Evaluating Psychometric Properties of Parent- or Caregiver-Report Instruments on Child Maltreatment

Systematic Reviews Using the COSMIN Methodology

Sangwon Yoon

Doctoral Thesis

Submitted in partial fulfilment of the requirements for
the degree of Doctor of Philosophy (PhD)
Department of Special Needs Education
Faculty of Educational Sciences

UNIVERSITY OF OSLO

2022

Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 343

ISSN 1501-8962

Cover: Hanne Baadsgaard Utigard.

Print production: Reprosentralen, University of Oslo.

Acknowledgements

Throughout the writing of this dissertation, I have received a great deal of support and assistance, for which I am sincerely grateful.

I would first like to thank my main supervisor, Professor Renée Speyer, whose expertise was invaluable in formulating the research questions and methodology. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Many thanks go to my co-supervisors, Professors Reinie Cordier and Pirjo Aunio, for their valuable guidance throughout my PhD journey. They provided me with the tools that I needed to choose the right direction for and successfully complete my dissertation.

I especially want to thank my Finnish colleague, Dr Airi Hakkarainen, for her wonderful collaboration on my research. To my friends Laoura Ziaka and Oleg Zacharov, I thank you for your encouragement and friendship in our little office.

Finally, I must give my utmost thanks to my wife, Bohey Kim, for her devotion. Without her understanding and support, I could not have completed this dissertation. I must also thank my daughter, Wimin Yoon, who provided many happy distractions, which allowed me to rest my mind outside my research.

Kind regards,

Sangwon Yoon

Oslo, August 2021

Summary

Background: Child maltreatment (CM) is a public health problem with devastating lifelong consequences for victims of CM. The United Nations (UN) launched an initiative to eliminate CM as part of their 2030 Agenda for Sustainable Development Goals. To monitor progress towards achieving the goal of eradicating CM, all UN member states should annually report their national CM prevalence and progress in reducing CM. However, no consensus has been reached on which instruments are best for investigating CM.

Aim: This thesis aimed to evaluate the psychometric properties of all currently available parent- or caregiver-report instruments on any type of CM and recommend those with the best psychometric quality.

Method: A systematic search of six databases (CINAHL, Embase, ERIC, PsycINFO, PubMed and Sociological Abstracts) was conducted by following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement. The assessment of psychometric properties was performed using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology for assessing the psychometric properties of patient-report outcome instruments in a systematic review. The scale, scope, and sophistication of reporting lead to the reporting of psychometric properties in three separate review papers: Paper 1 addressed the content validity (i.e., the extent to which the content of an instrument adequately reflects the construct measured) of identified measures; Paper 2 covered construct validity (i.e., the extent to which an instrument is consistent with a hypothesis regarding the relationships with other instruments or differences between groups), criterion validity (i.e., the extent to which an instrument adequately reflects a gold standard), and reliability (i.e., the extent to which the measurement is free from measurement error) of identified measures; and Paper 3 addressed the responsiveness (i.e., the ability of an instrument to detect changes in the measured construct over time) of identified measures.

Results: In total, 109 development and validation studies reporting on the psychometric properties of 15 selected instruments were included: 15 studies reported on the content validity; 25 studies reported on the construct validity, criterion validity, and reliability; and 69 studies reported on the responsiveness. The methodological quality of the studies was generally adequate; however, the quality of the studies reporting on content validity was poor overall. The psychometric quality of the instruments’ content validity was generally sufficient, but sufficient quality was determined based on reviewers’ subjective opinions of the content of the instrument itself (items, response options, and instructions) due to the lack of direct evidence from the studies. The psychometric quality of the construct validity, criterion validity, and reliability were overall either indeterminate or not reported because of incomplete or missing data on the psychometric properties. The quality of the responsiveness was also overall either insufficient or not reported. High-quality evidence on all psychometric properties was limited.

Recommendations: None of the included instruments can be recommended as the most suitable for use in clinical practice and research. Nine instruments are promising based on the available psychometric evidence, but need additional psychometric evidence before they can be recommended.

Keywords: *assessment; child abuse; child neglect; COSMIN; measurement properties; parent report.*

List of Articles

Article 1:

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 1: Content Validity. *Trauma, Violence, & Abuse.* Advanced online publication. https://doi.org/10.1177/1524838019898456 (Impact Factor: 10.570)

Article 2:

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 2: Internal Consistency, Reliability, Measurement Error, Structural Validity, Hypothesis Testing, Cross-Cultural Validity, and Criterion Validity. *Trauma, Violence, & Abuse*. Advanced online publication. https://doi.org/10.1177/1524838020915591 (Impact Factor: 10.570)

Article 3:

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Responsiveness of Parent- or Caregiver-Reported Child Maltreatment Instruments to Parenting Interventions. *Trauma, Violence, & Abuse*. Manuscript submitted for publication. (Impact Factor: 10.570)

Table of Contents

List of Tables

List of Figures

Abbreviations

AAPI-2	Adult Adolescent Parenting Inventory-2
ANOVA	Analysis of variance
APT	Analog Parenting Task
CAPTA	Child Abuse and Prevention Treatment Act
CFA	Confirmative factor analysis
CM	Child maltreatment
CNQ	Child Neglect Questionnaire
CNS-MMS	Child Neglect Scales-Maternal Monitoring and Supervision Scale
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
CPS	Child Protection Services
CTS-ES	Child Trauma Screen-Exposure Score
CTSPC	Conflict Tactics Scales: Parent-Child version
DIF	Differential item functioning
DSM-5	Diagnostic and Statistical Manual-5th revision
EFA	Exploratory factor analysis
FM-CA	Family Maltreatment-Child Abuse criteria
GRADE	Grading of Recommendations Assessment, Development and Evaluation
ICAST-Trial	International Society for the Prevention of Child Abuse and Neglect (ISPCAN) Child Abuse Screening Tool-for use in Trials
ICC	Intraclass correlation coefficient
ICD-11	International Classification of Diseases-11th revision
IPPS	Intensity of Parental Punishment Scale
IRT	Item response theory
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,
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomised controlled trial
RQ	Research question
SBS-SV	Shaken Baby Syndrome awareness assessment-short version
U.S.	United States
UN	United Nations
USDHHS	U.S. Department of Health Human Services
WHO	World Health Organization

PART Ⅰ

EXTENDED ABSTRACT

1 Introduction

This introduction briefly presents the background and rationale for this thesis. Furthermore, the aims and research questions of this systematic review are presented. Finally, the outline of this thesis is presented to summarise its overall structure.

1.1 Brief Background and Rationale

Worldwide more than one billion children between 2 and 17 years of age suffer from child maltreatment (CM; Hillis et al., 2016) and most cases of CM are perpetrated by parents or caregivers (Devries et al., 2018; Sedlak et al., 2010). Early exposure to CM can lead to long-term chronic illness, injuries and other physical damage, damage to vital organs, including the brain, and even death in severe cases (Anda et al., 2008; Corso et al., 2008; Repetti et al., 2002; Scarborough et al., 2009; Taylor et al., 2004). Severe cases of CM are common; approximately 155,000 children under 15 years of age die from CM worldwide every year (Gilbert, Widom, et al., 2009), which makes CM the second leading cause of death in childhood (Johnson, 2002) following unintentional injuries caused by incidents such falls and road traffic accidents (Liu et al., 2012). Furthermore, early exposure to CM is associated with serious psychosocial difficulties (e.g., aggression, depression, antisocial behaviour, self-destructive behaviour, and inappropriate sexual behaviour; Dhingra et al., 2015; Jaffee et al., 2004; Jones et al., 2004; Vachon et al., 2015), as well as cognitive developmental delay (e.g., lower IQ scores, language development delay, and poorer academic achievement; Pechtel & Pizzagalli, 2011).

Due to the widespread global prevalence and severe consequences of CM, the United Nations (UN, 2015) launched an initiative to eradicate CM as part of their 2030 Agenda for Sustainable Development Goals, in Target 16.2 *"…end abuse, exploitation, trafficking and all forms of violence against and torture of children"* (p. 25). To achieve the global goal towards ending CM, many countries have legally obligated all professionals (e.g., health professionals, social workers, and school teachers) working with children to report any suspected cases of CM (Greco et al., 2017; Pelletier & Knox, 2017) and have endeavoured to develop and implement effective interventions to prevent CM (Molnar et al., 2016). In addition, to monitor the progress towards ending CM, all member governments should report the estimated CM prevalence and improvements in terms of the reduction in CM after their governmental intervention every year from 2016 to 2030 (World Health Organization

[WHO], 2020). Thus, measuring the number of children exposed to CM and the intervention effects in reducing CM provides important data to support global efforts to eradicate CM.

However, measuring the prevalence of CM has been hampered by the use of non-standardised instruments (Hovdestad et al., 2015), which leads to wide variation in estimates within and between groups (Fang et al., 2015). In addition, the prevalence estimates for CM differ significantly depending on the informants. Child- or caregiver-reported CM prevalence is higher than that reported by professionals, including health professionals or child protection workers (Stoltenborgh et al., 2015). Since CM commonly occurs in private spaces (such as homes) without witnesses and is most often perpetrated by parents (Institute of Medicine and National Research Council, 2014), the actual incidences of CM are difficult to be accurately reported by individuals other than parents, caregivers, or children. Consequently, professionals tend to report only severe CM cases and not suspected mild cases (Negriff et al., 2016). In contrast, young children are likely to have more difficulties recalling abusive and neglecting behaviours than adult caregivers (Devries et al., 2018). Although caregiver-reported CM prevalence using the most standardised form of CM instruments appears to be less influenced by underreporting (Devries et al., 2018; Stoltenborgh et al., 2015) compared with CM prevalence measured with child- or professional-report instruments (Meinck et al., 2016), the accuracy of parent reports of their own CM perpetration is controversial as parents tend to respond in socially desirable ways (i.e., social desirability bias; Milner & Crouch, 1997). Thus, selecting reliable and valid parent- or caregiver-report instruments is critical for accurately estimating the prevalence of CM.

Apart from measuring parent-reported CM prevalence, it is critical to measure parents' attitudes towards CM (i.e., parents' values, beliefs, or feelings in relation to maltreating behaviour towards a child) to prevent CM (Altmann, 2008). Parents' attitudes towards CM are an important factor in predicting parental maltreating behaviour (Stith et al., 2009). A number of studies have found that parents who have more positive beliefs or values regarding CM are likely to engage in maltreating behaviours more frequently than parents with negative attitudes towards CM (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, several studies on preventing CM have used instruments to assess parents' attitudes towards CM as outcome instruments to evaluate the effectiveness of prevention programs (Chen &

Chan, 2015; Gershoff et al., 2017; Holden et al., 2014; Voisine & Baker, 2012). Thus, to assess the outcomes of evidence-based programs to prevent CM, reliable and valid instruments are needed to assess parents' attitudes towards CM, as well as parents' maltreating behaviours towards their children.

The best way to select the most reliable and valid evidence-based instruments is to conduct a systematic review to evaluate the instruments' psychometric properties (Scholtes et al., 2011), including validity (i.e., the degree to which an instrument measures the construct it purposes to measure), reliability (i.e., the degree to which scores are the same for repeated measurements), and responsiveness (i.e., the ability to detect clinically important changes over time in the construct of interest; Mokkink et al., 2010). In the selection of an instrument, the most important psychometric property is its content validity (i.e., the extent to which the content of an instrument adequately reflects the construct measured; Mokkink et al., 2010). If the construct(s) that the instrument measures (i.e., content validity) is unclear, then it is meaningless to evaluate its reliability, responsiveness, and other types of validity (beyond content validity), including its construct validity (i.e., the extent to which an instrument is consistent with a hypothesis on relationships with other instruments or differences between groups; Patrick et al., 2011; Prinsen et al., 2018; Streiner et al., 2015) and criterion validity (i.e., the extent to which an instrument adequately reflects a gold standard as a single error-free reference measure; Naaktgeboren et al., 2013). No systematic review has been conducted to date on the psychometric properties of parent- or caregiver-report CM instruments published to date.

1.2 Aims and Research Questions

The overall aim of this thesis was to recommend the most suitable parent- or caregiver-report CM instruments for use in clinical practice and research based on their psychometric quality. To achieve this overall aim, the following four research questions (RQs) for this thesis were formulated:

- **RQ 1.** Which parent- or caregiver-report instruments have been published to measure their attitudes towards CM or maltreating behaviours towards their children?
- **RQ 2.** What is the quality of studies and psychometric evidence on the content validity of the existing parent- or caregiver-report CM instruments?

- **RQ 3.** What is the quality of studies and psychometric evidence on the construct validity, criterion validity, and reliability of the existing parent- or caregiver-report CM instruments?
- **RQ 4.** What is the quality of studies and psychometric evidence on the responsiveness of the existing parent- or caregiver-report CM instruments?

To address the overall aim and research questions of this thesis, three systematic reviews were conducted to identify the existing instruments and evaluate their psychometric properties (see Figure 1.1).

RESEARCH QUESTIONS	ARTICLES
RQ 1 & 2	**Article 1.** Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 1: Content Validity. *Trauma, Violence, & Abuse.* Advanced online publication. https://doi.org/10.1177/1524838019898456
RQ 3	**Article 2.** Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 2: Internal Consistency, Reliability, Measurement Error, Structural Validity, Hypothesis Testing, Cross-Cultural Validity, and Criterion Validity. *Trauma, Violence, & Abuse*. Advanced online publication. https://doi.org/10.1177/1524838020915591
RQ 4	**Article 3.** Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Responsiveness of Parent- or Caregiver-Reported Child Maltreatment Instruments to Parenting Interventions. *Trauma, Violence, & Abuse.* Manuscript submitted for publication.

Figure 1.1. Overview of Articles and Research Questions

1.3 Outline of the Thesis

This thesis consists of two main parts: an extended abstract (Part 1) and three articles (Part 2). Part 1 provides the background information, theoretical framework, relevant literature, research methodology, and a discussion of the main results to ensure the internal coherence of the submitted articles throughout the thesis. Part 2 comprises the three submitted and/or published articles.

