A systematic review evaluating psychometric properties of parent or caregiver report instruments on child maltreatment: part 1: content validity

Yoon, S.; Speyer, R.; Cordier, R.; Aunio, P.; Hakkarainen, A.

A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 1: Content Validity

Sangwon Yoon[1] , Renée Speyer[1,2,3,4], Reinie Cordier[1,2] ,
Pirjo Aunio[1,5], and Airi Hakkarainen[6]

Abstract

Aims: Child maltreatment (CM) is a serious public health issue, affecting over half of all children globally. Although most CM is perpetrated by parents or caregivers and their reports of CM is more accurate than professionals or children, parent or caregiver report instruments measuring CM have never been systematically evaluated for their content validity, the most important psychometric property. This systematic review aimed to evaluate the content validity of all current parent or caregiver report CM instruments. **Methods:** A systematic literature search was performed in CINAHL, Embase, ERIC, PsycINFO, PubMed, and Sociological Abstracts; gray literature was retrieved through reference checking. Eligible studies needed to report on content validity of instruments measuring CM perpetrated and reported by parents or caregivers. The quality of studies and content validity of the instruments were evaluated using the COnsensus-based Standards for the selection of health Measurement INstruments guidelines. **Results:** Fifteen studies reported on the content validity of 15 identified instruments. The study quality was generally poor. The content validity of the instruments was overall sufficient, but most instruments did not provide high-quality evidence for content validity. **Conclusions:** Most instruments included in this review showed promising content validity. The International Society for the Prevention of Child Abuse and Neglect Child Abuse Screening Tool for use in Trial appears to be the most promising, followed by the Family Maltreatment–Child Abuse criteria. However, firm conclusions cannot be drawn due to the low quality of evidence for content validity. Further studies are required to evaluate the remaining psychometric properties for recommending parent or caregiver report CM instruments.

Keywords

assessment, child abuse, COSMIN, measure, measurement properties, parent report

Child maltreatment (CM) is defined by the World Organization (WHO, 2016) as:

> the abuse and neglect of children under 18 years of age. It includes all forms of physical and/or emotional ill treatment, sexual abuse, neglect, negligence, and commercial or other exploitation, which results in actual or potential harm to the child's health, survival, development, or dignity in the context of a relationship of responsibility, trust, or power. (p. 94)

This broad definition can be distinguished into four subtypes of CM (Krug et al., 2002; WHO, 1999): (1) physical abuse (PA: acts causing actual or potential physical harm); (2) emotional abuse (EA: acts having adverse impact on a child's emotional development); (3) sexual abuse (SA: acts using a child for sexual gratification); and (4) neglect (failure in providing for the development of a child in health, education, emotional development, nutrition, shelter, and safe living conditions).

CM causes significant public health problems and socioeconomic burden. CM can cause physical injuries, psychosocial difficulties, and lower academic achievement during childhood

[1] Department of Special Needs Education, Faculty of Education, University of Oslo, Norway
[2] School of Occupational Therapy, Social Work and Speech Pathology, Faculty of Health Sciences, Curtin University, Perth, Australia
[3] Department of Otorhinolaryngology and Head and Neck Surgery, Leiden University Medical Centre, the Netherlands
[4] Faculty of Health, School of Health and Social Development, Deakin University, Victoria, Australia
[5] Department of Education, University of Helsinki, Finland
[6] Open University, University of Helsinki, Finland

Corresponding Author:
Sangwon Yoon, Department of Special Needs Education, Helga Engs Hus, University of Oslo, Sem Sælands vei 7, 0371 Oslo, Norway.
Email: sangwon.yoon@isp.uio.no

(Boden et al., 2007; Glaser, 2000; Teicher et al., 2016; van Harmelen et al., 2010). Moreover, adults with histories of childhood abuse tend to have higher risk of mortality, lower educational attainment, and lower income compared with adults without a history of CM (Anda et al., 2010; Currie & Spatz Widom, 2010; Danese & McEwen, 2012; Felitti et al., 1998).

The prevalence of CM in the general population has been estimated at 57.6% of all children in the world (Hillis et al., 2016), and most CM is perpetrated by parents or caregivers (Devries et al., 2018; Sedlak et al., 2010). A recent meta-analysis on global prevalence of CM suggests that the overall prevalence rates are 12.7% for SA, 22.6% for PA, 36.3% for EA, and 34.7% for neglect (Stoltenborgh et al., 2015). While the most common perpetrators of SA are nonfamily members (Finkelhor et al., 2014), at least 50% of PA and EA or neglect is perpetrated by caregivers (Devries et al., 2018). For example, in the United States of America, parents are the perpetrators of 72% of all physically abused children, 73% of emotionally abused children, and 92% of neglected children, compared with 37% of sexually abused children (Sedlak et al., 2010). Thus, CM perpetrated by parents or caregivers is an important construct of interest.

However, estimates of the prevalence of CM vary markedly depending on who the informants are. Meta-analyses have shown that self-reported or caregiver-reported prevalence of CM is greater than prevalence reported by professionals such as doctors or child protection workers (Stoltenborgh et al., 2015). Furthermore, the prevalence rate of most forms of CM reported by children is far lower when compared with caregiver reports, with SA the notable exception (Devries et al., 2018). In contrast to self-report and caregiver report, lower professional–reported prevalence rates may be the result of professionals more likely to report severe CM cases, as mild cases may be considered as not important enough to report (Negriff et al., 2017). Conversely, young children may have more trouble recalling abusive and neglecting behaviors than adult caregivers (Devries et al., 2018). While caregiver-reported prevalence on CM appears to be less affected by underestimation of CM (Devries et al., 2018; Stoltenborgh et al., 2015), accuracy and reliability of a caregiver report instrument on CM are still an ongoing debate due to caregivers' general tendency to respond in socially desirable ways (Compier-de Block et al., 2017). Therefore, identifying reliable and valid parent or caregiver report measures is essential to estimate accurate prevalence of CM.

While directly measuring the prevalence of parental CM is important, there is a need to measure parents' attitude toward CM for the purpose of CM prevention, that is, parental values, beliefs, or feelings in relation to abusive and neglecting behavior toward a child (Altmann, 2008). Since parents are the main perpetrators of CM (Devries et al., 2018; Sedlak et al., 2010), prevention efforts need to focus on parents. Parents' attitude toward CM is a critical predictive factor of parental child abuse behavior (Stith et al., 2009). Several studies have shown that parents with more positive beliefs or values toward CM tend to show more child abusive behaviors than parents with a negative attitude (Asadollahi et al., 2016; Ateah & Durrant, 2005; Bower-Russa, 2005; Chavis et al., 2013; Stith et al., 2009; Vittrup et al., 2006). For this reason, a number of studies on CM prevention used instruments to measure parents' attitude toward CM as an outcome measure to establish whether the programs being evaluated are effective (Chen & Chan, 2016; Gershoff et al., 2017; Holden et al., 2014; Voisine & Baker, 2012). Therefore, to measure the outcomes for evidence-based CM prevention programs, reliable and valid instruments to measure parents' attitude toward CM are needed, as well as suitable instruments to measure parents' actual maltreating behaviors toward their children.

Even though the selection of a high-quality instrument is critically important for accurate and reliable assessment of CM, there is no universally accepted gold standard for measuring CM (Bailhache et al., 2013). The best way for selecting suitable evidence-based instruments is by evaluating the instruments' psychometric properties through a systematic review (Scholtes et al., 2011). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) group has developed and published comprehensive guidelines for conducting systematic reviews on psychometric properties of patient-reported outcome instruments (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). The COSMIN methodological guidelines include a taxonomy defining each psychometric property (Mokkink et al., 2010b), a checklist to assess the methodological quality of psychometric studies (Mokkink et al., 2018), criteria to evaluate the psychometric quality of instruments (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018), and a rating system to summarize psychometric evidence and grade quality of all evidence used for the psychometric quality assessment of instruments (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018).

The COSMIN taxonomy distinguishes nine psychometric properties across three domains: (1) validity (i.e., the extent to which an instrument measures the construct it is intended to measure); (2) reliability (i.e., the extent to which scores for patients who have not changed are the same for repeated measurements); and (3) responsiveness (i.e., the ability to detect clinically important change over time in the construct measured; Mokkink et al., 2010b). The domain of validity contains five psychometric properties: content validity (i.e., the extent to which the content of an instrument adequately reflects the construct to be measured), structural validity (i.e., the extent to which the scores adequately reflect the dimensionality of the construct to be measured), cross-cultural validity (i.e., the extent to which a translated or culturally adapted version of an instrument adequately reflects the performance of the items of the original instrument), hypothesis testing for construct validity (i.e., the extent to which the scores are consistent with hypotheses on differences between relevant groups and relations to scores of other instruments), and criterion validity (i.e., the extent to which the scores adequately reflect a "gold standard"; Mokkink et al., 2010b). Next, the reliability domain

contains three psychometric properties: internal consistency (i.e., the degree of the interrelatedness of items), reliability (i.e., the proportion of total score variance which is due to true differences among respondents), and measurement error (i.e., the systematic and random error of a respondent's score that is not because of true changes in the construct measured; Mokkink et al., 2010b). Lastly, the domain of responsiveness includes only one psychometric property that is also called responsiveness, which has the same definition as the domain (Mokkink et al., 2010b).

When selecting an instrument, the most important psychometric property is its content validity (Prinsen et al., 2018; Prinsen et al., 2016); if it is unclear what construct(s) the instrument is actually measuring, then the evidence of the remaining psychometric properties is not valuable (Patrick et al., 2011; Streiner et al., 2015). For example, a high Cronbach's α does not guarantee that all important concepts are included. Similarly, a high test–retest reliability or adequate responsiveness does not imply that all items are relevant to the construct being measured (Cortina, 1993; Sijtsma, 2009).

Content validity pertains to three aspects of the content of an instrument: (1) relevance (i.e., the degree to which all items of an instrument are relevant for the construct of interest within a target population and purpose of use), (2) comprehensiveness (i.e., the degree to which all key concepts of the construct are included in an instrument), and (3) comprehensibility (i.e., the degree to which items of an instrument are easy to understand by respondents; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018). Weaknesses in any of these three aspects of content validity can impact on all other psychometric properties (Wiering et al., 2017) in the following ways: If items of an instrument are irrelevant (poor relevance), it may decrease interrelatedness among the items (internal consistency), structural validity, and interpretability of an instrument, and if an instrument misses some key concepts of the construct (poor comprehensiveness), it may reduce the ability of an instrument to detect real change in the construct of interest before and after intervention (poor responsiveness; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Since content validity can have a significant influence on all other psychometric properties, the COSMIN methodological guidelines recommend evaluating the content validity of an instrument first and to not evaluate other psychometric properties if reviewers have high-quality evidence that the instrument has insufficient content validity (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018).

