 65)


Appendix D: Detailed content validity ratings according to type of study/rating.

<table>
<thead>
<tr>
<th>Outcome measure development</th>
<th>RELEVANCE</th>
<th>COMPREHENSIVENESS</th>
<th>COMPREHENSIBILITY</th>
<th>TOTAL CONTENT VALIDITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAB</td>
<td>[?] None</td>
<td>[++; +] +</td>
<td>VERY LOW</td>
<td>[+]</td>
</tr>
<tr>
<td>MFAS-24</td>
<td>[?; ±; ±] [++; +] +</td>
<td>VERY LOW</td>
<td>[?] [?; ?] [++; +] +</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MFAS-20</td>
<td>NA [±] [++; +] +</td>
<td>VERY LOW</td>
<td>NA [?] [++; ±] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>PAI</td>
<td>[?] [+] [++; +] +</td>
<td>VERY LOW</td>
<td>[?] [?] [++; +] [+]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MAAS</td>
<td>[+] [±] [++; +] +</td>
<td>VERY LOW</td>
<td>[+] [?] [++; +] [+]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>PPBS</td>
<td>[+] None  [++; +] +</td>
<td>VERY LOW</td>
<td>[-] None [++; -] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>PFAS</td>
<td>[?] None  [++; +] +</td>
<td>VERY LOW</td>
<td>[+; None [++; +] [+]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>PAAS</td>
<td>[+] [-] [++; +] +</td>
<td>VERY LOW</td>
<td>[+] [-] [++; +] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>K-PAFAS</td>
<td>[+] [±] [++; +] +</td>
<td>VERY LOW</td>
<td>[+] [±] [++; +] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MAI</td>
<td>[?] [±; ±] [++; +] +</td>
<td>VERY LOW</td>
<td>[+] [?; ?] [++; +] [+]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MIAS</td>
<td>[?] None  [++; ±] [±]</td>
<td>VERY LOW</td>
<td>[+] None [++; +] [+]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MABISC</td>
<td>[+] None  [++; +] +</td>
<td>LOW</td>
<td>[+; None [++; +] [+]</td>
<td>LOW</td>
</tr>
<tr>
<td>MPAS</td>
<td>[+] [±] [++; +] +</td>
<td>LOW</td>
<td>[+; -] [++; +] + [±]</td>
<td>LOW</td>
</tr>
<tr>
<td>PBQ-25</td>
<td>[±; ±; +] [++; +] +</td>
<td>VERY LOW</td>
<td>[+] [++; ±] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>PBQ-19</td>
<td>NA [-] [++; +] ±</td>
<td>VERY LOW</td>
<td>NA [?] [++; +] [+]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MIIBS</td>
<td>[±] None  [++; +] +</td>
<td>VERY LOW</td>
<td>[-] None [++; +] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MIRFS</td>
<td>[?] [±; ±] [++; +] +</td>
<td>VERY LOW</td>
<td>[+] [?; ±] [++; +] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>MORS-SF</td>
<td>[?] None  [++; +] +</td>
<td>VERY LOW</td>
<td>[-] None [++; +] [±]</td>
<td>VERY LOW</td>
</tr>
<tr>
<td>PPAS</td>
<td>[+] [±] [++; +] +</td>
<td>LOW</td>
<td>[+; -] [++; +] [±]</td>
<td>LOW</td>
</tr>
</tbody>
</table>

Notes: [+] = sufficient. [?] = indeterminate. [±] = inconsistent. [-] = insufficient. NA – not applicable. None = no content validity studies conducted. The multiple ratings per box indicate either multiple content validity studies or multiple reviewers’ ratings.