The extended abstract comprises six chapters. This introductory chapter (*Chapter 1*) provides a brief background and rationale for the research topic, aims, and research questions. *Chapter 2* reviews the relevant literature associated with a detailed description of child maltreatment, two approaches to prevent CM effectively, a need to measure both parental behaviours and attitudes on CM, a reason why parent or caregiver reports on CM are significant, and a current gap in evaluation of psychometric properties of CM instruments. Based on the relevant literature, the second chapter aims to address the concepts of CM and the reasons why the research topic for this thesis is important. *Chapter* 3 describes the theoretical frameworks of the taxonomy on psychometric properties and the social ecological model for measuring CM, which are applied to discuss the results and implications of this thesis at the end of this extended abstract. *Chapter 4* presents an overview of the research methods used for this thesis, including a systematic review and the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology. The aim of the fourth chapter is to explain why the methods were appropriate and how the COSMIN method was applied to collect and analyse the data presented in this thesis. *Chapter 5* provides a summary of the main results of the three articles presented in Part 2 and the recommendation of the most suitable CM instruments based on these results. Finally, *Chapter 6* discusses the results of the three articles (i.e., the characteristics and the psychometric properties of the identified CM instruments) and offers recommendations of CM instruments in relation to the overarching aims and research questions of this thesis. The methodological challenges, limitations, and implications for future research and practice are also discussed in detail.

2 Review of Relevant Literature

This review of relevant literature is divided into five subchapters. The first subchapter (2.1) discusses child maltreatment (CM), including its definition, prevalence, and consequences. Subchapter aims to address what CM is, why the perpetration of CM by parents or caregivers is a key construct of interest, and how serious the consequences of CM are. The second subchapter (2.2) describes how to prevent CM effectively through public health approaches and what should be considered to monitor the prevention of CM accurately and reliably through an evidence-based assessment approach. Next, Subchapter 2.3 emphasises the need to measure both maltreating behaviours and attitudes towards CM to investigate the current state of CM and prevent future CM. Subchapter 2.4 describes why parent or caregiver reports of CM are more important than other informant reports of CM. This review chapter concludes by outlining the current research gaps (Subchapter 2.5) in systematic literature reviews that evaluate the psychometric properties of instruments to measure CM.

2.1 Definition, Prevalence, and Consequences of Child Maltreatment (CM)

This subchapter begins by discussing the definition of CM (2.1.1) and then describes the prevalence of CM (2.1.2). The first two sections (2.1.1 and 2.1.2) explain why this thesis considers CM perpetrated by parents or caregivers as a construct of interest. In addition, Section 2.1.3 presents the consequences of CM and how it can influence the health of victimised children and even the next generation.

2.1.1 Definition

There is no international universally acknowledged definition of CM due to intercultural differences in what exactly is considered harmful treatment of children in parenting practices (Parsons et al., 2020). For instance, several countries, such as Sweden, Croatia, and the United Kingdom, have clearly outlawed all types of corporal punishment of children, while the United States (U.S.) legally allows disciplinary spanking of children (Ripoll-Núñez & Rohner, 2006). Most U.S. parents spank their children at least once before the children reach school age, because the parents believe that spanking can be helpful in disciplining their children without actually harming the children (Gershoff, 2013). Even within the U.S., most states (42/50) include ‘threatened harm’ or ‘risk of harm’ in their definition of physical abuse, while the other 8 states limit their definition to actual harm (Child Welfare Information Gateway, 2019). This discrepancy in the definition of CM strongly affects accurately estimating the number of victims of CM (Parsons et al., 2020).

Despite cultural variations in operationalising CM, partial consensus on the definition of CM has been reached (Cicchetti & Toth, 2005). The first consensus is that CM can be divided into two broad subcategories: abuse (acts of commission) and neglect (actions of omission) (Barnett et al., 1993). Another consensus is that child abuse and neglect are more frequently perpetrated by parents or caregivers than by peers or strangers. In line with these two consensuses on the definition of CM, the U.S. Child Abuse and Prevention Treatment Act (CAPTA) defines CM as, *“Any recent act or failure to act on the part of a parent or caretaker, which results in death, serious physical or emotional harm, sexual abuse or exploitation, or an act or failure to act which presents an imminent risk of serious harm”* (USDHHS, 2018, p. 15). Compared with the CAPTA focusing on only current harm related to CM, the WHO (1999) more broadly defines CM as, *“All forms of physical and/or emotional ill-treatment, sexual abuse, neglect or negligent treatment or commercial or other exploitation, resulting in actual or potential harm to the child´s health, survival, development or dignity in the context of relationship of responsibility, trust or power”* (p. 15).

Furthermore, the WHO (2006) distinguishes between four CM subtypes: physical abuse, emotional abuse, sexual abuse, and neglect (see Table 2.1). As this classification is by far the most common taxonomy of CM (Barnett et al., 1993; Cicchetti & Toth, 2005), these four subtypes of CM were used in the present thesis.

Table 2.1. Definitions of the Subtypes of CM adapted from WHO (2006)

Subtype	Definition
Physical Abuse	Physical abuse is defined as the intentional use of physical force against a child that results in—or has a high likelihood of resulting in—harm to the child's health, survival, development, or dignity. This type of abuse includes hitting, beating, kicking, shaking, biting, strangling, scalding, burning, poisoning, and suffocating. Much physical violence against children in the home is inflicted with the object of punishing.
Emotional Abuse	Emotional abuse involves both isolated incidents as well as a pattern of failure over time on the part of a parent or caregiver to provide a developmentally appropriate and supportive environment to a child. Acts in this category may have a high probability of damaging the child's physical or mental health or the child's physical, mental, spiritual, moral, or social development. Abuse of this type includes the following: the restriction of movement; patterns of belittling, blaming, threatening, frightening, discriminating against, or ridiculing; and other non-physical forms of rejection or hostile treatment.
Sexual Abuse	Sexual abuse is defined as the involvement of a child in sexual activity that the child does not fully comprehend; that the child is unable to give informed consent to; for which the child is not developmentally prepared; or that violates the laws or social taboos of society. Children can be sexually abused by both adults and other children who are—by virtue of their age or stage of development—in a position of responsibility, trust, or power over the victim.
Neglect	Neglect includes both isolated incidents as well as a pattern of failure over time on the part of a parent or other family member to provide for the development and well-being of a child—where the parent is in a position to do so—in one or more of the following areas: health, education, emotional development, nutrition, shelter, and safe living conditions. The parents of neglected child are not necessarily poor; they may equally be financially well-off.

2.1.2 Prevalence

The global prevalence of CM has been estimated to be 57.6% of all children worldwide, and most victims of CM are exposed to more than one type of CM (Hillis et al., 2016). To estimate the prevalence of CM subtypes, a recent meta-analysis combined the results of several meta-analyses on the global CM prevalence (Stoltenborgh et al., 2015). Stoltenborgh et al. (2015) found that emotional abuse was most common, accounting for 36.3% of CM incidents; the next most common was neglect, accounting for 34.7% of incidents, followed by physical abuse at 22.6% and sexual abuse at 12.7%. However, estimates of the prevalence of CM vary between studies and across countries (Hillis et al., 2016; Stoltenborgh et al., 2015) due to the use of different methods and questions to measure CM (Janson, 2018). For instance, the question "Have you ever been sexually abused?" will yield fewer "Yes" responses than specifically worded questions about acts of sexual abuse, such as sexual penetration, fondling of the genitals, and involvement of a child in an act of masturbation (Stoltenborgh et al., 2011).

The global prevalence of CM victimisation also varies between different age groups. Across the globe, studies on CM consistently report that the CM victimisation rate in young children is higher than that in adolescents (Bae & Kindler, 2017; Euser et al., 2010; Kim et al., 2019). For instance, a study of CM prevalence in the Netherlands found that the risk of CM victimisation was greatest for children aged 0 to 3 years (Euser et al., 2010). In the U.S., similar trends were found with approximately 30% of the victims reported to the Child Protection Services (CPS) in 2018 being under 3 years of age (USDHHS, 2020); a national annual report on CM in 2020, confirmed that very young children faced the highest risk of CM victimisation, with the CM rate decreasing with the child's age (USDHHS, 2020). In addition, the type of CM to which children are most vulnerable varies depending on the child’s age. For example, physical abuse is most prevalent among young children, while sexual abuse is most common among adolescents (WHO, 2002). Last, a child’s disability is another significant risk factor for CM victimisation. Children with disabilities are three to four times more likely to experience CM than their peers without disabilities worldwide according to a meta-analysis of the prevalence of CM against children with disabilities (Jones et al., 2012).

Although parents or caregivers perpetrate CM most frequently (Devries et al., 2018; Sedlak et al., 2010), the relationship with perpetrators differs depending on the CM subtype. The most common perpetrators of sexual abuse are non-family members (Finkelhor et al., 2014). However, for the other three types of CM (physical abuse, emotional abuse, and neglect), more than half of the perpetrators are parents or caregivers (Devries et al., 2018). For example, in the U.S., parents are the perpetrators of 92% of all cases of neglect, 73% of emotional abuse, and 72% of physical abuse, but only 37% of sexual abuse (Sedlak et al., 2010). Therefore, CM perpetrated by parents or caregivers should be considered a key construct of interest.

2.1.3 Consequences

Early exposure to CM is linked to a number of undesirable and severe outcomes, hampering children’s social, psychological, and physiological functioning. The impact of CM is often lifelong and severe, and it is fatal for some children (Gilbert, Kemp, et al., 2009). Exposure to CM leads to a higher risk of developing mental disorders, lifestyle-related diseases (e.g., liver, heart, and lung diseases), risky sexual behaviour, substance abuse (e.g., drug and alcohol abuse), and even suicide attempts (Felitti et al., 1998; Gilbert, Widom, et al., 2009; Leitzke & Pollak, 2017; Norman et al., 2012; Thornberry & Henry, 2013). In addition, persistent exposure to CM is linked to criminal, violent, and delinquent acts during adolescence (Gilbert, Widom, et al., 2009; Ireland et al., 2002). Last, early exposure to CM increases the risk of negative academic outcomes (Ryan et al., 2018). Children exposed to CM are more likely to have lower grade point averages and lower school attendance rates, as well as to experience grade retention, suspension, expulsion, and dropping out of school (Fry et al., 2018; Tessier et al., 2018). The negative academic outcomes are not limited to the primary or secondary school years, but have long-term impacts on lower entrance rates to university and lower socioeconomic status of young adults exposed to CM during childhood (Ryan et al., 2018).

Furthermore, the negative consequences of CM can influence the next generation (i.e., the intergenerational transmission of CM effects). In particular, victims of childhood maltreatment are more likely to abuse or neglect their own children (Leitzke & Pollak, 2017; Thornberry & Henry, 2013). Some longitudinal studies have shown that fewer than a quarter of CM victims grow up to be resilient adult survivors who can perform well in all major daily tasks, despite their childhood traumatic experience (Banyard & Williams, 2007; Ben-David & Jonson-Reid, 2017; McGloin & Widom, 2001). In addition, recent CM studies have indicated that traumatic symptoms related to childhood maltreatment may be passed from one generation to the next because certain neurogenetic variants that are caused by traumatic memories of maltreatment may be inherited by offspring (Buss et al., 2017; Yehuda & Lehrner, 2018). In particular, a strong genetic connection has been observed between a maternal history of CM victimisation and their child’s mental health problems such as suicide attempts, anxiety, depression, and maladaptive behaviour over time (Brent et al., 2004; Brodsky et al., 2008; Collishaw et al., 2007; Plant et al., 2013; Roberts et al., 2013). Even when a victimised mother has never maltreated her child, the child may experience a higher risk of mental disorders due to the intergenerational transmission of the mother’s traumatic memory (Plant et al., 2013; Rijlaarsdam et al., 2014).

2.2 Prevention of CM

This subchapter describes public health approaches (Section 2.2.1) to preventing CM effectively at the population level by comparing it with the current CPS approach, which provides its service only for targeted caregivers or parents at risk. In addition, an evidence-based approach (Section 2.2.2) to measuring CM is suggested for accurately and reliably monitoring CM prevention.

2.2.1 Public Health Approaches to Preventing CM

Contemporary approaches to CPS predominantly involve investigation and intervention after CM has occurred (Scott et al., 2016). However, when data are obtained only from children who are officially reported as CM victims after maltreatment occurs, it can result in the substantial underestimation of the prevalence of CM due to the data's limited scope (Putnam-Hornstein et al., 2011). Furthermore, a current statutory intervention, focusing more on punishment than support and targeting only parents suspected of perpetrating CM, can unnecessarily stigmatise parents receiving the intervention services to improve their parenting practices; hence, the intervention can make them reluctant to seek such services (O'Donnell et al., 2008). The statutory intervention can also make it difficult to support non-suspected parents who voluntarily request assistance in changing their discipline style to one that is more positive and less harsh/punitive (O'Donnell et al., 2008).

To overcome the challenges faced by the current CPS system, the WHO (2005) recommended that each member country implement public health approaches to CM that focus on preventive measures at the population level (O'Donnell et al., 2008; Putnam-Hornstein et al., 2011; Scott et al., 2016). The public health approaches can be conceptualised as a four-step process (see Figure 2.1) according to Putnam-Hornstein et al. (2011) and the WHO (2005):

- Step 1: *Define the problem* through data collection for surveillance.
- Step 2: *Uncover the possible causes* of CM through the identification of risk and protective factors.
- Step 3: *Develop and test interventions* through efficacy and effectiveness research.
- Step 4: *Implement the most effective intervention* through the dissemination and monitoring of interventions.

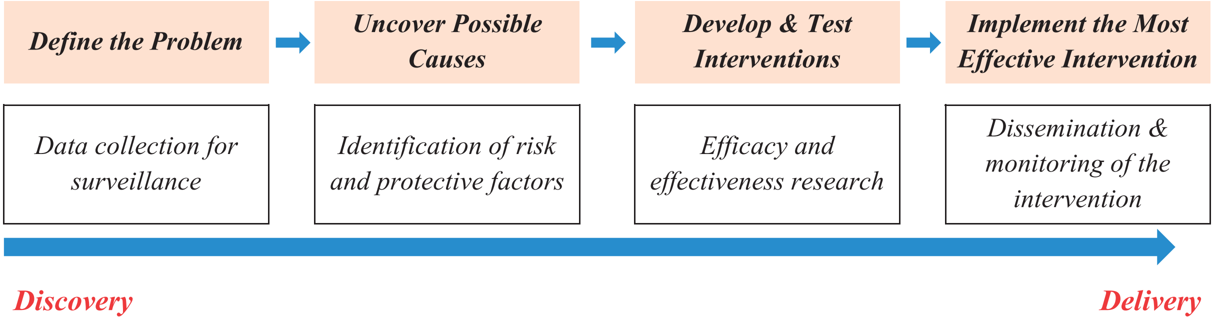

Define the Problem
Uncover Possible Causes
Develop & Test Interventions
Implement the Most Effective Intervention
Data collection for surveillance
Identification of risk and protective factors
Efficacy and effectiveness research
Dissemination & monitoring of the intervention
Discovery
Delivery

Figure 2.1. *Public Health Framework adapted from Putnam-Hornstein et al. (2011) and WHO (2005)*

To define the problem of CM (Step 1), data collection for surveillance of CM should be conducted first, with the aim of collecting data to estimate the prevalence of CM at the population level (Putnam-Hornstein et al., 2011). A precise estimate of the prevalence of CM can help detect the scope and magnitude of the health threat related to CM at the population level (Thacker & Berkelman, 1988). Step 2 involves the identification of risk factors placing individual children at risk of CM and protective factors serving to protect the children from CM. Next, based on the information of CM prevalence as well as risk and protective factors of CM, Step 3 involves the development and testing of intervention strategies to prevent and reduce CM (Diez-Roux, 2000). Even though public health approaches focus on the health of the entire population, interventions may target different segments of the population, such as primary interventions focused on the general population, secondary interventions focused on targeted populations at risk for CM, and tertiary interventions focused on victim children or perpetrator parents in CM (Putnam-Hornstein et al., 2011). Finally, Step 4 involves implementing effective interventions at the community level (Peden et al., 2008). Dissemination is an essential element of this step, while continued surveillance is also needed over time (Peden et al., 2008). Within the public health approaches, the cycle then returns the surveillance of the full population for the wide adoption of the most effective interventions to monitor its effectiveness (Putnam-Hornstein et al., 2011).