To have good content validity, instrument items and instructions should be sufficiently relevant, comprehensive, and comprehensible, based on high-quality evidence (Chiarotto, 2019). According to the COSMIN criteria, for a measure to be rated as having good content validity, the measure should have (1) items relevant to the construct of interest in a specific population and purpose of use and appropriate response options and a recall period (relevance), (2) comprehensive items covering all key concepts (comprehensiveness), and (3) instructions, items, and response options that are understandable to the target population (comprehensibility; Terwee, Prinsen, Chiarotto,

Westerman, et al., 2018). Evidence for rating these three aspects of content validity is mainly derived from instrument development and content validity studies (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018). The development study refers to a study generating relevant items based on input from the target population for a new instrument (item generation) and evaluating comprehensiveness and comprehensibility of a draft instrument by interview or survey with the target population (cognitive interview or pilot test). The content validity study refers to a study asking target population and professionals about relevance, comprehensiveness, and comprehensibility of an existing instrument. As additional evidence, the original instrument (i.e., content of instrument itself) should also be rated based on subjective opinion of reviewers in terms of relevance, comprehensiveness, and comprehensibility (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Summarizing all evidence from the studies and content of instrument itself, overall relevance, comprehensiveness, and comprehensibility of an instrument need to be determined (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Furthermore, the level of quality of all evidence used to determine overall relevance, comprehensiveness, and comprehensibility should be summarized (graded) to show how confident we are in the overall ratings on the three aspects of content validity, respectively. When the overall relevance, comprehensiveness, and comprehensibility are all sufficient and the levels of quality of evidence for the overall ratings are all high, we can decisively conclude that the instruments have good content validity (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018).

Only one study to date has conducted a systematic review on content validity of CM instruments (Saini et al., 2019). However, the review identified only child self-report and clinician interview instruments, which tend to underestimate the actual incidence of CM compared to parent report instruments (Devries et al., 2018) and one parent proxy-report instrument (asking parents about their children's maltreated experience by any adults, not about their own perpetration of CM; Saini et al., 2019; Sprangers & Aaronson, 1992). None of the instruments and studies included in the review by Saini et al. (2019) overlapped with this current review for parent- or caregiver-reported CM instruments. Furthermore, the authors did not use the latest, thoroughly revised COSMIN methodological guidelines (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018) but instead used the old version of the COSMIN checklist (Mokkink et al., 2010a) and criteria (Terwee et al., 2007) for assessing the methodological quality of studies on content validity and the quality of content validity of instruments. The old version of COSMIN checklist consists of a simplified 5-item for assessing only content validity studies and does not contain any standards for assessing the methodological quality of instrument development studies. Moreover, the early COSMIN criteria do not have specific consensus-based criteria for rating the relevance, comprehensiveness, and comprehensibility of an instrument (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). To address these shortcomings, the COSMIN methodological guideline for
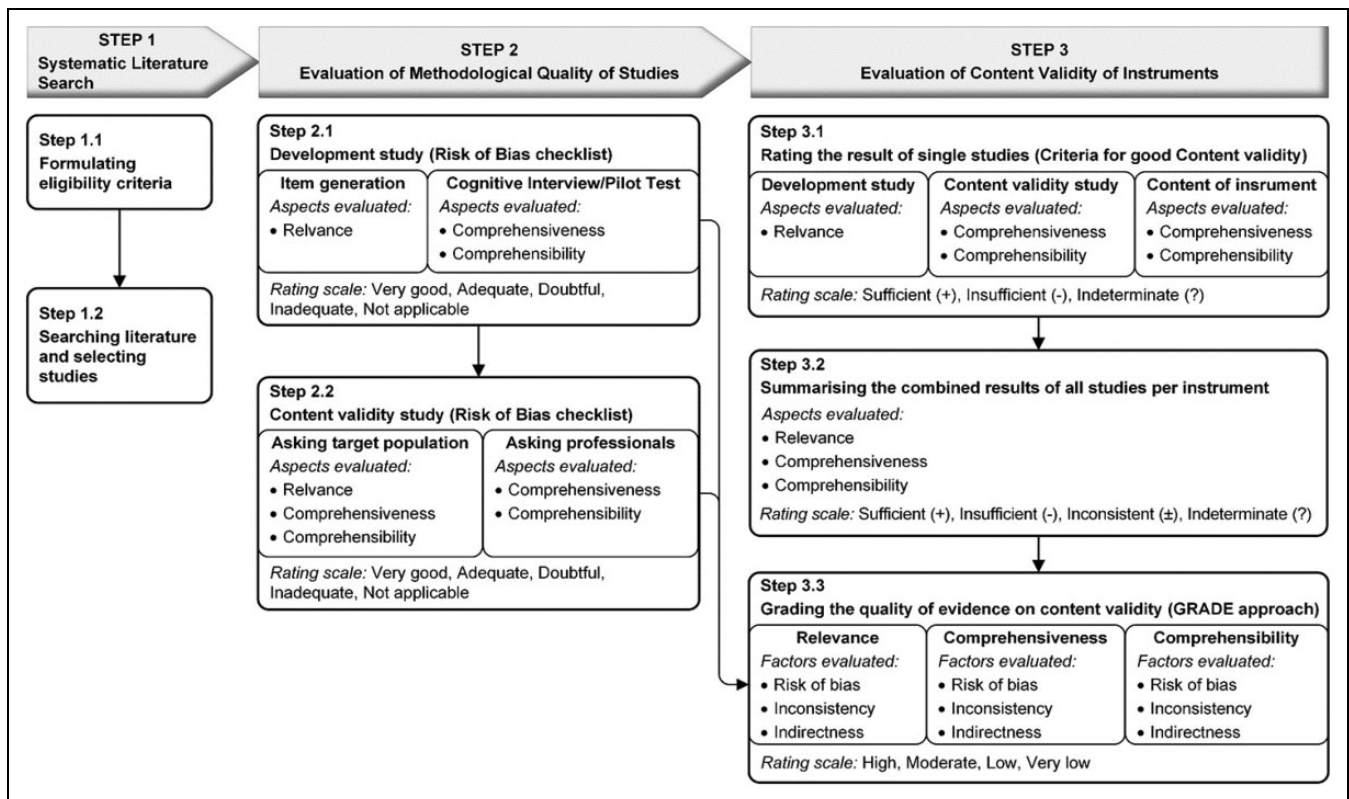
Figure 1. Study design: Steps for Preferred Reporting Items for Systematic Reviews and Meta-Analyses and COnsensus-based Standards for the selection of health Measurement INstruments processes.

assessing content validity of an instrument has been recently developed to provide a detailed and standardized checklist and criteria (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). No other systematic reviews on content validity or any of the other psychometric properties of parent or caregiver report instruments on CM have been published.

Study Aim

The aim of this systematic review was to evaluate content validity of all current parent or caregiver report CM instruments using the updated COSMIN methodological guidelines (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Due to the size, scope, and complexity of reporting the remaining psychometric properties, we aim to report the quality of studies and psychometrics of instruments identified in this systematic review in a companion paper (Part 2), excluding those instruments found to have high-quality evidence for insufficient content validity in this article.

Method

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Moher et al., 2009) and the COSMIN methodological guidelines (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al.,

2018). This review consists of three consecutive steps (see Figure 1):

- Step 1: *Systematic literature search* formulating eligibility criteria (Step 1.1) and searching literatures and selecting studies (Step 1.2; Moher et al., 2009);
- Step 2: *Evaluation of the methodological quality of studies* on instrument development (Step 2.1) and content validity (Step 2.2) using the COSMIN Risk of Bias checklist (Mokkink et al., 2018); and
- Step 3: *Evaluation of the content validity of instruments* rating the result of single studies against the criteria for good content validity (Step 3.1), summarizing all results of studies per instrument (Step 3.2), and grading quality of evidence on content validity (Step 3.3; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018).

Each of these steps will be explained in more detail in the following sections.

Systematic Literature Search (Step 1)

The systematic literature search was conducted for both this article on content validity (Part 1) and a companion paper on other psychometric properties (Part 2) by formulating eligibility criteria (Step 1.1) and searching literature and selecting studies (Step 1.2).

Eligibility criteria (Step 1.1). To select instruments and studies for this current review, the following five eligibility criteria for inclusion were used: (1) parent or caregiver report instruments assessed their own attitudes toward CM or maltreating behaviors toward their children; (2) at least one subscale or a minimum of 30% of all items within an instrument referred to one or more types of CM (i.e., PA, EA, SA, and neglect; Krug et al., 2002; WHO, 1999), as a criterion to ensure the contribution to the overarching construct of an instrument was involved CM; (3) instruments were developed and studies were published in English; (4) studies reported on psychometric data of at least one of the nine psychometric properties of eligible instruments as defined in the COSMIN taxonomy (Mokkink et al., 2010b) that were published as original journal articles, manuals, book chapters or conference papers; and (5) studies on content validity reported on the development of new items of eligible instruments, and/or evaluated the relevance, comprehensiveness, or comprehensibility of the content of the eligible instruments as reported by parents or caregivers and/or professionals.

Literature search and study selection (Step 1.2). To identify eligible instruments and journal articles that reported on any psychometric properties of the instruments as defined in the COSMIN taxonomy (Mokkink et al., 2010b), systematic literature searches were performed in six electronic databases (CINAHL, Embase, ERIC, PsycINFO, PubMed, and Sociological Abstracts) on January 29, 2018, with an update on October 5, 2019. Search terms consisted of subject headings and free-text words (see Online Appendix A). All publications prior to October 2019 were considered for inclusion.