### Appendix E: Risk of bias and measurement property results rated by study

<table>
<thead>
<tr>
<th>#</th>
<th>Type 1</th>
<th>Type 2</th>
<th>Measure</th>
<th>Language of scale</th>
<th>Paper (by first author and year)</th>
<th>n</th>
<th>Structural validity</th>
<th>n</th>
<th>Internal consistency</th>
<th>n</th>
<th>Reliability</th>
<th>n</th>
<th>Hypothesis testing</th>
<th>n</th>
<th>Measurement invariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pre</td>
<td>Mat</td>
<td>FAB</td>
<td>English</td>
<td>Leifer, 1977</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>Pre</td>
<td>Mat</td>
<td>MFAS-24</td>
<td>English</td>
<td>Cranley, 1981</td>
<td>–</td>
<td>–</td>
<td>71</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>71</td>
<td>Inadequate [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>Pre</td>
<td>Mat</td>
<td>Tamil</td>
<td>Language of scale</td>
<td>Languageswaran, 2012</td>
<td>–</td>
<td>–</td>
<td>230</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>71</td>
<td>Inadequate [-]</td>
</tr>
<tr>
<td>4</td>
<td>Pre</td>
<td>Mat</td>
<td>Hungarian</td>
<td>Andrek, 2016</td>
<td>Doster, 2018</td>
<td>324</td>
<td>Adequate [-]</td>
<td>324</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>324</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>5</td>
<td>Pre</td>
<td>Mat</td>
<td>MFAS-23</td>
<td>English</td>
<td>Müller and Ferketch, 1993</td>
<td>681</td>
<td>Adequate [-]</td>
<td>681</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>6</td>
<td>Pre</td>
<td>Mat</td>
<td>English</td>
<td>Müller, 1993</td>
<td>–</td>
<td>336</td>
<td>Adequate [?]</td>
<td>336</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>336</td>
<td>Very good [+]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>Pre</td>
<td>Mat</td>
<td>MFAS-20</td>
<td>Italian</td>
<td>Busonera, 2016a</td>
<td>482</td>
<td>Very good [-]</td>
<td>482</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>482</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>8</td>
<td>Pre</td>
<td>Mat</td>
<td>MFAS-17</td>
<td>Swedish</td>
<td>Seimyr, 2009</td>
<td>298</td>
<td>Adequate [-]</td>
<td>298</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>298</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>9</td>
<td>Pre</td>
<td>Mat</td>
<td>Swedish</td>
<td>Sjögren, 2004</td>
<td>–</td>
<td>76</td>
<td>Inadequate [+]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>76</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>10</td>
<td>Pre</td>
<td>Mat</td>
<td>PAI-21</td>
<td>English</td>
<td>Müller, 1993</td>
<td>336</td>
<td>Adequate [-]</td>
<td>336</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>336</td>
<td>Adequate [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>11</td>
<td>Pre</td>
<td>Mat</td>
<td>English</td>
<td>Gau, 2003</td>
<td>–</td>
<td>349</td>
<td>Very good [-]</td>
<td>349</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>349</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>12</td>
<td>Pre</td>
<td>Mat</td>
<td>Swedish</td>
<td>Pallant, 2014</td>
<td>–</td>
<td>775</td>
<td>Very good [-]</td>
<td>775</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>775</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>13</td>
<td>Pre</td>
<td>Mat</td>
<td>Italian</td>
<td>Barone, 2014</td>
<td>–</td>
<td>130</td>
<td>Adequate [+]</td>
<td>130</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>130</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>14</td>
<td>Pre</td>
<td>Mat</td>
<td>Italian</td>
<td>Busonera, 2017a</td>
<td>–</td>
<td>535</td>
<td>Very good [-]</td>
<td>535</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>535</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>15</td>
<td>Pre</td>
<td>Mat</td>
<td>Italian</td>
<td>Della Vedova, 2008</td>
<td>–</td>
<td>214</td>
<td>Adequate [-]</td>
<td>214</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>214</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>16</td>
<td>Pre</td>
<td>Mat</td>
<td>Persian</td>
<td>Samani, 2016</td>
<td>–</td>
<td>322</td>
<td>Very good [-]</td>
<td>322</td>
<td>Very good [?]</td>
<td>322</td>
<td>Adequate [+]</td>
<td>322</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>17</td>
<td>Pre</td>
<td>Mat</td>
<td>English</td>
<td>Condon, 1993</td>
<td>–</td>
<td>150</td>
<td>Adequate [-]</td>
<td>150</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
<td>150</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>18</td>
<td>Pre</td>
<td>Mat</td>
<td>English</td>
<td>Condon, 1997</td>
<td>–</td>
<td>150</td>
<td>Adequate [-]</td>