However, there are critical concerns about data collection for surveillance in the first step of public health approaches to CM prevention. Although high-quality data are needed for CM prevention within public health approaches, almost half of all countries in the world have failed to report robust prevalence estimates of CM (Hillis et al., 2016). This failure to accurately estimate the prevalence has occurred because survey questionnaires frequently contain irrelevant questions or have incomplete coverage of the construct of interest (i.e., poor content validity; Mathews et al., 2020). Accordingly, prevalence estimates are often inadequately specified and underestimate the actual frequency of CM (Mathews et al., 2020). In addition, the use of non-standardised instruments is common (Moore et al., 2015), which carries an increased risk of failing to capture experiences of CM and of capturing experiences not involving CM, which produces unreliable estimates of CM prevalence (Mathews et al., 2020).

In addition, research on parenting interventions to reduce CM in the third and fourth steps of public health approaches, has been hampered by the lack of consensus on which CM instrument is the most responsive or sensitive in detecting treatment effects following interventions for reducing CM by parents (Fluke et al., 2020). Many CM effectiveness studies have used parental questionnaires to measure the current state or prevalence of CM. However, these questionnaires may be less sensitive to capturing changes over time in CM occurring both before and after parenting interventions aimed at preventing CM (Cluver et al., 2016). Therefore, selecting and using high-quality CM instruments that are sensitive enough to measure change over time in response to a parenting intervention is essential to monitoring CM prevention efforts accurately.

2.2.2 Evidence-Based Assessment Approach to Monitoring CM Prevention

An evidence-based assessment approach to monitoring CM prevention refers to an approach to clinical evaluation in which CM practitioners actively use research evidence to guide the selection of CM instruments for assessing the effectiveness of an intervention to prevent CM (Hunsley & Mash, 2007). If CM instruments fail to accurately estimate the scope and magnitude of the current state of CM, practitioners may provide abusive and neglectful parents with ineffective or inappropriate interventions to reduce their CM, placing them at risk of further perpetration of CM (Mash & Hunsley, 2005). Thus, to determine whether an intervention is effective, the effects of an intervention on CM should be evaluated by using robust, evidence-based instruments to measure CM.

For the evidence-based assessment of CM, the emphasis shifts from the selection of empirically supported CM interventions to the selection of appropriate CM instruments (Achenbach, 2017). Such selection requires researchers and practitioners to consider the following three factors (Hunsley & Mash, 2007): (1) development of the relevant CM constructs (i.e., content validity) based on theoretical and empirical research; (2) good psychometric properties (other than content validity) of CM instruments; and (3) appropriate assessment processes for CM instruments in terms of the administration time, cost, and interpretation of instrument scores. Compared with other psychometric properties and assessment processes, the content validity in the development of the relevant constructs of CM is the most important factor to consider for establishing an evidence-based assessment of CM. Constructs or items to be measured in a CM instrument should be derived from relevant theories (e.g., a theoretical model related to the constructs) or empirical study results on CM for instrument development (e.g., questionnaires or interviews with professionals or parents). If the content validity of a CM instrument is poor, then the evaluation of its other psychometric properties and assessment processes are meaningless (Patrick et al., 2011; Prinsen et al., 2018; Streiner et al., 2015). For example, if a CM instrument includes irrelevant items such as items related to parental stress, one may measure an incorrect or incomplete construct of CM very reliably (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018), while failing to assess the targeted construct. Furthermore, in terms of responsiveness, an actual change in the CM construct may be overestimated or underestimated because of irrelevant or missing CM concepts (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Above all, parents (i.e., respondents) might be frustrated when questions about CM are irrelevant to them are asked or when important questions about CM are not asked, which can result in biased responses or low response rates (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). However, which constructs are relevant for measuring CM in interventions to prevent parental CM is still a matter of debate (Mathews et al., 2020; Meinck et al., 2018; Meinck et al., 2016).

2.3 Measurement of CM

This subchapter describes the need to measure maltreating behaviours directly (Section 2.2.1) using criteria with specific target behaviours to avoid underestimating the current prevalence of CM. In addition, the need to measure indirect attitudes towards CM (Section 2.2.2) is also explained. Both direct behaviour and indirect attitudes concerning CM are constructs of interest in this review.

2.3.1 Measuring Direct Maltreating Behaviours

The currently available CM prevalence estimates underestimate CM (Al-Eissa et al., 2015). To address the underestimation issue, items of CM instruments used to estimate prevalence must reflect observable specific behaviours instead of only abstract or unobservable concepts such as sexual abuse (Fisher, 2008). Notably, the use of CM instruments with nonspecific or unobservable items in CM prevalence studies may result in the underestimation of CM prevalence (Finkelhor et al., 2007; Hamby et al., 2010; Hillis et al., 2016; Sumner et al., 2015). In contrast, the use of behaviourally specific questions in CM instruments is more likely to help CM victims or perpetrators recall what they experienced or what actions they took, which can result in a more accurate estimation of CM prevalence than the use of non-specific questions (Fisher, 2009). In addition, as children may have been victimised through multiple types of CM simultaneously, one or more types of CM need to be considered when measuring its prevalence (Finkelhor et al., 2007; Gilbert, Widom, et al., 2009; Hughes et al., 2017). Furthermore, to provide useful and nuanced information for the prevention of CM, instruments need to ask about CM frequency (how often maltreating behaviours have occurred), severity (how serious the maltreating behaviour was), and timing (when the maltreating behaviour occurred; Manly, 2005). These factors impact health outcomes, and the measurement of these factors offers necessary information on the risks and protective factors for the prevention of CM. Although rigorously measuring specific maltreating behaviours is quite complex, it is very important to plan, implement, and monitor prevention based on precise data or evidence on CM (Anda et al., 2010; Hillis et al., 2016).

Specific criteria to assess direct maltreating behaviours are suggested in the Diagnostic and Statistical Manual-5th revision (DSM-5; American Psychiatric Association, 2013) and the International Classification of Diseases-11th revision (ICD-11; WHO, 2018), which are the most commonly used health-related classifications. Both the DSM-5 and ICD-11 were developed to support screening and identifying health-related problems by clinicians and researchers; they also include a list of criteria to define the subtypes of CM (i.e., physical abuse, emotional abuse, sexual abuse, and neglect). Within each subtype, a threshold is defined to distinguish between suboptimal but non-abusive parenting versus CM (examples are provided in Table 2.2).

Table 2.2. *DSM-5 and ICD-11 Criteria for CM adapted from Slep et al. (2015)*

Subtype	Health Classification	Behaviour Criteria
Physical Abuse	DSM-5	Beating or punching a child; biting or kicking; throwing or shaking; stabbing; hitting (with a hand, with a strap, a stick, or another object); choking; burning
	ICD-11	Suspected or confirmed intentional act of physical force, such as slapping or hitting a child
Emotional Abuse	DSM-5	Humiliating, disparaging, or berating a child; harming/abandoning things or people who are important to the child or threatening the child; threatening future abandonment, harm or confinement of the child (e.g., tying the child to a piece of furniture or another object, tying the child's arms or legs together, confining the child to a tight space [e.g., a closet]); scapegoating the child egregiously; excessively disciplining the child in a physical or non-physical way (e.g., for an extremely long duration or frequency but without the disciplining being considered physical abuse); coercing the child to inflict pain on themselves
	ICD-11	Engaging in suspected or confirmed symbolic or verbal acts that may cause a child psychological harm, such as humiliating, degrading, disparaging, or berating the child; threatening the child with future harm, sexual assault, or abandonment, harming/abandoning the child or indicating that the parent/care provider will inflict harm on or abandon things or people who the child cares about, such as loved ones, pets, or objects (including exposing the child to subthreshold or criteria-meeting partner maltreatment); confining the child (e.g., confining the child in a tight space [e.g., a closet]); tying the child to a piece of furniture or another object; tying the child's arms or legs together; scapegoating the child (blaming the child for something for which the child could not possibly bear responsibility); pressuring the child to inflict pain on the child himself or herself; excessively disciplining the child through physical or non-physical means (e.g., for an extremely long duration or frequency but without the disciplining being considered physical abuse); intentionally indoctrinating the child to make him or her believe a parent is evil, dangerous, or not worthy of the child's love and trust
Sexual Abuse	DSM-5	Rape or fondling of the genitals; incest, penetration; sodomy; indecent exposure; exploitation that does not involve contact (e.g., pressuring, forcing, coercing, or tricking a child to take part in acts of a sexual nature [for others' gratification])
	ICD-11	Actual or attempted anal or vaginal penetration or another physical contact between a child and an adult of a sexual nature; oral-anal or oral-genital contact; fondling through the clothing or directly on the skin Noncontact exploitation, such as pressuring, forcing, coercing, or tricking the child to take part in acts of a sexual nature for another person's gratification without there being physical contact directly between the victim and the offender, such as exposing the child's breasts, anus, or genitals; making the child masturbate or watch someone else masturbate; making the child participate in sexual acts with someone else (including child prostitution); making the child perform in a sexual way, pose, or undress (including child pornography)
Neglect	DSM-5	Failure to provide a child with the education needed; abandonment of the child; absence of appropriate supervision; failure to take care of basic emotional or psychological needs; failure to provide the necessary clothing, shelter, and/or nourishment, failure to provide necessary medical care
	ICD-11	At least one suspected or confirmed egregious omission or act by a child's care provider depriving the child of the age-appropriate care the child needs, such as a lack of appropriate supervision; abandonment; exposure to physical hazards; a lack of necessary healthcare, education, clothing, shelter, or nourishment

Note. The behaviour criteria were paraphrased from "Child maltreatment in DSM-5 and ICD-11", by A. M. Slep et al., 2015, *Family Process*, 54(1), pp. 20—23 (https://doi.org/10.1111/famp.12131). Copyright 2015 by the Family Process Institute. The licence agreement between Sangwon Yoon and John Wiley and Sons for reuse of the content of Slep et al. (2015) in this thesis was obtained from Copyright Clearance Center on the 10th of August in 2021.

Child neglect is more difficult to assess than child abuse as neglect involves omissions or failure to act. It is much more difficult to report something one has not done (i.e., acts of omission) in particular circumstances than what one has done (i.e., acts of commission; Slep et al., 2015). For this reason, developing a well-operationalised parent-report instrument to measure neglect has been challenging (Slep et al., 2015). Therefore, measuring parental attitudes towards neglect is recommended as a way of assessing neglect, rather than measuring their neglectful behaviours directly.

2.3.2 Measuring Indirect Attitudes Towards CM

To prevent CM, measuring parental attitudes towards CM, including a parent's values, beliefs, or feelings related to abusive and neglectful parenting behaviour towards a child, is also important (Altmann, 2008; Holden & Buck, 2002). Parental maltreating behaviours and attitudes towards such behaviours are strongly correlated: i.e., parental attitudes towards CM drive parental maltreating behaviours. This association between parents' attitudes and actual maltreating behaviours has been supported by empirical research (Jabraeili et al., 2015). For example, Ashton (2001) and Jackson et al. (1999) examined and noted the relationship between attitudes and behaviours. In addition, Vittrup et al. (2006) provided evidence of a significant relationship between maternal attitudes towards corporal punishment and their actual use of corporal punishment. Mothers who have positive attitudes towards corporal punishment often use this kind of punishment to discipline their children (Vittrup et al., 2006).

Social information processing theory is one of the leading theoretical models that has been applied to understand the relationship between parental maltreating behaviour and parental attitudes towards CM (Del Vecchio et al., 2012; Milner, 2000). That is, parents have pre-existing attitudes towards parenting behaviour before any concrete situation in which they might discipline their child (Milner, 2000). Then, when a parent with more accepting attitudes towards physical disciplines is confronted with a potential disciplinary decision, the following four stages may occur (Rodriguez et al., 2019). Initially, parents may misperceive the situation (Stage 1) and they may form biased, negative appraisals and expectations regarding their child's behaviour (Stage 2). Parents may then fail to integrate all relevant information before engaging in the physical discipline of their child, including considering their non-physical disciplinary options (Stage 3). Once parents begin administering physical discipline, they may fail to adequately monitor its intensity, escalating towards physical abuse (Stage 4).

As stated above, attitudes are an empirically and theoretically important factor in predicting and controlling behaviour. Therefore, measuring parental attitudes towards maltreating behaviours has great importance in preventing CM.

2.4 Parent or Caregiver CM Reports

This subchapter introduces the characteristics of three different informant reports of CM: professional, child, and parent reports (Section 2.2.1). Furthermore, comparisons of professional, child, and parent reports of CM are discussed to justify why this thesis focused on parent or caregiver CM reports instead of professional or child CM reports.

2.4.1 Three Types of Informants Reporting CM: Professionals, Children, and Parents

The main informants who report CM are professionals, children, and parents/caregivers (Cooley & Jackson, 2020). Professionals reporting CM include child protection workers, psychologists, health professionals, or teachers, who provide services for children. Professional CM reports can capture only alleged cases of maltreatment reported to CPS agencies (Huffhines et al., 2016). In many countries, when health professionals or teachers suspect that children or students are being maltreated, they are legally obliged to report any suspected cases of CM by either calling a hotline or completing a CM screening questionnaire for referral to CPS (Greco et al., 2017; Pelletier & Knox, 2017). Furthermore, child protection workers can report CM through direct observation of the parenting behaviours of caregivers who are referred to CPS (Cañas et al., 2020). These observational instruments are substantially more complex, costly, and time-consuming to administer than phone calls and questionnaires (Morsbach & Prinz, 2006).