Abstracts and articles retrieved from database searches were screened to identify eligible instruments and journal articles on any psychometric property by two reviewers independently. One reviewer screened all abstracts, while the other reviewer screened a random selection of approximately half of all abstracts; all full texts of eligible abstracts were retrieved and screened by both independent reviewers. Any discrepancies between both reviewers were resolved by involving a third reviewer. The degree of agreement between the two reviewers was assessed using Cohen's weighted κ (Cohen & Humphreys, 1968); agreement was very good (Altman, 1991): (1) weighted κ for abstract selection = .87 (95% confidence interval [CI] = [.83, .90]) and (2) weighted κ for article selection = .86 (95% CI [.77, .94]).Reference lists of all included full-text articles on any psychometric property were hand searched to identify additional eligible instruments and psychometric studies on the instruments. Websites of Pearson and Western Psychological Services, two major measurement publishers in social science, were also searched to retrieve potential instruments and manuals not identified in previous databases and reference searches. Both of the reference lists and websites were searched by one reviewer, and the additionally retrieved instruments and psychometric studies were checked by another reviewer. If instruments were not published or freely available, the developers of the instruments were contacted by e-mail to retrieve the original instruments.

Finally, among all eligible psychometric studies, only studies on content validity (i.e., instrument development and content validity studies) were included in this review (Part 1) for the evaluation of content validity. Studies on other psychometric properties were excluded in this article (Part 1), as these findings will be reported on in a companion paper (Part 2).

Evaluation of Methodological Quality of Studies (Step 2)

The methodological quality of included studies on instrument development (Step 2.1) and content validity (Step 2.2) was assessed using the COSMIN Risk of Bias checklist (Mokkink et al., 2018). First, the development studies were assessed using 35 items from the checklist, which consists of a separate rating of the quality of the "instrument design" (item generation) to ensure relevance of a new instrument and "cognitive interview or pilot test" to evaluate comprehensiveness and comprehensibility of a draft instrument (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018). Next, content validity studies were assessed using 38 items from the checklist, comprised of one set of items assessing quality of studies that ask parents or caregivers about relevance, comprehensiveness, and comprehensibility, and another set assessing quality of studies that ask professionals about relevance and comprehensiveness (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018). Total ratings for each aspect of content validity (i.e., relevance, comprehensiveness, and comprehensibility) were determined separately. Separate total ratings were also determined for the two parts of the development study (instrument design and cognitive interview or pilot test) as well as for two types of content validity study ("asking parents or caregivers" and "asking professionals"; Mokkink et al., 2018).

When rating the methodological quality of the instrument development and content validity studies, each checklist item was ranked on a 4-point rating scale (1 = *inadequate*, 2 = *doubtful*, 3 = *adequate*, and 4 = *very good*). A total rating for relevance, comprehensiveness, or comprehensibility was obtained by calculating the percentage of the ratings based on the following formula (Cordier et al., 2015), instead of a worst score counts method (reporting total ratings gained by taking the lowest rating among any of the checklist items) recommend by the COSMIN methodological guidelines (Mokkink et al., 2018). This approach was adopted as determining total scores of methodological quality of studies that are entirely based on the lowest rating of single items impedes the detection of subtle differences in methodological quality between studies (Speyer et al., 2014).

Total score for methodological quality (%)

$$= \frac{(\text{total score obtained} - \text{min score possible})}{(\text{max score possible} - \text{min score possible})} \times 100.$$

The total percentage score is then categorized into the following four scores: inadequate (from 0% to 25%), doubtful (from 25.1% to 50%), adequate (from 50.1% to 75%), and very good (from 75.1% to 100%). Two reviewers rated the methodological quality independently where after consensus ratings

were determined between the two reviewers. The interrater reliability was calculated using weighted κ (Cohen & Humphreys, 1968) between both reviewers.

After assessment of methodological quality on the included instrument development and content validity studies, the following data were extracted from the included studies and instruments: (1) study characteristics (i.e., study purpose, study population, and parents or professionals involvement); (2) instrument characteristics (i.e., instrument names and acronyms, measured constructs, targeted population, purpose of use, number of [sub] scales, number of items, response options and recall period); and (3) study results on all three aspects of content validity (relevance, comprehensiveness, and comprehensibility). All relevant data were extracted by one reviewer and rechecked for accuracy by another reviewer.

Evaluation of Content Validity of Instruments (Step 3)

The content validity of instruments was assessed for three separate aspects of content validity (relevance, comprehensiveness, and comprehensibility) in three sequential steps: Step 3.1, Step 3.2, and Step 3.3. All ratings were conducted by two reviewers independently, and any discrepancies were resolved by consensus.

Rating the result of single studies (Step 3.1). Rating the results of single studies was conducted for each instrument development study, content validity study, and content of the instrument itself separately. The results of each development and content validity study were rated based on the qualitative or quantitative data obtained by asking parents or caregivers and/or professionals about content validity of an instrument, using the 10 predefined criteria on relevance (5), comprehensiveness (1), and comprehensibility (4; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). By using the same criteria, the content of the original instrument itself (items, response options, and recall period) was also rated based on the subjective judgment of the reviewers. The reviewers received extensive training in appraising content validity of instruments using the COSMIN criteria under supervision of the second author who has considerable expertise in psychometrics and the COSMIN framework. Ratings for each source of evidence on content validity were given as sufficient (85% or more of the instrument items meet the criterion: +), insufficient (less than 85% of the instrument items meet the criterion: −), or indeterminate (lack of evidence to determine the quality or inadequate methodological quality of studies?). More detailed information on these criteria and how to apply these criteria can be found in the user manual on COSMIN methodology for assessing content validity (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018).

Summarizing the results of all studies per instrument (Step 3.2). All results from available studies on development and content validity per instrument and the reviewers' ratings on content of the instrument were qualitatively summarized into overall ratings for relevance, comprehensiveness, and comprehensibility of the instrument (i.e., all ratings determined in the previous step were jointly assessed; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). The focus in this step was on the specific instrument, while in the previous step, the focus was on single studies. An overall sufficient (+), insufficient (−), inconsistent (±), or indeterminate (?) rating was given for relevance, comprehensiveness, and comprehensibility for each instrument (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). For example, if all relevance scores of development studies, content validity studies, and content of the instrument (reviewers' ratings) were sufficient, insufficient, or indeterminate, the overall relevance rating became sufficient (+), insufficient (−), or indeterminate (?). If, however, at least one of these three scores was inconsistent with the other two scores, the overall rating became inconsistent (±). An exception to this rule was when the scores of both development and content validity studies were all indeterminate and inconsistent with the reviewers' rating on content of the instrument. In this instance, the overall rating could be determined by solely the reviewers' rating. Further details on rating overall relevance, comprehensiveness, and comprehensibility can be founded in the user manual for assessing content validity (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018).

Grading the quality of evidence on content validity (Step 3.3). The quality of the evidence (i.e., the total body of evidence used for overall ratings on relevance, comprehensiveness and comprehensibility of an instrument) was graded (high, moderate, low, or very low) using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Guyatt et al., 2008; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). The GRADE approach is used to downgrade level of evidence when there are concerns about the quality of evidence. The starting point of the evidence quality rating is based on the assumption that the overall rating is of high quality. Next, ratings are downgraded one or more levels (to moderate, low, or very low) if there is serious or very serious risk of bias (i.e., limitations in the methodological quality of studies), inconsistency (i.e., unexplained heterogeneity in results of studies), and/or indirectness (i.e., evidence from different populations than the target population of interest in the review; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). The quality of evidence was not graded if the overall rating was indeterminate (?) due to lack of evidence. More specific information about grading the quality of evidence can be found in the COSMIN user manual for content validity (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018).

Results

Systematic Literature Searches

In total, 2,859 nonduplicate abstracts were identified from six databases: CINAHL (1,173 records), Embase (456 records), ERIC (523 records), PsycINFO (285 records), PubMed
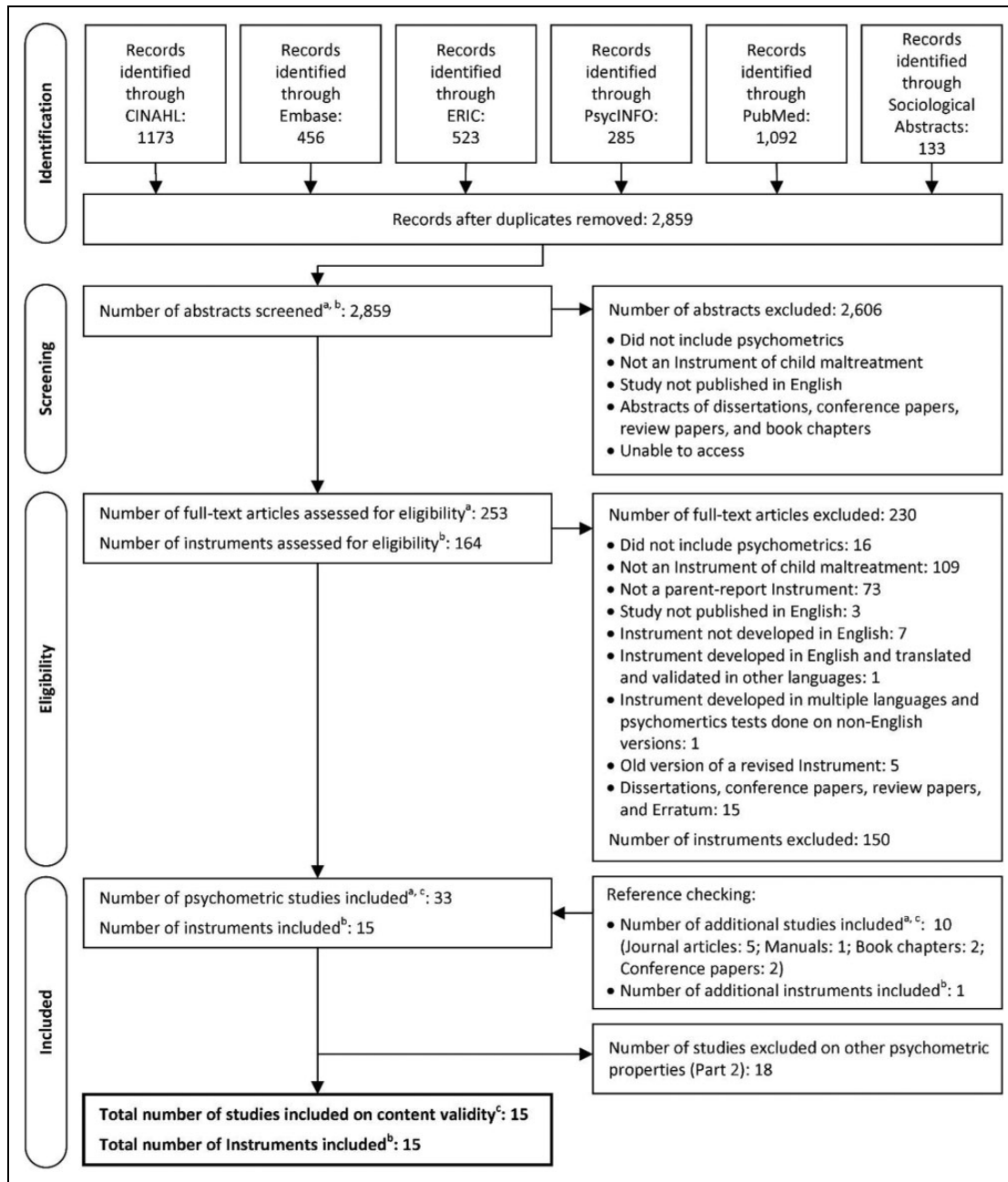
Figure 2. Flow diagram of the reviewing procedure based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Moher et al., 2009).
Notes. The literature searches and study selection were conducted for both this paper on content validity (Part 1) and a companion paper on other psychometric properties (Part 2).
[a]Studies on any psychometric property were eligible if they: (1) were journal articles and manuals published in English: (2) reported on psychometric data of any psychometric properties of eligible instruments.
[b]Instruments were eligible if: (1) attitude towards child maltreatment or maltreating behaviours towards children was assessed.