<td>150</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
<td>150</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>19</td>
<td>Pre</td>
<td>Mat</td>
<td>Italian</td>
<td>Busonera, 2016b</td>
<td>–</td>
<td>482</td>
<td>Very good [-]</td>
<td>482</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>482</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>20</td>
<td>Pre</td>
<td>Mat</td>
<td>Spanish</td>
<td>Navarro-Aresti, 2016</td>
<td>–</td>
<td>525</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>525</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>21</td>
<td>Pre</td>
<td>Mat</td>
<td>Turkish</td>
<td>Golbasi, 2015</td>
<td>–</td>
<td>190</td>
<td>Adequate [-]</td>
<td>190</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>190</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>22</td>
<td>Pre</td>
<td>Mat</td>
<td>Hungarian</td>
<td>Mako Deak, 2014</td>
<td>–</td>
<td>237</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>237</td>
<td>Very good [+]</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pre</td>
<td>Mat</td>
<td>Dutch</td>
<td>Van Bussel, 2010b</td>
<td>–</td>
<td>403</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>403</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>24</td>
<td>Pre</td>
<td>Mat</td>
<td>German</td>
<td>Göbel, 2019</td>
<td>–</td>
<td>263</td>
<td>Adequate [-]</td>
<td>263</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>263</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>25</td>
<td>Pre</td>
<td>Mat</td>
<td>Spanish</td>
<td>Navarro-Aresti, 2016</td>
<td>–</td>
<td>525</td>
<td>Very good [-]</td>
<td>525</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>525</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>26</td>
<td>Pre</td>
<td>Pat</td>
<td>PPBS</td>
<td>Dutch</td>
<td>Cuijpers, 2016</td>
<td>1050</td>
<td>Very good [+]</td>
<td>529</td>
<td>Very good [+]</td>
<td>–</td>
<td>–</td>
<td>1050</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>27</td>
<td>Pre</td>
<td>Pat</td>
<td>PFAS</td>
<td>English</td>
<td>Weaver, 1983</td>
<td>–</td>
<td>–</td>
<td>100</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
<td>100</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>28</td>
<td>Pre</td>
<td>Pat</td>
<td>Swedish</td>
<td>Seimyr, 2009</td>
<td>–</td>
<td>298</td>
<td>Adequate [-]</td>
<td>298</td>
<td>Very good [?]</td>
<td>–</td>
<td>–</td>
<td>298</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>29</td>
<td>Pre</td>
<td>Pat</td>
<td>PAAS-19</td>
<td>English</td>
<td>Condon, 1993</td>
<td>112</td>
<td>Doubtful [-]</td>
<td>112</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
<td>112</td>
<td>Inadequate [?]</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>Test</td>
<td>Year</td>
<td>Adequate</td>
<td>Very good</td>
<td>Doubtful</td>
<td>Inadequate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>------------</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>English</td>
<td>Pre Pat</td>
<td>Condon, 2013</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>204 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Italian</td>
<td>Pre Pat</td>
<td>Della Vedova, 2017</td>
<td>65</td>
<td>Doubtful [-]</td>
<td>65</td>
<td>Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>German</td>
<td>Pre Pat</td>
<td>Gobel, 2019</td>
<td>128</td>
<td>Adequate [-]</td>
<td>128</td>
<td>Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Korean</td>
<td>Pre Pat</td>
<td>Noh, 2017</td>
<td>200</td>
<td>Very good [-]</td>
<td>200</td>
<td>Very good [?]</td>
<td>200 Doubtful [+]</td>
<td>200 Very good [+]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>English</td>
<td>MAI-26</td>
<td>Müller, 1994²</td>
<td>-</td>
<td>196</td>
<td>Very good [?]</td>
<td>148 Doubtful [+]</td>
<td>196 Adequate [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>English</td>
<td>Post Mat</td>
<td>Müller, 1996</td>
<td>-</td>
<td>196</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>196 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Korean</td>
<td>Post Mat</td>
<td>Shin, 2007</td>
<td>196</td>
<td>Adequate [+]</td>
<td>196</td>
<td>Very good [-]</td>
<td>-</td>
<td>-</td>
<td>196 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Chinese</td>
<td>Post Mat</td>
<td>Chen, 2013</td>
<td>181</td>
<td>Adequate [+]</td>
<td>181</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>181 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Hindi</td>
<td>Post Mat</td>
<td>Bhakoo, 1994</td>
<td>100</td>
<td>Inadequate [?]</td>
<td>100 Adequate [?]</td>
<td>-</td>
<td>-</td>