Child reports of CM are obtained by asking individual children to identify their experiences of exposure to CM. However, compared with adults, young children often struggle with understanding what is being asked of them, remembering what they experienced, and verbalising what they remember (Lamb et al., 2007; Meinck et al., 2016; WHO, 2006). Furthermore, parents or caregivers, who are responsible for the child's welfare (McDonald, 2007), are also important informants to report their own maltreating behaviours of their children. For both child and parent or caregiver reports of CM, the most common method for measuring CM is through the use of self-administered questionnaires, although some studies have used interviews instead (Laurin et al., 2018; Moody et al., 2018). Self-administered questionnaires allow respondents to answer questions privately, instead of directly discussing their responses with a researcher. This method is useful because informants are more likely to disclose their experiences of victimisation or perpetration related to CM when asked in this manner than when asked similar questions in an interview (Meinck et al., 2018).

2.4.2 Comparison of Professional, Child, and Parent Reports of CM

A meta-analysis comparing CM prevalence rates among professional reports and child/caregiver self-administered reports (Stoltenborgh et al., 2015) found a tendency towards a lower prevalence of CM in professional reports than in either child or caregiver reports. This may be the result of professionals tending to report only more serious CM cases, since they may not consider mild cases to be significant enough to report (Negriff et al., 2016). For example, one study found that although 74% of schoolteachers had suspected more than one case of CM victimisation during their careers, only 27% had actually reported suspected cases to CPS agencies. This is because the teachers feared that reporting CM based on only their suspicions without clear evidence may have negatively affected the children’s lives (Greco et al., 2017). Another study found that approximately half of all medical doctors also felt uncomfortable discussing topics related to maltreatment with victimised children or their parents, making the doctors hesitant to report mild cases of CM (Foster et al., 2017). In addition, given that most CM occurs in private homes with no witnesses other than the victimised children or their caregivers (Institute of Medicine and National Research Council, 2014), professional-reported prevalence rates of CM likely represent only a fraction of CM cases, especially compared with child- or caregiver-reported CM (Fallon et al., 2010).

Another meta-analysis found that the prevalence estimates for most types of CM reported by caregivers were markedly higher than those reported by children, with the notable exception of sexual abuse (Devries et al., 2018). The underestimation of sexual abuse in caregiver reports might occur because perpetrators of sexual abuse mostly tend to be peers or adults other than the child’s parents or caregivers; most victims of sexual abuse are adolescents who tend to disclose their experience of exposure to sexual abuse to their caregivers (Devries et al., 2018). Conversely, the underreport of the three other types of CM (physical abuse, emotional abuse, and neglect) in child reports could be because most victims of CM are younger children (Euser et al., 2010; Kim et al., 2019), who may have more trouble recalling and disclosing their experiences of victimisation of CM than adult caregivers (Devries et al., 2018). Therefore, adult caregiver-report CM instruments are more likely to accurately estimate the prevalence of CM.

The precision and reliability of caregiver-report CM instruments, however, are still controversial because caregivers are most likely to respond in socially desirable ways (Compier-de Block et al., 2017). First, parents may not report their actual maltreating behaviours towards their children due to concerns about the legal consequences (Compier-de Block et al., 2017). Parents may be concerned that their child will be removed from their home or that they will be arrested for such abuse. Second, when parents feel either that their parenting is being questioned or that they are being accused of maltreatment, they may feel ashamed or guilty about their actions and deny any wrongdoing (Gibson, 2015). Both concerns may result in parents giving socially desirable responses rather than accurate descriptions of their actions.

Parent- or caregiver-report CM instruments are subject to social desirability bias (Compier-de Block et al., 2017), yet they are more feasible to administer than child-report CM instruments (Meinck et al., 2016). Children under nine years of age are the main victims of CM (e.g., in the U.S., more than two-thirds of CM victims are children under nine years of age; USDHHS, 2021); however, they may not understand the items and may not respond accurately to the items about their experience (Lamb et al., 2007; World Health Organization, 2006), making child-report CM instruments inappropriate for that age group (Meinck et al., 2016). In addition, it may be more difficult to obtain consent for administering child-report CM instruments than adult parent- or caregiver-report CM instruments. For these practical and ethical reasons, parent-report CM instruments are more easily administered, which can facilitate large-scale studies and survey research involving multiple follow-ups (Pallant et al., 2014; Wittkowski et al., 2020). Furthermore, in clinical practice, valid and reliable instruments that are easy to administer can facilitate both the screening of maltreating parents or caregivers and the detection of changes in their maltreating behaviours after interventions aimed at reducing CM (Brockington et al., 2001; Wittkowski et al., 2020). Due to their feasibility, parent- or caregiver-report CM instruments have been used most frequently to investigate and prevent CM in research and clinics, especially for young children (Meinck et al., 2016). Importantly, for optimal use in clinical practice and research, parent- or caregiver-report CM instruments should have robust validity and reliability (Streiner et al., 2015; Wittkowski et al., 2020). Hence, identifying parent- or caregiver-report instruments with good psychometric properties is essential for accurate estimation of CM prevalence and sensitive detection of CM intervention effects.

2.5 Current Gap in the Literature

For the selection of suitable instruments, either a systematic review evaluating the psychometric properties of existing CM instruments should be conducted or a relevant previously conducted review should be consulted (Scholtes et al., 2011). To date, only one systematic review has evaluated the psychometric properties of instruments assessing CM (Saini et al., 2019). However, the authors of the review identified mostly clinician interview instruments and child self-reports, which are more likely to underreport the actual occurrence of CM than caregiver-report instruments (Devries et al., 2018), and only one caregiver proxy-report instrument (i.e., asking caregivers about their child's experience of CM perpetrated by any adults, but not about their own perpetration of CM; Saini et al., 2019; Sprangers & Aaronson, 1992). None of the instruments and studies identified in the review by Saini et al. (2019) overlapped with this thesis on parent- or caregiver-report CM instruments. No other systematic reviews on the psychometric properties of parent- or caregiver-report CM instruments have been published to date. Therefore, to fill the current gap in the literature, this thesis systematically reviewed the psychometric properties of parent- or caregiver-report instruments measuring CM perpetrated by parents.

3 Theoretical Framework

This theoretical framework is divided into two subchapters. Subchapter 3.1 describes the application of the social ecological model to measure CM, which provides a framework for discussing the position of the included CM instruments within the model and the implications of the CM instruments for future practice in Chapter 6. Subchapter 3.2 presents a taxonomy of psychometric properties, which is a conceptual framework related to the terms and definitions of psychometric properties used throughout this thesis.

3.1 Social Ecological Model for Measuring CM

The social ecological model can be used as a theoretical framework to describe how individual children's experiences of CM are influenced by the various systems of society (Gershoff, 2013), such as the children themselves, their families (parents or caregivers), professionals (health professionals, child protection workers, or teachers), governments, and society or culture. That is, the model explains how these systems reciprocally influence the CM experiences of an individual child (see Figure 3.1).

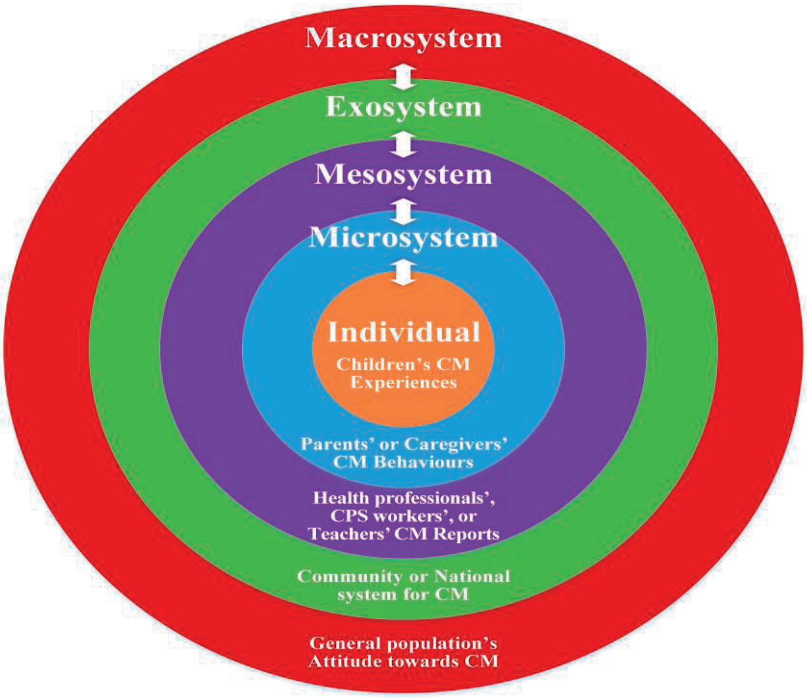

Macrosystem
Exosystem
Mesosystem
Microsystem
Individual
Children's CM Experiences
Parents' or Caregivers' CM Behaviours
Health professionals', CPS workers', or Teachers' CM Reports
Community or National system for CM
General population's Attitude towards CM

Figure 3.1. Social Ecological Model for Measuring CM adapted from Bronfenbrenner (1992)

Individual children are located at the centre of the model and are surrounded by various systems related to CM. Children's CM experiences are influenced directly and indirectly across the four levels of systems (Belsky, 1993; Bronfenbrenner, 1992). First, the microsystem refers to face-to-face influences on individual children's CM experiences, such as parents' or caregivers' maltreating behaviours towards their children or their attitudes towards CM (Belsky, 1993; Bronfenbrenner, 1992). Second, the mesosystem refers to the interrelations among the various agents who are involved in reporting and intervening in CM, such as health professionals, child protection workers, and teachers (Belsky, 1993; Bronfenbrenner, 1992). Third, the exosystem refers to factors within the community or national system related to CM, such as those that monitor the CM prevalence or the effectiveness of CM interventions at the population level (Belsky, 1993; Bronfenbrenner, 1992). Finally, the macrosystem refers to cultural beliefs and values towards CM that influence maltreating behaviours related to CM prevalence, such as the general population's attitudes towards CM in a country (Belsky, 1993; Bronfenbrenner, 1992). As the four systems reciprocally affect the CM experience and even affect one another, the social ecological model suggests that reciprocal relationships exist between individual children's CM experiences and the environmental factors related to those experiences (Belsky, 1993; Bronfenbrenner, 1992).

In addition, the social ecological model may imply that various perspectives on CM at each system level should be measured to understand the true state of CM. That is, multi-informant reports of CM from agents of each system may help compensate for the limitations of individual informant report (Belsky, 1993; Cooley & Jackson, 2020). For example, a paediatrician or teacher (i.e., mesosystem) who sees a child every day can identify and report suspected CM that has been hidden by parents or caregivers. Furthermore, at the population level (i.e., exosystem), questionnaires on the prevalence of CM may allow parents or caregivers to respond regarding their parenting behaviours more honestly (i.e., more free from social desirability bias; Milner & Crouch, 1997), because it is easier to guarantee anonymity at this level than in individual parent reports of CM at the microsystem level.

In summary, the social ecological model provides a meaningful framework for understanding where the CM instruments included in this thesis can be located among the four systems, how these instruments can be applied to culturally different parents or caregivers (due to cultural differences in the macrosystem) within the same system, how the instruments can be applied to the other systems, and how the instruments can be used to connect different systems. The answers to these questions will be further detailed in Subchapter 6.5 *Implications for Future Practice*.

3.2 Taxonomy of Psychometric Properties

'Psychometric properties' are an umbrella term used to refer to validity and reliability, which are often used interchangeably with terms such as 'measurement properties' (Mokkink et al., 2010). Different terminology and definitions have been used throughout the literature to describe psychometric properties. Variation in terminology and definitions for psychometric properties has led to inconsistent reporting in studies on the development and psychometric evaluation of measurement instruments (Mokkink et al., 2010). To overcome the absence of uniform terminology, an international Delphi study was conducted to achieve consensus on the definitions and domains of psychometric properties by the COSMIN group (Mokkink et al., 2010). The COSMIN terminology is used throughout this dissertation.

Figure 3.2 shows the COSMIN taxonomy, including three major domains of psychometric properties: (1) validity, (2) reliability, and (3) responsiveness. As each of the three domains includes one or more psychometric properties, the domains are subdivided into nine psychometric properties: content validity, criterion validity, structural validity, hypothesis testing for construct validity, cross-cultural validity, internal consistency, reliability, measurement error, and responsiveness (Mokkink et al., 2010). The definitions of the nine psychometric properties per domain are presented in Table 3.1.

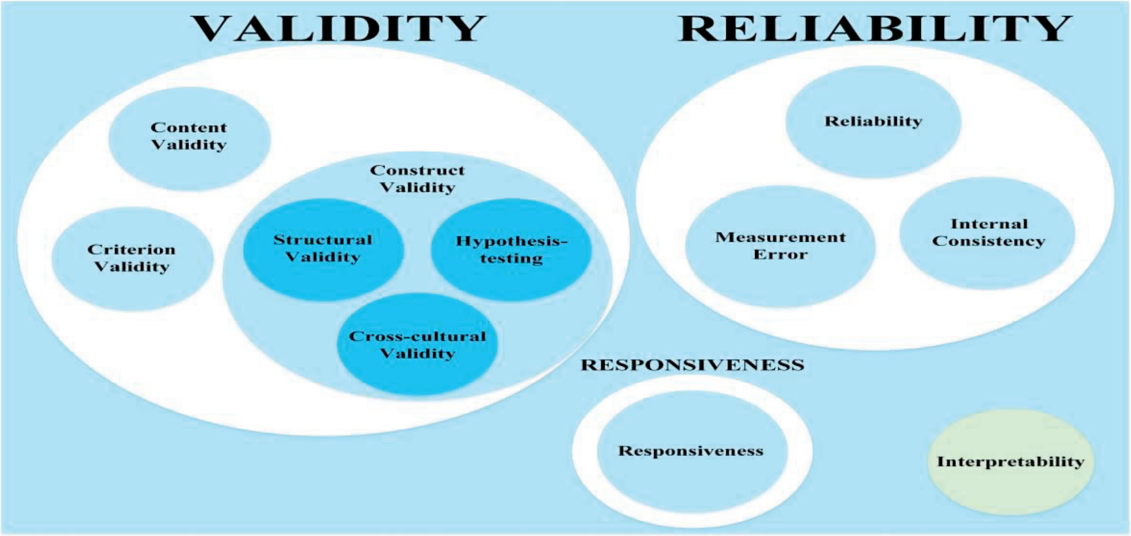


Figure 3.2. Overview of psychometric properties according to the COSMIN taxonomy adapted from Mokkink et al. (2010). *Notes.* Interpretability (i.e., the extent to which clinicians can interpret an instrument's quantitative scores as their qualitative meaning) is not considered a psychometric property. Nonetheless, good interpretability of a score is needed to support the usefulness of an instrument in clinical practice and research (Mokkink et al., 2010).