(1,092 records), and Sociological Abstracts (133 records). Figure 2 shows the flow diagram of the studies and instruments identified during the literature search and screening process in accordance with PRISMA (Moher et al., 2009). A total of 253 full-text articles and 164 instruments were assessed for eligibility, resulting in 23 full-text articles reporting on

psychometric properties and 14 instruments. Online Appendix B summarizes a list of the 150 excluded instruments and reasons for exclusion.

Reference checking of the 23 articles on psychometric properties resulted in one additional instrument and 10 additional psychometric studies being identified as meeting eligibility

criteria. A total of 33 psychometric studies evaluating 15 different instruments were identified. Fifteen of 33 psychometric studies reported on content validity (i.e., instrument development or content validity studies) and were included in this review (Part 1).

Characteristics of Included Studies and Instruments

Descriptions of the instrument development or content validity studies of the included CM instruments are presented in Online Appendix C. Table 1 provides a summary of the characteristics of all 15 instruments, including names and acronyms, construct of interest (subscales), target population, intended contexts for use, number of (sub)scales and items, response options, and recall periods. All 15 instruments measured at least one type of CM (construct of interest) for parents or caregivers (target population) with the purpose to identify maltreating parents, as well as abused children, and/or to evaluate intervention programs (purpose of use). Of the 15 instruments identified, no instrument measured only SA; 3 measured both SA and other types of CM (PA, EA, and/or neglect); and 12 measured other types of CM. The total number of subscales ranged from no subscales to six subscales; the total number of items varied between 4 and 60. All but one instrument used a Likert-type response scale, while only one used a reaction time response. Recall period varied between last week and last year for eight instruments (Child Neglect Questionnaire [CNQ], Child Neglect Scales–Maternal Monitoring and Supervision Scale [CNS-MMS], Conflict Tactics Scales: Parent–Child Version [CTSPC], Family Maltreatment–Child Abuse criteria [FM-CA], ISPCAN (International Society for the Prevention of Child Abuse and Neglect) Child Abuse Screening Tool for use in Trials [ICAST-Trial], Mother–Child Neglect Scale [MCNS], MCNS-Short Form [MCNS-SF], and Parental Response to Child Misbehavior questionnaire [PRCM]); the recall period was unspecified in the remaining seven instruments (Adult Adolescent Parenting Inventory-2 [AAPI-2], Analog Parenting Task [APT], Child Trauma Screen–Exposure Score [CTS-ES], Intensity of Parental Punishment Scale [IPPS], Parent–Child Aggression Acceptability Movie Task [P-CAAM], Parent Opinion Questionnaire [POQ], Shaken Baby Syndrome awareness assessment–Short Version [SBS-SV]).

Methodological Quality of Development and Content Validity Studies

The methodological quality of the 15 included studies on instrument development (14) and content validity (10) was assessed using the COSMIN checklist (Mokkink et al., 2018). All 10 content validity studies overlapped with the development studies; one study reported on more than one instrument. An overview of all methodological quality ratings is presented in Table 2. Only five development studies reported on either item generation or cognitive interviewing. Of those five studies, three studies used both item generation and cognitive interviews, whereas the other two studies conducted cognitive

interviews only. Of the 13 instrument development study quality ratings, a single rating for relevance and comprehensiveness was classified as doubtful, while all other 11 ratings were classified as inadequate. In content validity studies, all but five studies asked parents or carers and/or professionals about at least one of the three aspects on content validity (relevance, comprehensiveness, and comprehensibility). Of the 15 content validity study quality ratings, only 3 ratings (1 relevance and 2 comprehensibility) were rated as very good or adequate, whereas all other 12 ratings were rated as doubtful or inadequate. No information was retrieved on comprehensiveness in any content validity studies. The interrater reliability for study quality assessment between both reviewers was good (weighted κ .76; 95% CI [.68, .85]).

Content Validity of Instruments

Table 3 summarizes ratings on the content validity for development and content validity studies, respectively, as well as the content of instrument itself involving 15 studies and 15 instruments. The data of each single study and content of instruments were evaluated against the 10 criteria for good content validity for the following three separate aspects of content validity: relevance, comprehensiveness, and comprehensibility (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). All development and content validity studies received indeterminate ratings, except for the following two studies of FM-CA: one development study received sufficient rating in relevance and one content validity study received sufficient rating in comprehensibility. All but four instruments (CTS-ES, P-CAAM, POQ, and PRCM) were rated as sufficient for content of instruments based on the reviewers' expert opinion. Three instruments reported conflicting ratings in one of the three aspects of content validity (CTS-ES and POQ in relevance and PRCM in comprehensibility). Two instruments reported insufficient ratings in comprehensiveness (CTS-ES and POQ), and one instrument reported indeterminate ratings in all three aspects (P-CAAM).

Table 4 presents the overall ratings on content validity with quality of evidence for content validity. All but four instruments (CTS-ES, P-CAAM, POQ, and PRCM) received sufficient overall ratings in all three aspects of content validity (relevance, comprehensiveness, comprehensibility). Three instruments reported conflicting overall ratings in one of the three aspects of content validity (CTS-ES and POQ in relevance and PRCM in comprehensibility). Two instruments reported insufficient overall ratings in comprehensiveness (CTS-ES and POQ), and one instrument reported indeterminate overall ratings in all three aspects due to failure of retrieving the original instrument (P-CAAM).

High-quality evidence supporting overall ratings on content validity was only available for the FM-CA and the ICAST-Trial, whereas no high-quality evidence for content validity was found for the remaining 13 instruments. In fact, 67% (30/45) of all evidence quality ratings for content validity were rated as very low. For overall ratings of relevance, six

Table 1. Characteristics of the Included Instruments for the Assessment of Child Maltreatment.

Instrument (Acronym)	Main Constructs (Subscales)	Target Population (Child Age)	Purpose of Use	Number of Subscales (Total Number of Items); Range of Score	Response Options	Recall Period
Adult Adolescent Parenting Inventory-2 (AAPI-2; Bavolek & Keene, 1999; Bavolek et al., 1979)	Abusive and neglecting parenting practices (inappropriate parental expectations; parental lack of an empathic awareness of children's needs; strong belief in the use and value of corporal punishment; parent child role reversal; oppressing children's power and independence)	Current and prospective parent populations (NR)	To provide prevalence estimates of child maltreatment; to screen child maltreatment; to evaluate prevention and treatment of physical and psychological child abuse	5 (40); range: 0–50 (raw total scores per subscale are converted into standard scores: range 0–10)	5-point ordinal scale (1 = strongly disagree to 5 = strongly disagree)	Not specified
Analog Parenting Task (APT; Russa & Rodriguez, 2010; Zaidi et al., 1989)	Attitude toward physical discipline (physical discipline score: frequency of physical disciplinary response to alter children's behavior; escalation score: frequency of switching from nonphysical to physical disciplinary tactics when child persisting in behavior)	Prospective parent populations (NR)	To identify high-risk pre-parent populations for primary prevention programming	2 (26); range: 0–26	10 nominal scale (from nonphysical discipline tactics to physical discipline tactics)	Not specified
Child Neglect Questionnaire (CNQ; Stewart et al., 2015)	Child neglect (physical neglect; emotional neglect; educational neglect; supervision neglect)	Parents with older children (ages 10–12)	To detect children at high risk for parental neglect	4 (46); range: 46–184	4-point ordinal scale (1 = always to 4 = never)	Past 6 months
Child Neglect Scales–Maternal Monitoring and Supervision Scale (CNS-MMS; Kirisci et al., 2001; Loeber et al., 1998)	Child neglect by parents	Mothers (NR)	To quantify severity of child neglect by mothers	1 (11); range: 11–33	3-point ordinal scale (1 = hardly ever to 3 = often)	Past 6 months
Child Trauma Screen–Exposure Score (CTS-ES; Lang & Connell, 2017)	Potentially traumatic event including childhood physical abuse, sexual abuse, and domestic or community violence	Caregivers with children (ages over 6)	To screen children for trauma exposure	1 (4); range: 0–4	Dichotomous scale (no = 0 or yes = 1)	Not specified

Table I. (continued)