<td>Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>English</td>
<td>Post Mat</td>
<td>Hackney, 1996</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Inadequate [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Norwegian</td>
<td>Post Mat</td>
<td>Hoivik, 2013</td>
<td>76</td>
<td>Doubtful [+]</td>
<td>76</td>
<td>Doubtful [?]</td>
<td>-</td>
<td>-</td>
<td>76 Very good [+]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>English</td>
<td>Post Mat</td>
<td>Condon, 1998</td>
<td>212</td>
<td>Adequate [-]</td>
<td>212</td>
<td>Very good [?]</td>
<td>212 Adequate [+]</td>
<td>212 Very good [+]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>English</td>
<td>Post Mat</td>
<td>Feldstein, 2004</td>
<td>-</td>
<td>59</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>59 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>Italian</td>
<td>Post Mat</td>
<td>Scopesi, 2004</td>
<td>210</td>
<td>Very good [-]</td>
<td>210</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>Inadequate [-]</td>
<td>Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Dutch</td>
<td>Post Mat</td>
<td>Van Bussel, 2010a</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>263 Very good [+]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Spanish</td>
<td>Post Mat</td>
<td>Riera-Martín, 2016</td>
<td>571</td>
<td>Very good [-]</td>
<td>571</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Very good [?]</td>
<td>571 Adequate [+]</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>English</td>
<td>Post Mat</td>
<td>Brockington, 2001</td>
<td>104</td>
<td>Adequate [+]</td>
<td>-</td>
<td>-</td>
<td>104 Inadequate [+]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>English</td>
<td>Post Mat</td>
<td>Brockington, 2006</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>German</td>
<td>Post Mat</td>
<td>Reck, 2006</td>
<td>862</td>
<td>Adequate [-]</td>
<td>862</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>862 Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>English</td>
<td>Post Mat</td>
<td>Wittkowski, 2007</td>
<td>-</td>
<td>96</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>96 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>English</td>
<td>Post Mat</td>
<td>Wittkowski, 2010</td>
<td>132</td>
<td>Adequate [-]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>Chinese</td>
<td>Post Mat</td>
<td>Siu, 2010</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>Dutch</td>
<td>Post Mat</td>
<td>Van Bussel, 2010a</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>263 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>Japanese</td>
<td>Post Mat</td>
<td>Kaneko, 2014</td>
<td>1786</td>
<td>Doubtful [?]</td>
<td>1786</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>Japanese</td>
<td>Post Mat</td>
<td>Sueatsugu, 2015</td>
<td>199</td>
<td>Adequate [?]</td>
<td>199</td>
<td>Very good [?]</td>
<td>199 Doubtful [+]</td>
<td>199 Very good [+]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>Spanish</td>
<td>Post Mat</td>
<td>Garcia-Esteve, 2016</td>
<td>840</td>
<td>Very good [-]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>Portuguese</td>
<td>Post Mat</td>
<td>Baldisserotto, 2018</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>Italian</td>
<td>Post Mat</td>
<td>Busonera, 2017b</td>
<td>123</td>
<td>Adequate [+]</td>
<td>123</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>123 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Japanese</td>
<td>Post Mat</td>
<td>Ohashi, 2016</td>
<td>364</td>
<td>Very good [-]</td>
<td>364</td>
<td>Very good [?]</td>
<td>364 Adequate [+]</td>
<td>-</td>
<td>Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>German</td>
<td>Post Mat</td>
<td>Reck, 2006</td>
<td>-</td>
<td>862</td>
<td>Inadequate [?]</td>
<td>-</td>
<td>-</td>
<td>862 Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>English</td>
<td>Post Mat</td>
<td>Wittkowski, 2010</td>
<td>132</td>
<td>Adequate [+]</td>
<td>132</td>
<td>Very good [+]</td>
<td>-</td>
<td>-</td>
<td>132 Very good [+]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>Japanese</td>
<td>Post Mat</td>
<td>Kaneko, 2014</td>
<td>1786</td>
<td>Doubtful [?]</td>
<td>1786</td>
<td>Very good [?]</td>
<td>-</td>
<td>-</td>
<td>1786 Very good [-]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Japanese</td>
<td>Post Mat</td>
<td>Sueatsugu, 2015</td>
<td>199</td>
<td>Adequate [?]</td>
<td>199</td>
<td>Very good [?]</td>
<td>199 Doubtful [+]</td>
<td>199 Very good [+]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>Tamil</td>
<td>Post Mat</td>