Table 3.1. *Definitions of Domains and Psychometric Properties for Health-Related Patient-Reported Outcomes adapted from Mokkink et al. (2010)*

Domain	Properties	Definition[a]
Validity		The extent to which an instrument measures the construct(s) it is intended to measure.
	Content validity	The extent to which the content of an instrument adequately reflects the construct being measured.
	Criterion validity	The extent to which the scores of an instrument adequately reflect a "gold standard."
	Structural validity[b]	The extent to which the scores of an instrument adequately reflect the dimensionality of the construct being measured.
	Hypothesis testing[b]	The extent to which the scores of an instrument are consistent with a hypothesis based on the assumption that the instrument validly measures the construct being measured.
	Cross-cultural validity[b]	The extent to which the performance of the items of a translated or culturally adapted instrument adequately reflects the performance of the items of the original instrument.
Reliability		The extent to which the measurement is free from measurement error.
	Internal consistency	The extent to which the items of an instrument are interrelated.
	Reliability	The proportion of the total variance in the measurements that is due to "true" differences among patients.
	Measurement error	The systematic and random error of a patient's score which is not attributed to true changes in the construct being measured.
Responsiveness		The ability of an instrument to detect changes in the measured construct over time.
	Responsiveness	Idem responsiveness.

Notes.

[a] *Applies to health-related patient-reported outcome instruments.*

[b] *Aspects of construct validity (i.e., the degree to which the scores of an instrument are consistent with a hypothesis [e.g., internal associations, associations with scores of other instruments, or differences between relevant groups] based on the assumption that the instrument validly measures the construct being measured) under the domain of validity.*

3.2.1 Validity

Validity is a key psychometric property for any instrument because it determines the true association between the instrument and the construct of interest (de Vet et al., 2011). The validity domain defines the degree to which instruments actually measure the construct that they are supposed to measure (Mokkink et al., 2010). The validity domain contains three psychometric properties (Mokkink et al., 2010): content validity, criterion validity, and construct validity.

Content validity defines the extent to which the items/tasks of an instrument adequately reflect the construct to be measured (Mokkink et al., 2010). Content validity pertains to three aspects of the content of an instrument (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018): (1) relevance (i.e., the extent to which all items of an instrument are relevant for the construct of interest in a targeted population); (2) comprehensiveness (i.e., the extent to which all key concepts of the construct of interest are included in an instrument); and (3) comprehensibility (i.e., the extent to which all items of an instrument can be easily understood by the targeted respondents). Relevance, comprehensiveness, and comprehensibility are the main aspects to be considered in the development phase of an instrument to derive constructs of interest or generate items based on relevant theories or interviews from the target population (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). If an instrument is developed with irrelevant, unimportant, or excessively difficult questions, corresponding to the three aspects of content validity, the other psychometric properties do not require further consideration (Prinsen et al., 2018); hence, content validity is regarded as the most important psychometric property (Prinsen et al., 2016).

Criterion validity defines the extent to which the scores of an instrument adequately reflect a gold standard (Mokkink et al., 2010), which demonstrates the true state of the construct of interest (de Vet et al., 2011). Gold standards seldom exist for self-reported (or self-administered) instruments, which always collect subjective information (de Vet et al., 2011). However, if a researcher wants to develop a new short version of an existing long instrument, the original long version can be considered the gold standard for the shorter version (Mokkink et al., 2010).

Construct validity defines the extent to which the scores of the instrument being studied are consistent with a priori hypotheses on the association with the scores of other instruments that measure the same construct (Mokkink et al., 2010). Construct validity includes three psychometric properties. The first property is structural validity, which defines the degree to which the scores of an instrument can adequately reflect the dimensionality of the construct to be measured (Mokkink et al., 2010). The second is hypothesis testing, based on the idea that hypotheses are formulated and tested regarding the differences in the scores of an instrument between subgroups of the target population (i.e., discriminative validity) and the associations of the scores between two instruments to determine if the instruments measure the same construct of interest (i.e., convergent validity; de Vet et al., 2011). The convergent validity of an instrument does not need to be evaluated if a gold standard of the targeted construct is available and evidence of the criterion validity on the association between an instrument and the gold standard is available (de Vet et al., 2011). The last property is cross-cultural validity, which defines the degree to which the performance of the items of an instrument reflects the performance of the same items when the instrument is either translated into another language or adapted to capture cultural differences among respondents (Mokkink et al., 2010). Cross-cultural validity examines whether the translated instrument shows the expected associations with the related constructs and whether the instrument can discriminate between relevant subgroups of respondents.

3.2.2 Reliability

All instruments that are used in clinical practice and research must be reliable to ensure the accuracy of the scores being measured under different conditions when a person is stable on the construct to be measured (de Vet et al., 2011). As a domain, reliability is defined as the degree to which *". . . scores for people who have not changed are the same for repeated measurement under several conditions (e.g., using different sets of items from the same multi-item measurement instrument [internal consistency], over time [test-retest], by different persons in the same occasion [interrater], or by the same person in different occasions [intrarater])"* (Mokkink et al., 2010, p. 734). Three psychometric properties (internal consistency, reliability, and measurement error) constitute the reliability domain.

Internal consistency defines the degree to which the items of an instrument are interrelated (Mokkink et al., 2010). Internal consistency is a measure of the degree to which items test the same construct in a unidimensional (sub)scale of a multiple-item instrument (de Vet et al., 2011).

As a psychometric property, reliability refers to the proportion of the total variance in the measurement due to "true" differences among people (Mokkink et al., 2010). Reliability concerns how consistent the scores obtained from repeated measurements about the construct of interest of people with stable condition are over time (test-retest reliability), between different raters (interrater reliability), and within one rater (interrater reliability) (de Vet et al., 2011).

Measurement error defines the error that is not attributable to true changes in the construct to be measured, but that is due to the systematic and random error of a respondent's score (Mokkink et al., 2010). It is the absolute measurement error over repeated measurements of the construct of interest when the construct of interest is stable between measurements (de Vet et al., 2011). Furthermore, compared with reliability, which depends on the variability between individuals (de Vet et al., 2011), measurement error is affected by the variability within individuals (de Vet et al., 2006). Thus, measurement error is more useful for explaining how reliably an instrument assesses the intra-individual variability between repeated measurements (for monitoring change in an individual person's trait, such as evaluation of change in a child's weight over time), while reliability is useful for explaining how reliably an instrument assesses the inter-individual variability (for screening a group, such as discrimination between overweight and obese children; de Vet et al., 2006; Verweij et al., 2013).

3.2.3 Responsiveness

The domain of responsiveness defines the sensitivity of an instrument in detecting changes in the construct of interest over time (Mokkink et al., 2010). Accordingly, evaluative instruments used for clinical and research purposes must be able to detect and quantify changes in status of people (as the construct of interest) over time (de Vet et al., 2011). Responsiveness requires a longitudinal study design with repeated measurements to be conducted to calculate the change between baseline and follow-up scores when changes in people's construct of interest are expected (i.e., a proportion of people will worsen or improve). If no change in the instrument's scores between repeated measurements were expected, it would be impossible to determine whether the unchanged scores were due to the stable status of the people or the poor responsiveness of the instruments (de Vet et al., 2011).

To test for responsiveness, the following two approaches can be applied: criterion and construct approaches (Mokkink, de Vet, et al., 2018; Mokkink, Prinsen, et al., 2018). The criterion approach tests the association of changes in scores between an instrument and a gold standard to detect the effect of an intervention for the prevention of CM (Mokkink, de Vet, et al., 2018; Mokkink, Prinsen, et al., 2018). If no gold standard is available for an instrument to measure the construct of interest, the criterion approach cannot be used to assess the instrument. The construct approach includes the following three aspects (Mokkink, Prinsen, et al., 2018): (1) comparison of the instrument with other outcome instruments (i.e., the association of the changes in scores between the instrument under review and other instruments used to measure a similar construct); (2) comparison between subgroups (i.e., the mean difference in change scores for the instrument between different subgroups); and (3) comparison before and after an intervention (i.e., the mean difference in the change scores on the instrument from before and after the intervention). The construct approach may be more feasible in the evaluation of responsiveness for self-reported or self-administered instruments than the criterion approach due to the lack of gold standards for instruments that collect subjective information (de Vet et al., 2011).

4 Methodology

This methodology chapter is divided into three subchapters. Subchapter 4.1 describes what a systematic review is and explains how it is conducted. Subchapter 4.2 discusses the systematic review of psychometric properties and briefly introduces the COSMIN methodology for evaluating the psychometric properties of measurement instruments (Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). Subchapter 4.3 discusses why the COSMIN methodology is an appropriate method for this thesis by comparing its strengths and limitations with those of other similar methodologies. Finally, Subchapter 4.4 presents the application of the COSMIN methodology in this thesis, including the collection and analysis of data.

4.1 Systematic Reviews

This subchapter defines systematic reviews (4.1.1) and describes how systematic reviews are conducted (4.1.2).

4.1.1 Definition of Systematic Review

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) review group (Moher et al., 2009), which has produced guidelines for reporting and conducting systematic reviews and meta-analyses, defined a systematic review as follows: *"A systematic review is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review"* (p. 1). To specify this broad definition, the PRISMA review group (Liberati et al., 2009) suggested that a systematic review should have the following six key characteristics:

- Clearly stated research questions;
- Pre-defined eligibility criteria for the included studies;
- A systematic literature search that attempts to identify all the studies that would meet the eligibility criteria;
- An assessment of the methodological quality (or risk of bias) of the included studies;
- A systematic summary of the results of the included studies; and

- Systematic reporting of the summarised results and study characteristics.

As the PRISMA definition is by far the most commonly accepted definition of a systematic review (Liberati et al., 2009; Moher et al., 2009), the PRISMA definition was applied to conduct the systematic review of this thesis.

4.1.2 PRISMA Procedure for Systematic Reviews

Based on the definition (Moher et al., 2009) and characteristics (Liberati et al., 2009) of systematic reviews proposed by the PRISMA review group, the following five phases should be carried out in systematic reviews (Liberati et al., 2009; Moher et al., 2009): (1) *formulating research questions and eligibility criteria* to include the literature that is relevant to the research questions; (2) *performing a systematic literature search* to identify all literature that would meet the eligibility criteria; (3) *assessing the methodological quality* (or risk of bias) of the included studies; (4) *summarising the results* of the included studies; and (5) *reporting the summarised results* and study characteristics.

As a guideline for conducting systematic reviews following the suggested PRISMA procedure, the PRISMA statement (Liberati et al., 2009; Moher et al., 2009) provides a detailed checklist (PRISMA checklist) of the minimum information that needs to be reported in each phase of a systematic review, as well as a specific workflow (PRISMA flow chart; Moher et al., 2009) for performing a systematic literature search in Phase 2. Phases 1 and 2 are more relevant to all types of systematic reviews than the other phases (Phases 3, 4 and 5), which are more appropriate for a meta-analysis (i.e., a statistical method that combines the results from several included studies to obtain a single summarised effect size of such an intervention; Liberati et al., 2009). As a meta-analysis of an intervention was not the main purpose of this thesis, the PRISMA statement was used to formulate eligibility criteria and perform a systematic literature search (Phases 1 and 2, respectively). The PRISMA flow chart (Moher et al., 2009) was particularly applied to Phase 2 for the systematic literature search.

In a systematic literature search (Phase 2), the PRISMA flow chart (Moher et al., 2009) suggests the following four consecutive stages: (1) *identification*, (2) *screening*, (3) *eligibility*, and (4) *inclusion*. *Identification* refers to identifying relevant literature through database searching and other sources of literature (Moher et al., 2009). Identifying appropriate databases related to the review topic should be conducted first, followed by searching with relevant subject headings and free texts in databases (Moher et al., 2009).

Compared with free texts that are non-standardised but commonly used terms to describe a concept, subject headings are standardised and assigned terms used in databases to uniformly describe a concept, which relieves researchers from considering synonyms and spelling variations when searching databases. *Screening* refers to assessing the abstracts and titles of identified literature to either include or exclude them based on the pre-defined eligibility criteria (Moher et al., 2009). This stage starts with removing duplicates among the identified literature from the database search, followed by the review of the titles and abstracts of the identified literature by two independent reviewers to include the eligible abstracts (Moher et al., 2009). *Eligibility* refers to conducting a more comprehensive evaluation of the full-text articles and determining whether the full texts should be included or excluded (Moher et al., 2009). Finally, the *inclusion* stage involves determining how many articles will be included in the data analysis, which is critical for assessing the methodological quality and summarising the results of the included studies (Moher et al., 2009).

4.2 Systematic Reviews of Psychometric Properties

Systematic reviews for evaluating psychometric properties of instruments involve identifying, critically appraising and summarising evidence from the literature of an instrument's psychometric properties (de Vet et al., 2011; Mokkink et al., 2009). The results from psychometric reviews help practitioners and researchers make informed decisions about whether an instrument should be used (Prinsen et al., 2016). The quality of the results in psychometric reviews mainly relies on critical appraisals to assess and summarise the quality of evidence supporting the psychometric properties of the reviewed instruments (de Vet et al., 2011; Mokkink et al., 2009). Critical appraisals usually involve evaluating the psychometric quality (i.e., validity, reliability and responsiveness) against pre-defined criteria and assessing the study quality for issues such as risk of bias (de Vet et al., 2011). Critical appraisal of an instrument's interpretability and feasibility should also be conducted (Prinsen et al., 2016). While other critical appraisal methodologies have been developed, the COSMIN methodology (Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018) remains the benchmark in the field of psychometric review due to its comprehensiveness and standardisation (Aromataris & Munn, 2020; Rosenkoetter & Tate, 2018); therefore, the COSMIN methodology was chosen to guide this thesis.