Instrument (Acronym)	Main Constructs (Subscales)	Target Population (Child Age)	Purpose of Use	Number of Subscales (Total Number of Items); Range of Score	Response Options	Recall Period
Conflict Tactics Scales: Parent–Child Version (CTSPC; Straus et al., 1998, 2003)	Physical and psychological child abuse (nonviolent discipline; psychological aggression; physical assault) (Optional supplementary three subscales: weekly discipline; neglect; sexual abuse)	Parents (NR)	To provide prevalence estimates of child maltreatment; to screen child maltreatment; to evaluate prevention and treatment of physical and psychological child abuse	3 (22); range: 0–550 (raw scores per item are converted into frequency scores: $0 = 0$, $1 = 1$, $2 = 2$, $3–5 = 4$, $6–10 = 8$, $11–20 = 15$, and $>20 = 25$) (Supplementary subscales: 3 (13); 0–233)	8-point ordinal scale ($0 = never\ happened$; $1 = once\ in\ the\ past\ year$; $2 = twice$; $3 = 3–5\ times$; $4 = 6–10\ times$; $5 = 11–20\ times$; $6 = more\ than\ 20\ times$; $7 = not\ in\ the\ past\ year,\ but\ it\ happened\ before$) (Supplementary subscales: 3 to 7-point ordinal scale)	Past 1 year (Optional supplementary subscales: past 1 week to lifetime before 18 years old)
Family Maltreatment–Child Abuse criteria (FM-CA; Heyman et al., 2019)	Clinically significant child abuse and neglect (physical child abuse; psychological child abuse)	Parents (NR)	To screen clinically significant child abuse	2 (27); range: 0–63	Dichotomous scale for physical child abuse subscale ($0 = I\ did\ or$ $1 = I\ never\ did$); 6-point ordinal scale for psychological child abuse subscale ($0 = never\ to$ $5 = more\ than\ once\ a\ day$)	Past 1 year
International Society for the Prevention of Child Abuse and Neglect Child Abuse and Neglect Child Abuse Screening Tool for use in Trials (ICAST-Trial; Meinck et al., 2018; Runyan et al., 2009)	Child abuse and neglect (physical abuse; emotional abuse; contact sexual abuse; neglect)	Caregivers (ages 10–18)	To evaluate effectiveness of child abuse prevention program	4 (14); range: 0–112	9-point ordinal scale ($0 = never\ to\ 8 = more$ $than\ 8\ times$)	Past 1 month
Intensity of Parental Punishment Scale (IPPS; Gordon et al., 1979)	Intensity of parent behavioral responses to hypothetical child misbehavior situations (school misbehavior; disobedience after a recent reminder; public disobedience; crying; destructiveness)	Parents of children (ages 5–10)	To provide investigators with cost-effective information of long-term effects on parental punishments than time-consuming interview and observation without any demonstrable reduction in accuracy	5 (33); range: 33–231	7-point ordinal scale ($1 = no$ $reaction\ to\ 7 = very\ strong$ $punishment$)	Not specified
Mother–Child Neglect Scale (MCNS; Lounds et al., 2004; Straus et al., 1995)	Maternal neglectful behavior toward their children (emotional neglect; cognitive neglect; supervisory neglect; physical needs neglect)	Mothers (NR)	To screen parents at highest risk of child neglect for prevention of its future occurrence	4 (20); range: 20–80	4-point ordinal scale ($1 = strongly\ disagree$ to $4 = strongly\ agree$)	Past 1 year

Table 1. (continued)

Instrument (Acronym)	Main Constructs (Subscales)	Target Population (Child Age)	Purpose of Use	Number of Subscales (Total Number of Items); Range of Score	Response Options	Recall Period
Mother–Child Neglect Scale–Short Form (MCNS-SF; Lounds et al., 2004; Straus et al., 1995)	Maternal neglectful behavior toward their children (emotional neglect; cognitive neglect; supervisory neglect; physical needs neglect)	Mothers (NR)	To screen parents at highest risk of child neglect for prevention of its future occurrence	2 (8); range: 4–32	4-point ordinal scale (1 = *strongly disagree* to 4 = *strongly agree*)	Past 1 year
Parent–Child Aggression Acceptability Movie Task (P-CAAM; Rodriguez et al., 2011)	Acceptance of parent–child aggression (physical discipline; physical abuse)	Current and prospective parent populations (NR)	To assess intervention programming outcomes	2 (8 video clips: 90 s each); range: 0–NR	Clips build toward "initial physical contact between caregiver and child"; rater should identify that moment and stop video; delay between actual physical contact and stop video = score (per video)	Not specified
Parent Opinion Questionnaire (POQ; Twentyman et al., 1981, November)	Parental expectations of child behavior (self-care; family responsibility and care of siblings; help and affection to parents; leaving children alone; proper behavior and feelings; punishment)	Parents (NR)	To identify abusive parents for child maltreatment service	6 (60); range: 0–60	Dichotomous scale (0 = *disagree* or 1 = *agree*)	Not specified
Parental Response to Child Misbehavior Questionnaire (PRCM; Holden & Zambarano, 1992; Vittrup et al., 2006)	Discipline techniques used by parents in response to their children's misbehaviors	Parents with young children (NR)	To obtain information regarding the frequency of specific discipline techniques	1 (12); range: 0–72	6-point ordinal scale (0 = *never* to 6 = ≥9 *times per week*)	Past 1 week
Shaken Baby Syndrome awareness assessment–Short Version (SBS-SV; Russell, 2010; Russell & Britner, 2006)	Shaken baby syndrome awareness (soothing techniques; discipline techniques; potential for injury)	Parents, babysitters, and childcare providers of young children (ages younger than 2)	To provide a measure for caregiver education and other service provision concerning the care of infants younger than 2 years	3 (36); range: 36–216	6-point ordinal scale (1 = *strongly disagree* to 6 = *strongly agree*)	Not specified

Note. All information was derived from all eligible studies and the original included instruments; NR = not reported.

Table 2. Methodological Quality Assessment of Development and Content Validity Studies on Content Validity of the Included Instruments.

| | | Development Study Quality[a] | | | Content Validity Study Quality[a] | | | | |
| | | Item Generation[b] | Cognitive Interview[b] | | Asking Parents or Carers[b] | | | Asking Professionals[b] | |
Instrument	Reference	Relevance	Comprehensiveness	Comprehensibility	Relevance	Comprehensiveness	Comprehensibility	Relevance	Comprehensiveness
AAPI-2	Bavolek et al. (1979)	NR	Inadequate (4.8%)	Inadequate (21.6%)	NR	NR	Doubtful (42.9%)	Doubtful (40.0%)	NR
APT	Zaidi et al. (1989)	NR	NR	NR	NR	NR	NR	NR	NR
CNQ	Stewart et al. (2015)	NR	NR	NR	NR	NR	NR	Doubtful (33.3%)	NR
CNS-MMS	Loeber et al. (1998)	NR	NR	NR	NR	NR	NR	NR	NR
CTS-ES	Lang and Connell (2017)	NR	NR	NR	NR	NR	NR	Doubtful (33.3%)	NR
CTSPC	Straus et al. (1998)	Inadequate (20.0%)	Inadequate (7.1%)	Doubtful (36.4%)	NR	NR	Doubtful (33.3%)	NR	NR
FM-CA	Heyman et al. (2019)	Doubtful (50.0%)	Inadequate (9.5%)	Inadequate (9.5%)	Doubtful (38.1%)	NR	Adequate (66.6%)	NR	NR
ICAST-Trial	Runyan et al. (2009)	NR	NR	NR	NR	NR	NR	NR	NR
	Meinck et al. (2018)	NR	NR	NR	Very good (76.2%)	NR	Very good (76.2%)	Doubtful (33.3%)	NR
IPPS	Gordon et al. (1979)	Inadequate (3.5%)	Inadequate (7.1%)	Inadequate (4.8%)	Inadequate (12.5%)	NR	NR	NR	NR
MCNS	Straus et al. (1995)	NR	NR	NR	NR	NR	NR	NR	NR
MCNS-SF	Straus et al. (1995)	NR	NR	NR	NR	NR	NR	NR	NR
P-CAAM	Rodriguez et al. (2011)	NR	NR	NR	NR	NR	NR	Doubtful (40.0%)	NR
POQ	Twentyman et al. (1981, November)	NR	NR	NR	Doubtful (38.1%)	NR	NR	Doubtful (40.0%)	NR
PRCM	Holden and Zambarano (1992)	NR	NR	NR	NR	NR	NR	NR	NR
SBS-SV	Russell and Britner (2006)	NR	Inadequate (7.1%)	Inadequate (7.1%)	NR	NR	NR	Doubtful (33.3%)	NR

Note. AAPI-2 = Adult Adolescent Parenting Inventory-2; APT = Analog Parenting Task; CNQ = Child Neglect Questionnaire; CNS-MMS = Child Neglect Scales–Maternal Monitoring and Supervision Scale; CTS-ES = Child Trauma Screen–Exposure Score; CTSPC = Conflict Tactics Scales: Parent–Child version; FM-CA = Family Maltreatment–Child Abuse criteria; ICAST-Trial = ISPCAN (International Society for the Prevention of Child Abuse and Neglect) Child Abuse Screening Tool for use in Trials; IPPS = Intensity of Parental Punishment Scale; MCNS = Mother–Child Neglect Scale; MCNS-SF = Mother–Child Neglect Scale–Short Form; P-CAAM = Parent–Child Aggression Acceptability Movie task; POQ = Parent Opinion Questionnaire; PRCM = Parental Response to Child Misbehavior Questionnaire; SBS-SV = Shaken Baby Syndrome awareness assessment–Short Version.
[a]The methodological quality per development and content validity study was rated using the COnsensus-based Standards for the selection of health Measurement INstruments checklist (Mokkink et al., 2010a). The overall methodological quality per study was presented as a percentage of the ratings (Cordier et al., 2015): inadequate = 0–25%; doubtful = 25.1–50%; adequate = 50.1–75%; very good = 75.1–100%; NR = not reported.
[b]The methodological quality was rated in the three aspects of content validity: relevance, comprehensiveness, and comprehensibility. The development study was rated in the two parts (item generation and cognitive interview); the content validity study was rated in the two study categories asking parents or carers and asking professionals about the relevance, comprehensiveness, and comprehensibility.

Table 3. Quality of Content Validity per Development and Content Validity Study, and Content of Instrument Itself.