<td>Vengavadaradan, 2019</td>
<td>250</td>
<td>Adequate [?]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Very good [?]</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post Mat</td>
<td>MM</td>
<td>Language</td>
<td>Measure</td>
<td>Year</td>
<td>N</td>
<td>Type</td>
<td>S-V</td>
<td>V-V</td>
<td>V-G</td>
<td>V-G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>----</td>
<td>----------</td>
<td>---------</td>
<td>------</td>
<td>---</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>Mat</td>
<td>S-PBQ</td>
<td>English</td>
<td>Kinsey, 2014</td>
<td>3005</td>
<td>Adequate (?)</td>
<td>3005</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>Mat</td>
<td>MIBS-8</td>
<td>English</td>
<td>Taylor, 2005</td>
<td>162</td>
<td>Adequate (?)</td>
<td>162</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>68</td>
<td>Mat</td>
<td>English</td>
<td>Wintkowski, 2007</td>
<td>–</td>
<td>–</td>
<td>96</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>96</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>Mat</td>
<td>Dutch</td>
<td>Van Bussel, 2010a</td>
<td>–</td>
<td>–</td>
<td>263</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>263</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>Mat</td>
<td>French</td>
<td>Bienfait, 2011</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>78</td>
<td>Very good [+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>Mat</td>
<td>MIBS-J-10</td>
<td>Japanese</td>
<td>Yoshida, 2012</td>
<td>554</td>
<td>Very good [-]</td>
<td>554</td>
<td>Very good (?)</td>
<td>554</td>
<td>Doubtful [-]</td>
<td>–</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>Mat</td>
<td>MIBS-J-7</td>
<td>Japanese</td>
<td>Ohara, 2016</td>
<td>375</td>
<td>Very good [+</td>
<td>751</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>Mat</td>
<td>MIRFS</td>
<td>Swedish</td>
<td>Thorstensson, Nissen 2012b</td>
<td>395</td>
<td>Adequate (?)</td>
<td>395</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>Mat</td>
<td>Swedish</td>
<td>Thorstensson, Hertfelt Wahn, 2012a</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>75</td>
<td>Mat</td>
<td>MORS-SF</td>
<td>English</td>
<td>Oates, 2018</td>
<td>–</td>
<td>–</td>
<td>100</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>Mat</td>
<td>Hungarian</td>
<td>Oates, 2018</td>
<td>–</td>
<td>–</td>
<td>331</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>Post Mat</td>
<td>English</td>
<td>Oates, 2019</td>
<td>100</td>
<td>Inadequate [-]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Post Mat</td>
<td>Hungarian</td>
<td>Oates, 2019</td>
<td>134</td>
<td>Inadequate [-]</td>
<td>–</td>
<td>–</td>
<td>36</td>
<td>Doubtful [+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>Post Mat</td>
<td>English and Hungarian</td>
<td>Oates, 2019</td>
<td>243</td>
<td>Adequate [-]</td>
<td>243</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>80</td>
<td>Post Pat</td>
<td>PPAS</td>
<td>English</td>
<td>Condon, 2008</td>
<td>461</td>
<td>Adequate [-]</td>
<td>461</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Post Pat</td>
<td>English</td>
<td>Feldstein, 2004</td>
<td>–</td>
<td>–</td>
<td>38</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>82</td>
<td>Post Pat</td>
<td>English</td>
<td>Condon, 2013</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>204</td>
<td>Inadequate [-]</td>
<td>204</td>
<td>Very good [-]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>83</td>
<td>Post Pat</td>
<td>Spanish</td>
<td>Riera-Martin, 2016</td>
<td>376</td>
<td>Very good [-]</td>
<td>376</td>
<td>Very good (?)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Very good (?)</td>
<td>376</td>
<td>Adequate [+</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**Notes.** [+ = sufficient. [- = insufficient. [?] = indeterminate. Structural validity ratings were based on the best fitting model presented in the paper (this was not necessarily the factor structure proposed by the original authors). As per the COSMIN criteria, internal consistency could only be rated as sufficient if there was at least low evidence of sufficient structural validity (otherwise an indeterminate rating was assigned). Some papers are listed more than once as the authors tested psychometric properties of more than one measure. 1 EFA was conducted on 29-item PAI, resulting in a 21-item scale. 2 EFA was conducted on a 31-item MAI, resulting in a 26-item scale. 3 PBQ-J-16 contains different items from the PBQ-16. 4 EFA conducted on full PBQ, resulting in 19-item scale. 5 MORS-SF EFA conducted on 44-items, resulting in 14-item scale.