4.3 Strengths and Limitations of the COSMIN Methodology

One of the main strengths of the COSMIN methodology (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018) is its standardised terms and definitions of psychometric properties, which counteracts confusion when extracting and reporting psychometric data (Prinsen et al., 2018). The COSMIN taxonomy of psychometric properties (Mokkink et al., 2010) was developed through consensus among 40 international experts on instrument development within the field of patient-report instruments. While other research groups developed psychometric taxonomies as well (Polit, 2015), the taxonomy developed by Polit (2015) was created based on the opinions of a small group of individual experts. Moreover, the taxonomy of Polit (2015) has not consistently been used to develop critical appraisal tools, such as the COSMIN Risk of Bias checklist (Mokkink, de Vet, et al., 2018), used for evaluating the quality of studies reporting on any of the nine psychometric properties of an instrument (Mokkink et al., 2016). Another research group developed a simplified checklist for assessing the quality of psychometric studies of patient-report instruments (Francis et al., 2016). However, due to its simplicity, the checklist developed by Francis et al. (2016) does not provide sufficient detail for unbiased and systematic ratings of study design quality (Mokkink, de Vet, et al., 2018). For example, criteria on which data analyses and techniques are suitable for good-quality studies on content validity, factor structure, and responsiveness are lacking (Terwee, de Vet, et al., 2016). In addition, several checklists have been designed for evaluating study quality, but all of these checklists include only limited psychometric properties. For instance, the updated Quality Assessment of Diagnostic Accuracy Studies checklist (Whiting et al., 2011) is primarily concerned with the single psychometric property of criterion validity (Christian et al., 2019), while the Quality Appraisal of Reliability Studies checklist (Lucas et al., 2010) was developed only to evaluate reliability (Abedi et al., 2019).

The COSMIN checklist (Mokkink, de Vet, et al., 2018) is the only consensus-based comprehensive checklist that contains detailed standards for the preferred designs of studies on any psychometric property (Terwee, Prinsen, et al., 2016). In addition, the COSMIN methodology (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018) provides consensus-based quality criteria for evaluating single-study results for each psychometric property separately, and a rating system that allows summarising all study results on each psychometric property and grading the *quality of evidence.* The quality of evidence refers to the level of confidence or certainty in the summarised results on each psychometric property; to determine the quality of evidence, all bodies of evidence used for assessing both the methodological and the psychometric quality are considered. All these critical appraisal tools are provided in the comprehensive COSMIN user manuals (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) to help reviewers avoid making subjective quality assessments.

However, the size and complexity of the COSMIN manuals (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) pose some challenges. For example, significant time and effort are needed to complete all stages of the quality assessments for study design, single study results, summarising all results, and grading the level of confidence in the summarised results (Kwok et al., 2021). Additionally, while the COSMIN group has claimed that its quality criteria of the COSMIN methodology are also applicable to non-patient-reported outcome instruments (Prinsen et al., 2018), it has also been argued that not all of the criteria are appropriate to be applied to other types of instruments (e.g., clinician-report instruments to measure speech performance in children; Kwok et al., 2021). Furthermore, the COSMIN methodology (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) lacks a rating scale to assess interpretability and feasibility, even though these characteristics are considered important for instrument selection (Kwok et al., 2021). Last, the COSMIN methodology (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) does not provide detailed guidelines for a systematic literature search (except some examples of search terms for psychometric properties that are available for different databases), including formulating eligibility criteria, searching the literature, and selecting eligible studies (Aromataris & Munn, 2020). However, a systematic literature search can be performed using the PRISMA statement (Moher et al., 2009), which provides more detailed information on how to conduct the systematic literature search in various types of systematic reviews. For this reason, using the PRISMA statement (Moher et al., 2009) for the systematic literature search and the COSMIN tools (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) for psychometric quality assessments is recommended for performing a systematic review of the psychometric properties of instruments (Aromataris & Munn, 2020; Mokkink, Prinsen, et al., 2018).

In summary, as long as the PRISMA statement (Moher et al., 2009) is used for the systematic literature search in a systematic review of psychometric properties, the strengths of the COSMIN methodology (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) far exceed its weaknesses. For this reason, when conducting a systematic review to evaluate the psychometric properties of instruments, the use of the COSMIN methodology (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) with the PRISMA statement (Moher et al., 2009) for the systematic literature search has been officially recommended by the Joanna Briggs Institute (Aromataris & Munn, 2020), one of the leading international organisations that has developed guidelines for conducting systematic reviews. Therefore, the COSMIN methodology (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) and the PRISMA statement (Moher et al., 2009) were used in this thesis.

4.4 The COSMIN Method and the Current Thesis

This thesis followed the PRISMA statement (Moher et al., 2009) and the COSMIN methodology (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). The three reviews (Yoon et al., 2020a, 2020b, 2021) in this thesis were conducted in four consecutive steps (see Figure 4.1).

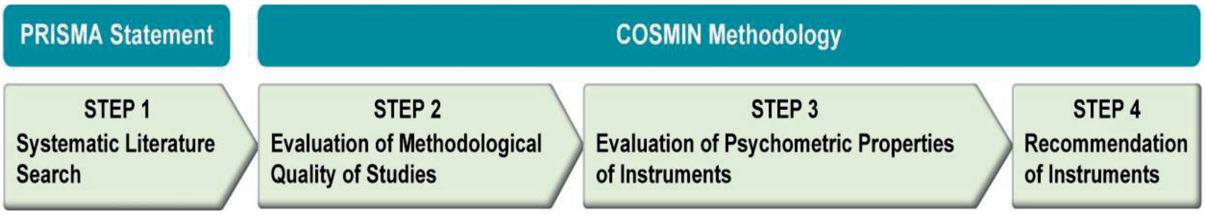

PRISMA Statement
COSMIN Methodology
STEP 1
Systematic Literature Search
STEP 2
Evaluation of Methodological Quality of Studies
STEP 3
Evaluation of Psychometric Properties of Instruments
STEP 4
Recommendation of Instruments

Figure 4.1. Study Design: Steps for the PRISMA Statement (Step 1; Moher et al., 2009) and the COSMIN Process (Steps 2, 3, and 4; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018)

Each of these steps is briefly explained in the following sections. A detailed explanation can be found in the Methods sections of the three review papers (Yoon et al., 2020a, 2020b, 2021).

4.4.1 Step 1. Systematic Literature Search

The systematic literature search (Step 1 in Figure 4.1) for the three papers (Yoon et al., 2020a, 2020b, 2021) was conducted by (1) formulating eligibility criteria, (2) searching the literature, and (3) selecting studies. Eligibility criteria for selecting instruments were formulated as follows: (1) instruments reported by parents or caregivers, (2) instruments measuring parents' or caregivers' own perpetration of CM or attitudes towards CM, (3) instruments developed and published in English; and (4) instruments measuring one or more subtypes of CM, including physical abuse, emotional abuse, sexual abuse, and/or neglect. To select psychometric studies, the following two additional inclusion criteria were formulated: (1) studies (journal articles and manuals) published in English and (2) studies reporting psychometric data on one or more of the eight psychometric properties of eligible instruments as defined in the COSMIN taxonomy (i.e., content validity, criterion validity, structural validity, cross-cultural validity, hypothesies testing for construct validity, internal consistency, reliability, and measurement error; Mokkink et al., 2010). To select studies on responsiveness in Paper 3 (Yoon et al., 2021), all studies reporting the *change scores of the included instruments before and after intervention* (i.e., responsiveness data) needed to be included; hence, different eligibility criteria than the review of the other psychometric properties in Paper 1 (Yoon et al., 2020a) and Paper 2 (Yoon et al., 2020b) were formulated for responsiveness (see *Eligibility criteria [Step 1.1]* in Paper 3; Yoon et al., 2021).

To retrieve eligible instruments and psychometric studies, systematic literature searches were conducted in six electronic databases (CINAHL, Embase, ERIC, PsycINFO, PubMed, and Sociological Abstracts) in October 2019 for both Paper 1 (Yoon et al., 2020a) and Paper 2 (Yoon et al., 2020b), and in March 2021 for Paper 3 (Yoon et al., 2021). As a review of responsiveness, Paper 3 required the review of all studies using the included instruments as an outcome measure. For this reason, searching the literature on responsiveness was performed after identifying all eligible parent- or caregiver-report CM instruments in Paper 1 (Yoon et al., 2020a).

Finally, the abstracts and full texts of the eligible studies identified through database searches were screened by two independent reviewers to retrieve eligible instruments and full-text articles on any psychometric property. Any discrepancies between the two reviewers were resolved through consensus involving a third reviewer. In addition, the reference lists of all selected full-text articles were hand searched to identify additional eligible instruments and psychometric studies.

4.4.2 Step 2. Evaluation of the Methodological Quality of Included Studies

The methodological quality of the included studies regarding at least one of the nine psychometric properties of the identified instruments was rated using the COSMIN Risk of

Bias checklist (Step 3 in Figure 4.1; Mokkink, de Vet, et al., 2018). The checklist contains between 3 and 38 items for each psychometric property (Mokkink, de Vet, et al., 2018). The checklist items were used to rate the quality of the study design and the robustness of the statistical methods conducted to investigate the nine psychometric properties assessed in this thesis (Mokkink, de Vet, et al., 2018). When rating the methodological quality of the included psychometric studies, each checklist item was scored on a four-point scale (inadequate = 1, doubtful = 2, adequate = 3, and very good = 4; Mokkink, de Vet, et al., 2018). A total rating for each of the nine psychometric properties was obtained by calculating the ratio of the ratings (Cordier et al., 2015). Thus, the total score of the methodological quality ratings for each psychometric property was reported as a ratio of the ratings: inadequate (0%–25%), doubtful (25.1%–50%), adequate (50.1%–75%), and very good (75.1%–100%). The ratings of the methodological quality were conducted by two reviewers independently, and any differences were resolved through consensus between the two reviewers.

Content validity was evaluated before the other psychometric properties because it is the most important psychometric property (Prinsen et al., 2018). If the content validity of an included instrument was poor in Paper 1 (Yoon et al., 2020a), the evaluation of its other psychometric properties was not conducted in either Paper 2 (Yoon et al., 2020b) or Paper 3 (Yoon et al., 2021).

4.4.3 Step 3. Evaluation of the Psychometric Properties of the Instruments

For evaluation of the instruments' psychometric properties (Step 3 in Figure 4.1), all results for each of the nine psychometric properties per instrument that were obtained from the included studies were combined. The combined results were scored as either overall sufficient (+ = above the threshold of the quality criteria), insufficient (– = below the threshold of the quality criteria), or indeterminate (? = a lack of robust data meeting the quality criteria) against the pre-defined criteria for good psychometric properties (Mokkink, Prinsen, et al., 2018).

In addition, to indicate the level of confidence in the combined results (or overall ratings) for each psychometric property, the quality of evidence was graded by considering all bodies of evidence used to assess both the methodological and psychometric quality. A high, moderate, low, or very low quality of evidence was graded using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Guyatt et al., 2008; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018). The initial quality of evidence used for the overall ratings was high, but the quality of evidence was subsequently downgraded by one or more levels (to moderate, low or very low) when there were serious concerns regarding the following four factors (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018): (1) risk of bias (limitations in the methodological quality of the included studies), (2) inconsistency (heterogeneity in the results of the included studies), (3) indirectness (evidence from populations other than the target population of interest), and (4) imprecision (a low total number of participants included in the studies). Evidence quality was not graded if the overall rating was indeterminate (?) due to a lack of robust evidence (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018).

4.4.4 Step 4. Recommendation of Instruments

The recommendation of suitable instruments for future use was conducted by combining the results of the overall ratings on each of nine psychometric properties (Step 3.2 in Figure 4.1) and the grades for the quality of evidence used for the overall ratings on each property (Step 3.3 in Figure 4.1; Prinsen et al., 2018). The recommendations were based on all results of the nine psychometric properties of the included instruments from the three papers (Yoon et al., 2020a, 2020b, 2021). Each of the 15 included instruments was classified into the following 3 categories for recommendation (Mokkink, Prinsen, et al., 2018): (A) most suitable (i.e., instruments having high-quality evidence supporting sufficient content validity in any aspect of relevance, comprehensiveness, and comprehensibility; and at least low-quality evidence supporting sufficient internal consistency); (B) promising but need further validation studies (i.e., instruments categorised as neither A nor C); and (C) not recommendable (i.e., instruments having high-quality evidence supporting an insufficient psychometric property).

To recommend suitable instruments, the decisive psychometric properties include content validity and internal consistency, because when it is unclear what the content of an instrument is measuring and how different items in the instrument are associated with the construct being measured, evaluating the other psychometric properties is meaningless (Prinsen et al., 2018). Moreover, when it is difficult to differentiate the quality of an instrument's psychometric properties, interpretability (the extent to which clinical meaning can be assigned to an instrument's quantitative scores or change scores) and feasibility (ease of use including the completion time, length, and cost of an instrument) can help in selecting the most suitable instruments. However, interpretability and feasibility are not considered psychometric properties (Prinsen et al., 2018); hence, both were not evaluated in this thesis.

5 Summary of Articles

The aim of this thesis was to recommend the most suitable parent- or caregiver-report CM instruments in terms of psychometric quality. This overarching aim was investigated through the three studies published in the journal *Trauma, Violence, and Abuse*, which specialises in review articles in the field of social work and has no strict word limit. The generous word limit of the journal allowed the three studies to explain all the details of the COSMIN methodology. The results from each of the three articles are summarised and presented in the following three subchapters: 5.1 Paper 1 on Content Validity; 5.2 Paper 2 on Construct Validity, Criterion Validity, and Reliability; and 5.3 Paper 3 on Responsiveness. Based on the summarised results from all three papers, Subchapter 5.4 provides recommendations of the most robust parent- or caregiver-report CM instruments in terms of their psychometric quality according to the COSMIN methodology.

5.1 Paper 1 on Content Validity

Paper 1 (Yoon et al., 2020a) full citation:

- Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 1: Content Validity. *Trauma, Violence, & Abuse.* Advanced online publication. https://doi.org/10.1177/1524838019898456

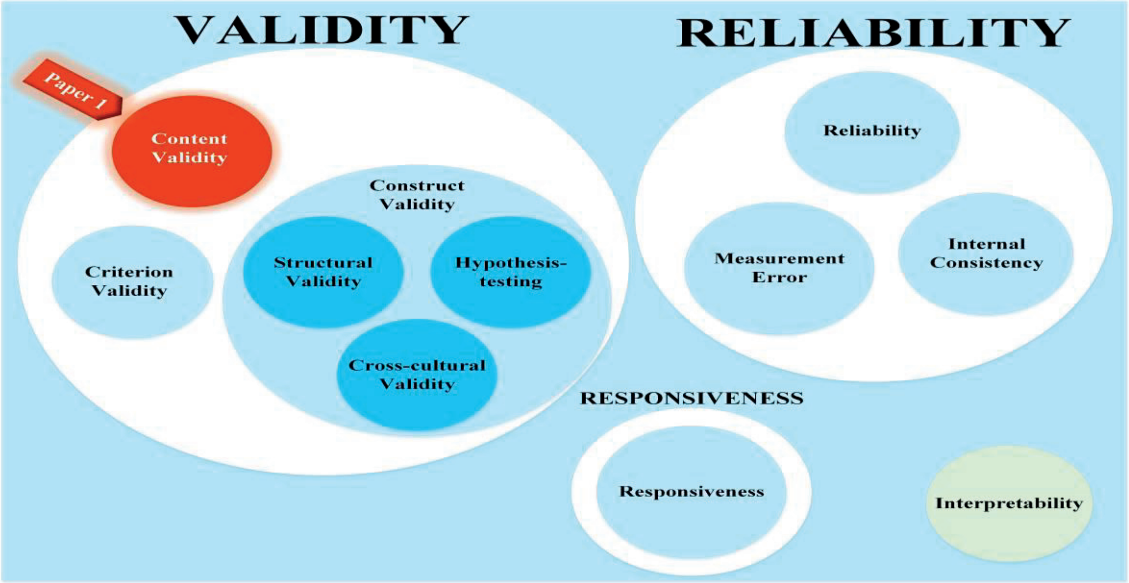


Figure 5.1. Position of Paper 1 on Content Validity within the COSMIN Taxonomy

Paper 1 (Yoon et al., 2020a) aimed to assess the content validity (see Figure 5.1) of all currently available parent- or caregiver-report CM instruments by following the COSMIN methodology (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018). The following two research questions (RQs 1 and 2) of this thesis guided the study of Paper 1 (Yoon et al., 2020a):

- **RQ 1.** Which parent- or caregiver-report instruments have been published to measure their attitudes towards CM or maltreating behaviours towards their children?
- **RQ 2.** What is the quality of studies and psychometric evidence on the content validity of the existing parent- or caregiver-report CM instruments?