Instrument	Reference	Relevance[a]			Comprehensiveness[a]			Comprehensibility[a]		
		Development Study	Content Validity Study	Content of Instrument	Development Study	Content validity Study	Content of Instrument	Development Study	Content Validity Study	Content of Instrument
AAPI-2	Bavolek et al. (1979)	?	?	+	?	?	+	?	?	+
APT	Zaidi et al. (1989)	?	?	+	?	?	+	?	?	+
CNQ	Stewart et al. (2015)	?	?	+	?	?	+	?	?	+
CNS-MMS	Loeber et al. (1998)	?	?	+	?	?	+	?	?	+
CTS-ES	Lang and Connell (2017)	?	?	±	?	?	−	?	?	+
CTSPC	Straus et al. (1998)	?	?	+	?	?	+	?	?	+
FM-CA	Heyman et al. (2019)	+	?	+	?	?	+	?	+	+
ICAST-Trial	Meinck et al. (2018); Runyan et al. (2009)	?	?	+	?	?	+	?	?	?
IPPS	Gordon et al. (1979)	?	?	+	?	?	+	?	?	+
MCNS	Straus et al. (1995)	?	?	+	?	?	+	?	?	+
MCNS-SF	Straus et al. (1995)	?	?	+	?	?	+	?	?	+
P-CAAM	Rodriguez et al. (2011)	?	?	?	?	?	?	?	?	?
POQ	Twentyman et al. (1981, November)	?	?	±	?	?	−	?	?	+
PRCM	Holden and Zambarano (1992)	?	?	+	?	?	+	?	?	±+
SBS-SV	Russell and Britner (2006)	?	?	+	?	?	+	?	?	+

Note. AAPI-2 = Adult Adolescent Parenting Inventory-2; APT = Analog Parenting Task; CNQ = Child Neglect Questionnaire; CNS-MMS = Child Neglect Scales–Maternal Monitoring and Supervision Scale; CTS-ES = Child Trauma Screen–Exposure Score; CTSPC = Conflict Tactics Scales: Parent–Child version; FM-CA = Family Maltreatment–Child Abuse criteria; ICAST-Trial = ISPCAN (International Society for the Prevention of Child Abuse and Neglect) Child Abuse Screening Tool for use in Trials; IPPS = Intensity of Parental Punishment Scale; MCNS = Mother–Child Neglect Scale; MCNS-SF = Mother–Child Neglect Scale–Short Form; P-CAAM = Parent–Child Aggression Acceptability Movie Task; POQ = Parent Opinion Questionnaire; PRCM = Parental Response to Child Misbehavior questionnaire; SBS-SV = Shaken Baby Syndrome awareness assessment–Short Version.

[a]The quality of content validity (relevance, comprehensiveness, and comprehensibility) per study and content of instrument was rated using the criteria for good content validity (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018): + = sufficient rating; − = insufficient rating; ? = indeterminate rating; ± = inconsistent rating. Rating for development and content validity studies was determined based on the data from development and content validity studies; rating for content of instrument was determined based on reviewers' subjective opinion on content of instrument itself (items and instructions).

Table 4. Overall Quality of Content Validity and Evidence Quality per Instrument.

Instrument	Relevance		Comprehensiveness		Comprehensibility	
	Overall Quality of Content Validity[a]	Quality of Evidence[b]	Overall Quality of Content Validity[a]	Quality of Evidence[b]	Overall Quality of Content Validity[a]	Quality of Evidence[b]
AAPI-2	+	Moderate	+	Very low	+	Very low
APT	+	Very low	+	Very low	+	Very low
CNQ	+	Moderate	+	Very low	+	Very low
CNS-MMS	+	Very low	+	Very low	+	Very low
CTS-ES	±	Low	−	Very low	+	Very low
CTSPC	+	Very low	+	Low	+	Low
FM-CA	+	Moderate	+	Very low	+	High
ICAST-Trial	+	High	+	Very low	+	High
IPPS	+	Moderate	+	Very low	+	Very low
MCNS	+	Very low	+	Very low	+	Very low
MCNS-SF	+	Very low	+	Very low	+	Very low
P-CAAM	?	NE	?	NE	?	NE
POQ	±	Low	−	Very low	+	Very low
PRCM	+	Very low	+	Very low	±	Very low
SBS-SV	+	Low	+	Very low	+	Very low

Note. AAPI-2 = Adult Adolescent Parenting Inventory-2; APT = Analog Parenting Task; CNQ = Child Neglect Questionnaire; CNS-MMS = Child Neglect Scales–Maternal Monitoring and Supervision Scale; CTS-ES = Child Trauma Screen–Exposure Score; CTSPC = Conflict Tactics Scales: Parent–Child version; FM-CA = Family Maltreatment–Child Abuse criteria; ICAST-Trial = ISPCAN (International Society for the Prevention of Child Abuse and Neglect) Child Abuse Screening Tool for use in Trials; IPPS = Intensity of Parental Punishment Scale; MCNS = Mother–Child Neglect Scale; MCNS-SF = Mother–Child Neglect Scale-Short Form; P-CAAM = Parent–Child Aggression Acceptability Movie Task; POQ = Parent Opinion Questionnaire; PRCM = Parental Response to Child Misbehavior questionnaire; SBS-SV = Shaken Baby Syndrome awareness assessment–Short Version.
[a] The overall quality of content validity (relevance, comprehensiveness, and comprehensibility) was determined by qualitatively summarizing all ratings on content validity per study of each instrument and reviewers' ratings on content of instrument itself (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018): + = sufficient rating; ? = indeterminate rating; − = insufficient rating; ± = inconsistent rating.
[b] The quality of evidence (confidence level for the overall quality rating of content validity) was rated using a modified Grading of Recommendations Assessment, Development and Evaluation approach (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018); high = high level of confidence; moderate = moderate level of confidence; low = low level of confidence; very low = very low level of confidence; NE = not evaluated (instruments could not be retrieved).

instruments received very low quality of evidence ratings (APT, CNS-MMS, CTSPC, MCNS, MCNS-SF, and PRCM). Three instruments were rated as having low quality of evidence (CTS-ES, POQ, and SBS-SV); four instruments were rated as having moderate quality of evidence (AAPI-2, CNQ, FM-CA, and IPPS); one instrument (ICAST-Trial) was rated as having high quality of evidence; and one instrument (P-CAAM) was not evaluated (NE) because of indeterminate overall ratings (i.e., lack of evidence). All instruments received a very low quality of evidence for the overall ratings in comprehensiveness, except for the following two instruments: CTSPC reported low-quality evidence and P-CAAM was not evaluated (NE). For overall ratings of comprehensibility, only two instruments received high quality of evidence ratings (FM-CA and ICATS-Trial), whereas all other instruments (except CTSPC and P-CAAM) received very low ratings.

Discussion

The aim of this systematic review was to determine the quality of content validity of all current parent or caregiver report instruments measuring CM by parents or caregivers. This review identified 15 instruments and 15 corresponding instrument development and content validity studies of the instruments. Findings from the systematic review

demonstrate lack of high-quality evidence, suggesting that none of the instruments received high-quality ratings for all three aspects of content validity (relevance, comprehensiveness, and comprehensibility). As such, none of the instruments have unequivocally support for their use in terms of the quality of content validity.

Instrument Development Study

The majority of instrument development studies did not address SA as a construct of interest to be measured. While most CM instruments had a scale or subscale related to PA, EA, and/or neglect, only three instruments had some items or a subscale related to SA: a single item of the CTS-ES, 2 items of the ICAST-Trial, and one optional supplementary subscale of the CTSPC. A recent meta-analysis on who perpetrates CM reported that most SA is perpetrated by people other than parents or caregivers compared with the other three types of CM, but this result was only based on child self-report and professional report instruments due to lack of studies reporting SA by using parent report instruments (Devries et al., 2018). To verify the exceptional lower prevalence rates of SA perpetrated by parents, comparison of prevalence rates reported by parents, children, and professionals should be conducted. However, based on the findings from this review, comparing the

prevalence rates of SA reported between parents or caregivers, children and professionals may be challenging because of the lack of parent report instruments on SA.

Many instrument development studies generated new items without involvement of the target population (parents or caregivers), that is, most instrument items were generated based on a review of relevant literature, commonly used instruments, or professional input by developers themselves. Involvement of the target population is essential to ensure adequate content validity in the generation of new instrument items (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Involving the target population through individual interviews or focus groups helps to identify items that are relevant to the target population, to ensure items are based on their own experience or perceptions related to the construct being measured (Ricci et al., 2018). If the respondents (target population) are of the opinion that the instrument items are irrelevant, the instrument could fail to measure respondents' attitudes and behaviors accurately (Wiering et al., 2017). Therefore, development studies of new instrument items as reported in this review may have significant methodological flaws given the lack of target population involvement.

Content Validity Study

Only a few content validity studies asked parents or caregivers about relevance, comprehensiveness, and comprehensibility of the instruments and reported specific research methods and results, which enabled the evaluation of the content validity of the instruments clearly. According to findings on the methodological quality of content validity studies, relevance of the final version of instruments was mostly evaluated by asking the professionals, whereas, surprisingly, the comprehensiveness of instruments was not evaluated by neither professionals nor parents or caregivers. Furthermore, the comprehensibility (i.e., how easy it is for respondents to understand instrument items) was rarely evaluated by parents or caregivers as respondents. The few studies that did evaluate the relevance and comprehensibility of instruments using parents or caregivers as respondents lacked the required detail when reporting on the methodology (e.g., insufficient reporting on study design and results). These weaknesses made it difficult to determine whether the content validity of instruments was positive or negative based on the evidence obtained from the content validity studies.

Synthesis of Evidence on Content Validity

Given that content validity is the first psychometric property to consider when selecting an instrument, the inadequate quality of evidence on content validity makes it difficult to select the best instrument(s); Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). The majority of ratings (88/99) on relevance, comprehensiveness, and comprehensibility based on the development and content validity studies were categorized as indeterminate. Due to these indeterminate study ratings, most

overall ratings on relevance, comprehensiveness, and comprehensibility were determined based on reviewers' subjective opinion about the content of instrument itself only. The results indicate lack of evidence on content validity or inappropriate methodological approaches used for instrument development and content validity studies (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Due to the largely inappropriate methodological approaches used when developing new instruments and assessing content validity of the instruments, in most instances, evidence on the quality of relevance, comprehensiveness, and comprehensibility was very low; high-quality evidence was found only for the relevance or comprehensibility for two instruments (FM-CA and ICAST-Trial). Therefore, findings from this review indicate that evidence of the quality of content validity of parent or caregiver report CM instruments is very uncertain.

Based on available evidence on content validity for the 15 included instruments, the ICAST-Trial seems to be the most promising instrument in terms of content validity; however, the evidence is not conclusive. The ICAST-Trial displayed high-quality evidence for sufficient relevance and comprehensibility and very low evidence for sufficient comprehensiveness. The next most promising instrument was the FM-CA with high-quality evidence for sufficient comprehensibility, moderate evidence for sufficient relevance, and very low evidence for sufficient comprehensiveness. While none of the remaining 13 instruments reported high-quality evidence on any aspects of content validity, they also have the potential to be used in terms of content validity because no high-quality evidence for insufficient relevance, comprehensiveness, or comprehensibility was found.