The systematic literature search (Step 1 in Figure 4.1) identified 15 studies on the content validity of 15 identified instruments (see Figure 2 in Paper 1; Yoon et al., 2020a). The characteristics of the identified studies and instruments can be found in Table 1 and Online Appendix C of Paper 1 (Yoon et al., 2020a). The methodological quality of the included studies (Step 2 in Figure 4.1) was generally poor (see Table 2 in Paper 1; Yoon et al., 2020a). The interrater reliability for the assessment of methodological quality between two independent reviewers was good (i.e., a weighted κ of 0.76 and a 95% CI of 0.68–0.85). Last, the evaluation of psychometric properties of the included instruments (Step 3 in Figure 4.1) found that the content validity of the 15 included instruments was generally sufficient, but most of the included instruments did not offer high-quality evidence (see Table 4 in Paper 1; Yoon et al., 2020a).

Based on the results, most of the instruments included in Paper 1 (Yoon et al., 2020a) demonstrated promising content validity. The International Society for the Prevention of Child Abuse and Neglect (ISPCAN) Child Abuse Screening Tool for use in Trials (ICAST-Trial) and the Family Maltreatment–Child Abuse (FM-CA) criteria appeared to be the most promising based on current evidence of content validity. However, strong conclusions cannot be drawn due to the overall low-quality of the evidence regarding content validity. Additional studies are needed to evaluate psychometric properties other than the content validity to recommend parent- or caregiver-report CM instruments.

5.2 Paper 2 on Construct Validity, Criterion Validity, and Reliability

Paper 2 (Yoon et al., 2020b) full citation:

- Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 2: Internal Consistency, Reliability, Measurement Error, Structural Validity, Hypothesis Testing, Cross-Cultural Validity, and Criterion Validity. *Trauma, Violence, & Abuse.* Advanced online publication. https://doi.org/10.1177/1524838020915591

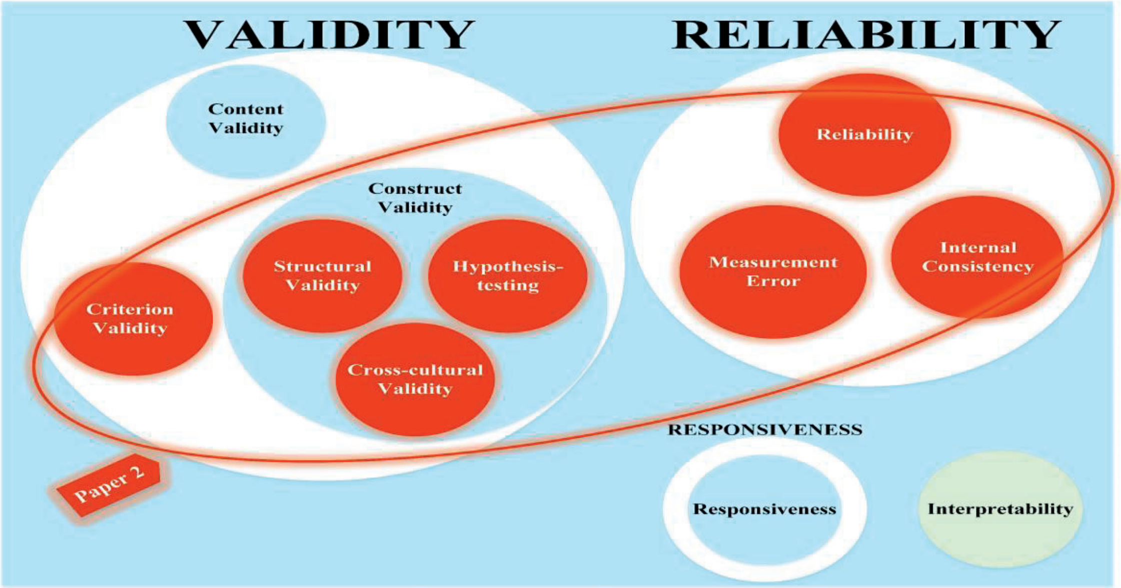


Figure 5.2. Position of Paper 2 on Construct Validity, Criterion Validity, and Reliability within the COSMIN Taxonomy

The aim of Paper 2 (Yoon et al., 2020b) was to evaluate the following seven psychometric properties (see Figure 5.2) of all currently available parent- or caregiver-report CM instruments using the COSMIN methodological manual (Mokkink, Prinsen, et al., 2018): structural validity, cross-cultural validity, and hypothesis testing (the three psychometric properties of construct validity); criterion validity; and internal consistency, reliability, and measurement error (the three properties of reliability). The following research question (RQ 3) of this thesis was addressed in Paper 2 (Yoon et al., 2020b):

- **RQ 3.** What is the quality of studies and psychometric evidence on the construct validity, criterion validity, and reliability of the existing parent- or caregiver-report CM instruments?

The systematic literature search (Step 1 in Figure 4.1) found 25 studies on the validity (other than content validity) and reliability of the 15 identified instruments (see Figure 2 in Paper 2; Yoon et al., 2020b). The characteristics of all identified studies and instruments can be found in Table 1 and Online Appendix C of Paper 2 (Yoon et al., 2020b). The methodological quality of the included studies (Step 2 in Figure 4.1) was adequate overall (see Table 2 in Paper 2; Yoon et al., 2020b). For the study quality assessment, the interrater reliability between the two independent reviewers was very good (i.e., a weighted κ of 0.86 and a 95% CI of 0.83–0.90). Last, the seven psychometric properties of the included instruments (Step 3 in Figure 4.1) were mostly not reported (NR) or indeterminate due to either missing or incomplete psychometric data; high-quality evidence for the seven psychometric properties was limited (see Table 4 in Paper 2; Yoon et al., 2020b).

Based on these results, 6 of the 15 instruments included in Paper 2 (Yoon et al., 2020b) could not be recommended, but further validation studies on hypothesis testing and/or internal consistency should be conducted to confirm whether these instruments should indeed not be recommended. The other nine instruments showed promising validity (other than content validity) and reliability, but still required further validation due to the lack of high-quality psychometric evidence. Additional studies are needed to evaluate the responsiveness of the 15 included instruments before the recommendation of the most suitable parent- or caregiver-report instruments measuring CM can be made.

5.3 Paper 3 on Responsiveness

Paper 3 (Yoon et al., 2021) full citation:

- Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2021). A Systematic Review Evaluating Responsiveness of Parent- or Caregiver-Reported Child Maltreatment Instruments to Parenting Interventions. *Trauma, Violence, & Abuse*. Manuscript submitted for publication.

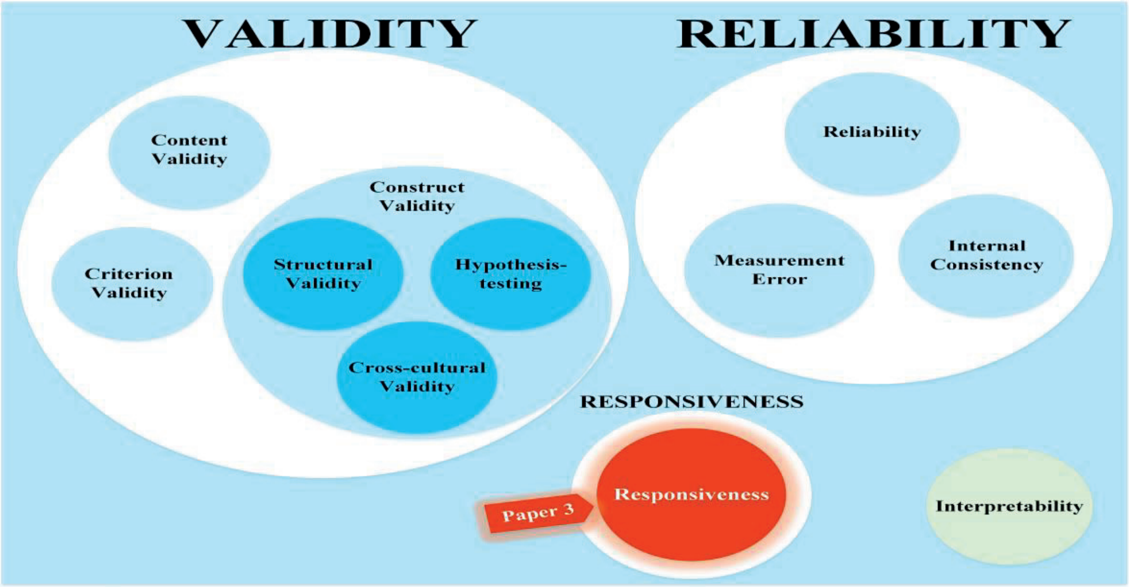


Figure 5.3. Position of Paper 3 on Responsiveness within the COSMIN Taxonomy

Paper 3 (Yoon et al., 2021) aimed to assess the responsiveness (see Figure 5.3) of all currently available parent- or caregiver-report CM instruments using the COSMIN methodological manual (Mokkink, Prinsen, et al., 2018). To achieve this aim, Paper 3 (Yoon et al., 2021) addressed the following research question (RQ 4) of this thesis:

- **RQ 4.** What is the quality of studies and psychometric evidence on the responsiveness of the parent- or caregiver-report CM instruments?

The systematic literature search (Step 1 in Figure 4.1) identified 69 journal articles on the responsiveness of the 15 included instruments (see Figure 2 in Paper 3; Yoon et al., 2021). The characteristics of the identified articles and instruments are presented in Table 1 and Online Supplemental Table S5 of Paper 3 (Yoon et al., 2021). The methodological quality of the identified studies (Step 2 in Figure 4.1) was generally adequate (see Table 2 in Paper 3; Yoon et al., 2021). For the study quality assessment, the interrater reliability between two independent reviewers was very good (i.e., weighted κ 0.83 and 95% CI of 0.75 to 0.90). Last, the responsiveness of the included instruments (Step 3 in Figure 4.1) was either insufficient overall or not reported (NR); no high-quality evidence of sufficient or insufficient responsiveness was found except for the Physical Abuse subscale of the ICAST-Trial (see Table 4 in Paper 3; Yoon et al., 2021).

Based on these results, only the Physical Abuse subscale of the ICAST-Trial (Meinck et al., 2018) can be recommended as the most responsive for use in parenting interventions, with high-quality evidence supporting it as having sufficient responsiveness. All other instruments were identified as promising based on the currently available data on responsiveness. However, further psychometric evidence on responsiveness is needed before their recommendation for use in parenting interventions to reduce CM.

5.4 Recommendation of Instruments

Table 5.1 presents the recommendations for the most suitable parent- or caregiver-report CM instruments for use in research and clinics based on the results from all three papers (Yoon et al., 2020a, 2020b, 2021). None of the 15 included instruments could be recommended as the most suitable for use (category A) due to a lack of high-quality evidence for sufficient content validity, as reported in Paper 1 (Yoon et al., 2020a); and lack of evidence or at least low-quality evidence for sufficient internal consistency, as reported in Paper 2 (Yoon et al., 2020b). Six instruments (CNQ, CTSPC, ICAST-Trial, MCNS, MCNS-SF, and POQ) could not be recommended at all (category C) due to high-quality evidence for an insufficient psychometric property (i.e., insufficient hypothesis testing for all six instruments and insufficient internal consistency for the ICAST-Trial only), as reported in Paper 2 (Yoon et al., 2020b) and Paper 3 (Yoon et al., 2021). The other nine instruments (AAPI-2, APT, CNS-MMS, CTS-ES, FM-CA, IPPS, P-CAAM, PRCM and SBS-SV) may have the potential to be recommended, but further validation studies are needed (category B) due to a lack of high-quality evidence for sufficient psychometric properties.

Table 5.1. *Recommendations for Suitable Instruments adapted from Prinsen et al. (2018)*

Category	Description on Category	Criteria	Instruments	
A: Most suitable	Instruments that have the potential to be recommended for use in respect of the construct and population of interest	High-quality evidence for sufficient content validity in any aspects AND at least low-quality evidence for sufficient internal consistency	None	
B: Promising but need further validation studies	Instruments that may have the potential to be recommended for use, but need further validation studies	Not categorised in A or C	• AAPI-2 • APT • CNS-MMS • CTS-ES • FM-CA	• IPPS • P-CAAM • PRCM • SBS-SV
C: Not recommendable	Instruments that should not be recommended for use	High-quality evidence for an insufficient psychometric property	• CNQ • CTSPC • ICAST-Trial	• MCNS • MCNS-SF • POQ

Notes. 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.

6 Discussion

This thesis aimed to recommend the most suitable parent- or caregiver-report CM instruments in terms of psychometric quality using the COSMIN methodology. To address this overarching purpose, three studies were undertaken with the following research questions: RQ 1. Which parent- or caregiver-report instruments have been published to measure their attitudes towards CM or maltreating behaviours towards their children? (Paper 1); RQ 2. What is the quality of studies and psychometric evidence on the content validity of the existing parent- or caregiver-report CM instruments? (Paper 1); RQ 3. What is the quality of studies and psychometric evidence on the construct validity, criterion validity, and reliability of the existing parent- or caregiver-report CM instruments? (Paper 2); and RQ 4. What is the quality of studies and psychometric evidence on the responsiveness of the existing parent- or caregiver-report CM instruments? (Paper 3). By summarising the results of the 3 papers (Yoon et al., 2020a, 2020b, 2021), this thesis found that none of the 15 identified instruments on CM have the potential to be recommended as the most suitable, as defined by the COSMIN methodology (Prinsen et al., 2018). While nine instruments have the potential for use in clinical practice and research, their psychometric properties need to be evaluated further, and the other six instruments could not be recommended at all. Notably, these recommendations were not based on high-quality evidence; the studies had either a lack of evidence or low-quality evidence.