Limitations

This systematic review has some limitations. Firstly, only instruments developed and validated in English and psychometric studies published in English were considered. Thus, findings on content validity of parent or carer report CM instruments developed in languages other than English may have been excluded. Secondly, despite contacting the developer of the P-CAAM, we failed to retrieve the original instrument from the authors or from literature and, therefore, could not determine the overall ratings on content validity of this instrument. Lastly, while rating the quality of the studies and psychometric properties using the COSMIN guidelines for assessing content validity required a degree of subjective judgment by reviewers, all ratings for this review were conducted by two reviewers independently and disagreements were resolved through consensus.

Conclusion

Fifteen parent or caregiver report CM instruments were retrieved. An evaluation of the content validity using the COS-MIN methodological guidelines found that the ICAST-Trial appears to be the most promising instrument, followed by the

FM-CA, but firm conclusions cannot be drawn because evidence concerning the content validity is limited and mostly of low quality. However, no high-quality evidence was found to indicate that the content validity is insufficient. As such, all identified instruments have the potential to be used, but their remaining psychometric properties should be evaluated. A companion paper (Part 2) will report on the evaluation of the remaining psychometric properties of the 15 included instruments to identify parent or caregiver report instruments of CM with robust psychometric properties based on current evidence.

Implication for Research and Practice

There is a need for follow-up studies on parent-reported CM questionnaires to be conducted with the following five recommendations in mind. First, future instrument development studies should include SA parent-reported items or subscales, especially in the case of early childhood SA where recall bias in young children is an important consideration. Second, development of a new instrument items should involve parents or caregivers (e.g., individual or group interviews) to identify relevant items from their perspective on CM. Third, additional validation studies are needed to evaluate content validity of the included instruments, as current evidence on their content validity is not enough to determine conclusively which of the instruments has good content validity. In particular, the comprehensibility of the instruments should be further evaluated from the perspectives of parents or caregivers. Fourth, it is recommended that future studies apply the COSMIN guidelines in their study design for the generation of new items and assessment of content validity of instruments. Finally, a review on quality of the remaining psychometric properties of current parent or caregiver report CM instruments is needed, as no high-quality evidence of insufficient content validity was found. This additional assessment of psychometric quality will help clinicians and researchers decided which instruments to use for their interventions and research on CM perpetrated by parents or caregivers.

ORCID iD

Sangwon Yoon ![ORCID] https://orcid.org/0000-0002-9959-3808
Reinie Cordier ![ORCID] https://orcid.org/0000-0002-9906-5300
Airi Hakkarainen ![ORCID] https://orcid.org/0000-0001-5199-3493

References

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman & Hall.

Altmann, T. K. (2008). Attitude: A concept analysis. *Nursing Forum*, *43*, 144–150. http://doi.org/10.1111/j.1744-6198.2008.00106.x

Anda, R. F., Butchart, A., Felitti, V. J., & Brown, D. W. (2010). Building a framework for global surveillance of the public health implications of adverse childhood experiences. *American Journal of Preventive Medicine*, *39*, 93–98. http://doi.org/10.1016/j.amepre.2010.03.015

Asadollahi, M., Jabraeili, M., Asghari Jafarabadi, M., & Hallaj, M. (2016). Parents' attitude toward child abuse conducted in the health centers of Tabriz. *International Journal of School Health*, *3*, e60221. http://doi.org/10.17795/intjsh-31198

Ateah, C. A., & Durrant, J. E. (2005). Maternal use of physical punishment in response to child misbehavior: Implications for child abuse prevention. *Child Abuse & Neglect*, *29*, 169–185. http://doi.org/10.1016/j.chiabu.2004.10.010

Bailhache, M., Leroy, V., Pillet, P., & Salmi, L. R. (2013). Is early detection of abused children possible? A systematic review of the diagnostic accuracy of the identification of abused children. *BMC Pediatrics*, *13*, 202. http://doi.org/10.1186/1471-2431-13-202

Bavolek, S. J., & Keene, R. G. (1999). *Adult-Adolescent Parenting Inventory-AAPI-2: Administration and development handbook*. Family Development Resources, Inc.

Bavolek, S. J., Kline, D. F., McLaughlin, J. A., & Publicover, P. R. (1979). Primary prevention of child abuse and neglect: Identification of high-risk adolescents. *Child Abuse & Neglect*, *3*, 1071–1080. http://doi.org/10.1016/0145-2134(79)90152-2

Boden, J. M., Horwood, L. J., & Fergusson, D. M. (2007). Exposure to childhood sexual and physical abuse and subsequent educational achievement outcomes. *Child Abuse & Neglect*, *31*, 1101–1114. http://doi.org/10.1016/j.chiabu.2007.03.022

Bower-Russa, M. (2005). Attitudes mediate the association between childhood disciplinary history and disciplinary responses. *Child Maltreatment*, *10*, 272–282. http://doi.org/10.1177/1077559505277531

Chavis, A., Hudnut-Beumler, J., Webb, M. W., Neely, J. A., Bickman, L., Dietrich, M. S., & Scholer, S. J. (2013). A brief intervention affects parents' attitudes toward using less physical punishment. *Child Abuse & Neglect*, *37*, 1192–1201. http://doi.org/10.1016/j.chiabu.2013.06.003

Chen, M., & Chan, K. L. (2016). Effects of parenting programs on child maltreatment prevention: A meta-analysis. *Trauma, Violence, & Abuse*, *17*, 88–104. http://doi.org/10.1177/1524838014566718

Chiarotto, A. (2019). Patient-reported outcome measures: Best is the enemy of good but what if good is not good enough? *Journal of Orthopaedic & Sports Physical Therapy*, *49*, 39–42. http://doi.org/10.2519/jospt.2019.0602

Cohen, J., & Humphreys, L. H. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.

Psychological Bulletin, *70*, 213–220. http://doi.org/10.1037/h0026256

Compier-de Block, L. H. C. G., Alink, L. R. A., Linting, M., van den Berg, L. J. M., Elzinga, B. M., Voorthuis, A., Tollenaar, M. S., & Bakermans-Kranenburg, M. J. (2017). Parent-child agreement on parent-to-child maltreatment. *Journal of Family Violence*, *32*, 207–217. http://doi.org/10.1007/s10896-016-9902-3

Cordier, R., Speyer, R., Chen, Y., Wilkes-Gillan, S., Brown, T., Bourke-Taylor, H., Doma, K., & Leicht, A. (2015). Evaluating the psychometric quality of social skills measures: A systematic review. *PLoS One*, *10*, e0132299–e0132299. http://doi.org/10.1371/journal.pone.0132299

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

Currie, J., & Spatz Widom, C. (2010). Long-term consequences of child abuse and neglect on adult economic well-being. *Child Maltreatment*, *15*, 111–120. http://doi.org/10.1177/1077559509355316

Danese, A., & McEwen, B. S. (2012). Adverse childhood experiences, allostasis, allostatic load, and age-related disease. *Physiology & Behavior*, *106*, 29–39. http://doi.org/10.1016/j.physbeh.2011.08.019

Devries, K., Knight, L., Petzold, M., Merrill, K. G., Maxwell, L., Williams, A., Cappa, C., Chan, K. L., Garcia-Moreno, C., Hollis, N., Kress, H., Peterman, A., Walsh, S. D., Kishor, S., Guedes, A., Bott, S., Riveros, B. C. B., Watts, C., & Abrahams, N. (2018). Who perpetrates violence against children? A systematic analysis of age-specific and sex-specific data. *BMJ Paediatrics Open*, *2*, e000180. http://doi.org/10.1136/bmjpo-2017-000180

Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., Koss, M. P., & Marks, J. S. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The Adverse Childhood Experiences (ACE) study. *American Journal of Preventive Medicine*, *14*, 245–258. http://doi.org/10.1016/S0749-3797(98)00017-8

Finkelhor, D., Shattuck, A., Turner, H. A., & Hamby, S. L. (2014). The lifetime prevalence of child sexual abuse and sexual assault assessed in late adolescence. *Journal of Adolescent Health*, *55*, 329–333. http://doi.org/10.1016/j.jadohealth.2013.12.026

Gershoff, E. T., Lee, S. J., & Durrant, J. E. (2017). Promising intervention strategies to reduce parents' use of physical punishment. *Child Abuse & Neglect*, *71*, 9–23. http://doi.org/10.1016/j.chiabu.2017.01.017

Glaser, D. (2000). Child abuse and neglect and the brain—A review. *Journal of Child Psychology and Psychiatry*, *41*, 97–116. http://doi.org/10.1111/1469-7610.00551

Gordon, D. A., Jones, R. H., & Nowicki, S. (1979). A measure of intensity of parental punishment. *Journal of Personality Assessment*, *43*, 485–496. http://doi.org/10.1207/s15327752jpa4305_9

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., Schünemann, H. J., & GRADE Working Group. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, *336*, 924–926. http://doi.org/10.1136/bmj.39489.470347.AD

Heyman, R. E., Snarr, J. D., Slep, A. M. S., Baucom, K. J. W., & Linkh, D. J. (2019). Self-reporting DSM–5/ICD-11 clinically significant intimate partner violence and child abuse: Convergent and response process validity. *Journal of Family Psychology*. http://doi.org/10.1037/fam0000560

Hillis, S., Mercy, J., Amobi, A., & Kress, H. (2016). Global prevalence of past-year violence against children: A systematic review and minimum estimates. *Pediatrics*, *137*, 1–13. http://doi.org/10.1542/peds.2015-4079

Holden, G. W., Brown, A. S., Baldwin, A. S., & Croft Caderao, K. (2014). Research findings can change attitudes about corporal punishment. *Child Abuse & Neglect, 38,* 902–908. http://doi.org/10.1016/j.chiabu.2013.10.013

Holden, G. W., & Zambarano, R. J. (1992). Passing the rod: Similarities between parents and their young children in orientations toward physical punishment. In I. E. Sigel, A. V. McGillicuddy-DeLisi, & J. J. Goodnow (Eds.), *Parental belief systems: The psychological consequences for children* (2nd ed., pp. 143–172). Lawrence Erlbaum Associates.

Kirisci, L., Dunn, M. G., Mezzich, A. C., & Tarter, R. E. (2001). Impact of parental substance use disorder and child neglect severity on substance use involvement in male offspring. *Prevention Science*, *2*, 241–255. http://doi.org/10.1023/a:1013662132189

Krug, E. G., Linda, L. D., James, A. M., Anthony, B. Z., & Rafael, L. (Eds.). (2002). *World report on violence and health*. Word Health Organization.

Lang, J. M., & Connell, C. M. (2017). Development and validation of a brief trauma screening measure for children: The child trauma screen. *Psychological Trauma: Theory, Research, Practice, and Policy, 9,* 390–398. http://doi.org/10.1037/tra0000235

Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., & Van Kammen, W. B. (1998). *Antisocial behavior and mental health problems: Explanatory factors in childhood and adolescence*. Lawrence Erlbaum Associates.

Lounds, J. J., Borkowski, J. G., & Whitman, T. L. (2004). Reliability and validity of the mother-child neglect scale. *Child Maltreatment*, *9*, 371–381. http://doi.org/10.1177/1077559504269536

Meinck, F., Boyes, M. E., Cluver, L., Ward, C. L., Schmidt, P., Destone, S., & Dunne, M. P. (2018). Adaptation and psychometric properties of the ISPCAN Child Abuse Screening Tool for use in trials (ICAST-Trial) among South African adolescents and their primary caregivers. *Child Abuse & Neglect*, *82*, 45–58. http://doi.org/10.1016/j.chiabu.2018.05.022

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*, e1000097. http://doi.org/10.1371/journal.pmed.1000097

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

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010a). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement

instruments: An international Delphi study. *Quality of Life Research*, *19*, 539–549. http://doi.org/10.1007/s11136-010-9606-8

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

Negriff, S., Schneiderman, J. U., & Trickett, P. K. (2017). Concordance between self-reported childhood maltreatment versus case record reviews for child welfare–affiliated adolescents: Prevalence rates and associations with outcomes. *Child Maltreatment*, *22*, 34–44. http://doi.org/10.1177/1077559516674596

Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011). Content validity—Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part 2—Assessing respondent understanding. *Value in Health*, *14*, 978–988. http://doi.org/10.1016/j.jval.2011.06.01

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

Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P. R., & Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set"—A practical guideline. *Trials*, *17*, 449. http://doi.org/10.1186/s13063-016-1555-2

Ricci, L., Lanfranchi, J., Lemetayer, F., Rotonda, C., Guillemin, F., Coste, J., & Spitz, E. (2018). Qualitative methods used to generate questionnaire items: A systematic review. *Qualitative Health Research*, *29*, 149–156. http://doi.org/10.1177/1049732318783186

Rodriguez, C. M., Russa, M. B., & Harmon, N. (2011). Assessing abuse risk beyond self-report: Analog task of acceptability of parent-child aggression. *Child Abuse & Neglect*, *35*, 199–209. http://doi.org/10.1016/j.chiabu.2010.12.004

Runyan, D. K., Dunne, M. P., Zolotor, A. J., Madrid, B., Jain, D., Gerbaka, B., Menick, D. M., Andreva-Miller, I., Kasim, M. S., Choo, W. Y., Isaeva, O., Macfarlane, B., Ramirez, C., Volkova, E., & Youssef, R. M. (2009). The development and piloting of the ISPCAN child abuse screening tool—Parent version (ICAST-P). *Child Abuse & Neglect*, *33*, 826–832. http://doi.org/10.1016/j.chiabu.2009.09.006

Russa, M. B., & Rodriguez, C. M. (2010). Physical discipline, escalation, and child abuse potential: Psychometric evidence for the Analog Parenting Task. *Aggressive Behavior*, *36*, 251–260. http://doi.org/10.1002/ab.20345

Russell, B. S. (2010). Revisiting the measurement of shaken baby syndrome awareness. *Child Abuse & Neglect*, *34*, 671–676. http://doi.org/10.1016/j.chiabu.2010.02.008

Russell, B. S., & Britner, P. A. (2006). Measuring Shaken Baby Syndrome awareness: Preliminary reliability of a caregiver attitudes

and beliefs survey. *Journal of Child and Family Studies*, *15*, 765–777. http://doi.org/10.1007/s10826-006-9050-0

Saini, S. M., Hoffmann, C. R., Pantelis, C., Everall, I. P., & Bousman, C. A. (2019). Systematic review and critical appraisal of child abuse measurement instruments. *Psychiatry Research*, *272*, 106–113. http://doi.org/10.1016/j.psychres.2018.12.068

Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011). What makes a measurement instrument valid and reliable? *Injury*, *42*, 236–240. http://doi.org/10.1016/j.injury.2010.11.042

Sedlak, A. J., Mettenburg, J., Basena, M., Petta, I., McPherson, K., Greene, A., & Li, S. (2010). *Fourth National Incidence Study of Child Abuse and Neglect (NIS–4): Report to Congress*. Administration for Children and Families.

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

Speyer, R., Cordier, R., Kertscher, B., & Heijnen, B. J. (2014). Psychometric properties of questionnaires on functional health status in oropharyngeal dysphagia: A systematic literature review. *BioMed Research International*, *2014*, 458–678. http://doi.org/10.1155/2014/458678

Sprangers, M. A. G., & Aaronson, N. K. (1992). The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. *Journal of Clinical Epidemiology*, *45*, 743–760. http://doi.org/10.1016/0895-4356(92)90052-O

Stewart, C., Kirisci, L., Long, A. L., & Giancola, P. R. (2015). Development and psychometric evaluation of the child neglect questionnaire. *Journal of Interpersonal Violence*, *30*, 3343–3366. http://doi.org/10.1177/0886260514563836

Stith, S. M., Liu, T., Davies, L. C., Boykin, E. L., Alder, M. C., Harris, J. M., Som, A., McPherson, M., & Dees, J. (2009). Risk factors in child maltreatment: A meta-analytic review of the literature. *Aggression and Violent Behavior*, *14*, 13–29. http://doi.org/10.1016/j.avb.2006.03.006

Stoltenborgh, M., Bakermans-Kranenburg, M. J., Alink, L. R. A., & Ijzendoorn, M. H. (2015). The prevalence of child maltreatment across the globe: Review of a series of meta-analyses. *Child Abuse Review*, *24*, 37–50. http://doi.org/10.1002/car.2353

Straus, M. A., Hamby, S. L., Finkelhor, D., Moore, D. W., & Runyan, D. (1998). Identification of child maltreatment with the Parent-Child Conflict Tactics Scales: Development and psychometric data for a national sample of American parents. *Child Abuse & Neglect*, *22*, 249–270. http://doi.org/10.1016/S0145-2134(97)00174-9

Straus, M. A., Hamby, S. L., & Warren, W. L. (2003). *The Conflict Tactics Scales handbook: Revised Conflict Tactics Scales (CTS2) and CTS—Parent-child version (CTSPC)*. Western Psychological Services.

Straus, M. A., Kinard, E. M., & Williams, L. M. (1995). *The multidimensional neglectful behavior scale, Form A: Adolescent and adult-recall version*. Family Research Laboratory, University of New Hampshire.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Teicher, M. H., Samson, J. A., Anderson, C. M., & Ohashi, K. (2016). The effects of childhood maltreatment on brain structure, function and connectivity. *Nature Reviews Neuroscience*, *17*, 652–666. http://doi.org/10.1038/nrn.2016.111

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*, 34–42. http://doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., de Vet, H. C. W., Bouter, L. M., Alonso, J., Westerman, M. J., Patrick, D. L., & Mokkink, L. B. (2018). *COSMIN methodology for assessing the content validity of PROMs—User manual* (Version 1.0). https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf

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

Twentyman, C. T., Plotkin, R., Dodge, D., & Rohrbeck, C. A. (1981, November). *Inappropriate expectations of parents who maltreat their children*. Paper presented at the Annual Meeting of the Association for Advancement of Behavior Therapy, Toronto.

van Harmelen, A., van Tol, M., van der Wee, N. J. A., Veltman, D. J., Aleman, A., Spinhoven, P., van Buchem, M. A., Zitman, F. G., Penninx, B. W. J. H., & Elzinga, B. M. (2010). Reduced medial prefrontal cortex volume in adults reporting childhood emotional maltreatment. *Biological Psychiatry*, *68*, 832–838. http://doi.org/10.1016/j.biopsych.2010.06.011

Vittrup, B., Holden, G. W., & Buck, J. (2006). Attitudes predict the use of physical punishment: A prospective study of the emergence of disciplinary practices. *Pediatrics*, *117*, 2055–2064. http://doi.org/10.1542/peds.2005-2204

Voisine, S., & Baker, A. J. L. (2012). Do universal parenting programs discourage parents from using corporal punishment: A program review. *Families in Society: The Journal of Contemporary Social Services*, *93*, 212–218. http://doi.org/10.1606/1044-3894.4217

Wiering, B., de Boer, D., & Delnoij, D. (2017). Patient involvement in the development of patient-reported outcome measures: A scoping review. *Health Expectations*, *20*, 11–23. http://doi.org/10.1111/hex.12442

World Health Organization. (1999). *Report of the consultation on child abuse prevention*. Author. https://apps.who.int/iris/handle/10665/65900

World Health Organization. (2016). *INSPIRE: Seven strategies for ending violence against children*. Author. http://apps.who.int/iris/bitstream/10665/207717/1/9789241565356-eng.pdf?ua=1

Zaidi, L. Y., Knutson, J. F., & Mehm, J. G. (1989). Transgenerational patterns of abusive parenting—Analog and clinical-tests. *Aggressive Behavior*, *15*, 137–152. http://doi.org/10.1002/1098-2337(1989)15:2<137::AID-AB2480150202>3.0.CO;2-O

Author Biographies

Sangwon Yoon, MPhil, is a PhD candidate at the Department of Special Needs Education, University of Oslo in Norway.

Renée Speyer, PhD, is a professor at the Department of Special Needs Education, University of Oslo in Norway.

Reinie Cordier, PhD, is a professor at the Department of Special Needs Education, University of Oslo in Norway.

Pirjo Aunio, PhD, is a professor at the Department of Education, University of Helsinki in Finland.

Airi Hakkarainen, PhD, is a university lecturer in the field of special needs education at the Open University, University of Helsinki in Finland.