This chapter begins by discussing which constructs were measured (i.e., the types of CM, the attitudes towards CM and the maltreating behaviours, and the severity, frequency, and timing of CM) in the included instruments in Subchapter 6.1 *Characteristics of the Included Instruments*. Next, the methodological flaws and evidence gaps in the included studies are identified and discussed for each psychometric property in Subchapter 6.2 *Psychometric Properties and Recommendations*. Third, the methodological challenges that emerged when applying the COSMIN method and the limitations in the results of this thesis are discussed in Subchapter 6.3 *Challenges and Limitations*. Based on the identified flaws and gaps discussed in Subchapter 6.2, the *Implications for Future Research* (6.4) to improve future development and validation studies are presented. Fourth, the *Implications for Future Practice* (6.5) are discussed for the identified instruments and the method used in this thesis in relation to the social ecological model, public health approaches, and evidence-based assessment. Finally, the *Concluding Remarks (6.6)* presents a brief summary of the major findings and the recommendations resulting from this thesis.

6.1 Characteristics of the Included Instruments

Regarding the main constructs of the instruments, most of the instruments (9/15) measured multiple types of CM (see Table 1 in Paper 3; Yoon et al., 2021): two instruments (CTS-ES and ICAST-Trial) measure all four types of CM; three (AAPI-2, POQ, and SBS-SV) measure physical abuse, emotional abuse, and neglect; and four (CTSPC, FM-CA, IPPS, and PRCM) measure physical and emotional abuse. The other six instruments (APT, CNQ, CNS-MMS, MCNS, MCNS-SF, and P-CAAM) measure only one type of CM. In addition, the response options presented in Table 1 of Paper 3 (Yoon et al., 2021) show which instruments measure either parental attitudes towards CM or maltreating behaviours towards their children. Eight instruments (AAPI-2, APT, IPPS, MCNS, MCNS-SF, P-CAAM, POQ, and SBS-SV) measure attitudes towards CM by asking parents or caregivers about the extent to which they agree with or prefer the use of CM. The other seven instruments measure maltreating behaviours: six (CNQ, CNS-MMS, CTSPC, FM-CA, ICAST-Trial, and PRCM) ask parents or caregivers how often they engage in maltreating behaviours towards their children; and one (CTS-ES) asks them whether their children have been exposed to their maltreating behaviours. These response options also show which instruments collect data on the severity, frequency, and timing of CM. All instruments on attitudes towards CM measure the severity (or degree) of the attitudes; all instruments on maltreating behaviours measure the frequency of CM, except for CTS-ES, which measures the exposure to CM. However, no instruments were identified to measure the timing of CM, which may be because parents cannot recall precisely when they perpetrated CM (Milner & Crouch, 1997). The severity of maltreating behaviours towards their children was also not identified, which may be because of parents' concerns about the legal consequences of reporting their severe maltreating behaviours towards their children (Compier-de Block et al., 2017).

6.2 Psychometric Properties and Recommendations

This subchapter discusses the results of the three psychometric reviews in relation to the methodological flaws of the included studies in their investigation of each psychometric property. The methodological flaws are discussed as follows: content validity (Section 6.2.1); construct validity, criterion validity, and reliability (Section 6.2.2); and responsiveness

(Section 6.2.3). In addition, the evidence gaps that need to be filled to determine the psychometric quality of instruments before reaching firm conclusions on the recommendation of the instruments are discussed in more detail (Section 6.2.4).

6.2.1 Content Validity

Most instrument development studies included in Paper 1, generated new items based on the relevant literature, existing instruments and/or professional input by the developers themselves, but not based on the input of the target population (parents or caregivers). Input from the target population is essential for generating new instrument items with good content validity (Terwee, Prinsen, Chiarotto, Westerman, et al., 2018). To generate relevant, comprehensive, comprehensible items for the respondents (target population), the respondents' own perceptions or experiences related to the construct of interest should be obtained through interviews or surveys (Ricci et al., 2018). If the respondents feel that the instrument items are irrelevant, unimportant, or too difficult, the instrument items will fail to precisely assess the respondents' attitudes and behaviours (Wiering et al., 2017). Thus, in terms of generating new items, instrument development studies may have important methodological flaws due to a lack of input from the target population.

Only a few content validity studies have asked parents or caregivers their opinions about the relevance, comprehensiveness and comprehensibility of the instrument items. The relevance of the final version of the instruments was assessed mainly based on input from professionals. The assessment of the comprehensiveness of instruments lacked input from professionals, parents or caregivers, and the comprehensibility was rarely assessed by asking parents or caregivers for input. In addition, the few content validity studies that assessed the relevance and comprehensibility of the instruments by asking parents or caregivers for input, mostly did not report the required details in the study design and results. However, these details are needed for a clear evaluation of the instruments' content validity. Thus, these methodological flaws made it difficult to determine whether the content validity of the instruments was sufficient based on the reported study evidence.

6.2.2 Construct Validity, Criterion Validity and Reliability

Regarding structural validity, the studies of most of the instruments (9/15) either did not report any psychometric data or analysed the factor structure of the instruments with a less preferred method (e.g., exploratory factor analysis [EFA]). EFA identifies a factor structure of a new instrument when there is no existing hypothesis of the structure. However, the structural validity is needed to test an existing hypothesis regarding the factor structure of an already developed instrument (Mokkink, Prinsen, et al., 2018). To test the existing hypothesis of the factor structure, either confirmative factor analysis (CFA) or item response theory (IRT) analysis is preferred in the COSMIN methodology (Mokkink, de Vet, et al., 2018). Although both CFA and IRT have the same overall purpose for testing how well the data fit a priori hypothesised factor structure (de Vet et al., 2011), the specific foci on methods for handling or interpreting the data in each type of analysis differ. CFA focuses on total responses or summed scores under the assumption that each response for all items is equally weighted in terms of difficulty or severity. In contrast, IRT analysis is focused on individual responses to items because it assumes that individual items have different difficulty or severity levels (Lo et al., 2015). Although these two analyses are preferred, they were not used to test the factor structure of most (10/15) of the instruments.

Hypothesis testing for construct validity was reported for all 15 instruments. However, the studies of most instruments (9/15) had imbalanced evidence for construct validity between convergent validity (i.e., analysing the correlations between the responses of the CM instrument under study and a comparator CM instrument) and discriminative validity (i.e., analysing the differences in responses between caregivers who maltreated their child and those who did not). Evidence on both convergent and discriminative validity was reported only for six instruments. In addition, most studies conducting hypothesis testing of instruments reported only a *t*-value or *F*-value to determine whether the responses between two groups, such as caregivers who maltreated their child and those who did not, were significantly different. Notably, both statistical values are dependent on sample size and do not explain the direction and/or magnitude of the difference (de Vet et al., 2011). To show the direction and magnitude of the difference between two groups regardless of the sample size, an effect size estimate such as Cohen's *d* needs to be calculated and reported (de Vet et al., 2011; Friedman, 1968).

The criterion validity in the comparison of a shortened version with the original long version was provided for only one instrument, the MCNS-SF, which is the shortened version of the MCNS. The correlation between the two versions was calculated, which is a preferred statistical method for establishing criterion validity in the COSMIN methodology. In addition, only one instrument (IPPS) was tested for cross-cultural validity, but incomplete information was provided on the measurement invariance of the instrument between two different groups. For good cross-cultural validity of an instrument regarding measurement invariance between culturally different groups in terms of gender, age, or socioeconomic status, evidence on either the instrument factor structures obtained from CFA (Gregorich, 2006) or the item difficulty or discrimination obtained from differential item functioning (DIF) analysis (Teresi et al., 2009) should be provided. However, none of the psychometric studies included in Paper 2 (Yoon et al., 2020b) reported preferred statistics on the measurement invariance between different groups by using either CFA or DIF analysis.

Within the domain of reliability (i.e., reliability, measurement error, and internal consistency), there were very large evidence gaps, except for internal consistency. Internal consistency was reported for most instruments (12/15) with the preferred statistic (i.e., Cronbach's α). None of the studies of the instruments provided any data on measurement error. Measurement error is clinically quite important because an instrument with a low error can sensitively detect clinically important changes, which can help the clinician determine when to either adjust or terminate treatment (Dvir, 2015; Guyatt et al., 1987). Of the four instruments reporting psychometric data on reliability (test–retest, interrater, and intrarater reliability), three reported different reliability statistics (e.g., Spearman's correlation coefficients and unweighted κ) from those preferred in the COSMIN methodology (Prinsen et al., 2018). The COSMIN methodology suggests the weighted κ or the intraclass correlation coefficient (ICC) as acceptable reliability statistical values. The ICC considers systematic error due to different test conditions and learning effects in repeated tests for continuous scales, while the Spearman's ρ coefficient does not (Scholtes et al., 2011). The weighted κ considers the extent of disagreement between the two raters for categorical scales, while the unweighted κ does not (Tang et al., 2015). However, the ICC was reported for only one instrument.

6.2.3 Responsiveness

Only a few of the included studies on the responsiveness of the included instruments tested the instruments' responsiveness through randomised controlled trials (RCTs), which allocate study samples to either an intervention or a control group randomly. RCTs help intervention studies minimise their selection bias and confounding variables (e.g., different sample characteristics; Altman, 1991). As a result, RCTs are recognised as the best study design for estimating the unbiased effect size of an intervention (Altman, 1991). However, most effectiveness studies on interventions for preventing CM were not designed based on RCTs due to practical (e.g., high cost) and ethical issues (e.g., socially sensitive research topics; van der Put et al., 2018). Therefore, the lack of RCTs is a methodological limitation in studies on the responsiveness of parent- or caregiver-report CM instruments.

Many studies on the responsiveness of the instruments tested the responsiveness with an inappropriate statistical method (e.g., the paired *t*-test or the repeated measures analysis of variance [ANOVA]), reporting only *p*-values (see Online Supplemental Table S6 of Paper 3 for details). The *p*-value is a less robust statistic of responsiveness (Mokkink, de Vet, et al., 2018) because it cannot explain whether the magnitude of the estimated mean difference is large enough to detect a clinically significant effect (i.e., clinical significance), and depends on sample size (Altman, 1991). For this reason, instead of a *p*-value, an effect size (e.g., Hedges' *g*; Hedges & Olkin, 2014) is suggested as a preferred measure of responsiveness in the COSMIN Risk of Bias checklist (Mokkink, Prinsen, et al., 2018), which provides information on clinical significance, regardless of the sample size (Altman, 1991). However, for most instruments, only *p*-values were reported based on paired *t*-tests or repeated-measure ANOVAs.

Last, there was generally either a lack of evidence or low-quality evidence on responsiveness. Only the Physical Abuse subscale of the ICAST-Trial had high-quality evidence of sufficient responsiveness among the overall scales or subscales of the 15 included instruments.

6.2.4 Evidence Gaps in the Recommendation of Instruments

No high-quality evidence for the content validity of most instruments (14/15) was reported (see Table 4 in Paper 1; Yoon et al., 2020a) because there were either missing data or a lack of robust evidence of the content validity (Yoon et al., 2020a). Evidence on the internal consistency of most instruments (14/15) either was not reported (NR) (see Table 4 in Paper 2; Yoon et al., 2020b) due to a lack of data on their internal consistency or was rated as indeterminate (?) due to a lack of data on their structural validity (Yoon et al., 2020b). Given the lack of evidence or low-quality evidence on both content validity and internal consistency, none of the 15 included instruments could be recommended as the most suitable for use (category A; see Table 5.1 in Section 5.4). To be the most suitable, the instruments should have both high-quality evidence for sufficient content validity and at least low-quality evidence for sufficient internal consistency (see Table 5.1 in Section 5.4). Moreover, as there was a lack of high-quality evidence to suggest that any of the psychometric properties are inherently insufficient (see Table 4 in Paper 1; Yoon et al., 2020a; Table 4 in Paper 2; Yoon et al., 2020b; Table 4 in Paper 3; Yoon et al., 2021), nine instruments might still have the potential to be recommended but would require further validation studies (category B; see Table 5.1 in Section 5.4). Last, six instruments could not be recommended (category C; see Table 5.1 in Section 5.4) because all but one (ICAST-Trial) had high-quality evidence supporting insufficient hypothesis testing, while the ICAST-Trial had high-quality evidence supporting both its insufficient internal consistency and hypothesis testing (see Table 4 in Paper 2; Yoon et al., 2020b). However, most of the hypothesis testing focused on convergent validity to test associations between different instruments rather than discriminative validity to test differences between groups (see Appendix F in Paper 2; Yoon et al., 2020b). For this reason, the evidence on the hypothesis testing of the six instruments provided only one side of the testing without evidence on discriminative validity.

Only the overall scales for the 15 included instruments were considered when recommending the most suitable parent- or caregiver-report CM instruments in this thesis. Paper 1 (Yoon et al., 2020a) and Paper 2 (Yoon et al., 2020b) evaluated the psychometric quality of the overall scales only (see Table 4 in Paper 1; Table 4 in Paper 2), while Paper 3 (Yoon et al., 2021) evaluated the overall scales and the unidimensional subscales (i.e., subscale[s] consisting of multiple items assessing a single underlying construct; de Vet et al., 2011; see Table 4 in Paper 3). Both the overall scales and the subscales tended to be used more in studies on the effectiveness of interventions than in studies on the construct validity, criterion validity, or reliability, which usually used the overall scales only. Therefore, the assessment of responsiveness in Paper 3 (Yoon et al., 2021) was conducted for all the overall scales and the unidimensional subscales thereof. The unidimensionality of a subscale was confirmed if data could be identified in the literature that supported the internal structure of the subscale (i.e., conducted either EFA or CFA and internal consistency using Cronbach's α for each subscale; Mokkink, de Vet, et al., 2018). The confirmed subscale can be used as an independent measure as an alternative to an overall scale; a convention sometimes used in studies to lessen participant burden (Mokkink, Prinsen, et al., 2018). As more data on both overall scales and the confirmed subscales were found for the responsiveness than the other psychometric properties (i.e., content validity, construct validity, criterion validity, and reliability), the quality assessment of responsiveness was conducted for both the scales and the subscales in Paper 3 (Yoon et al., 2021). However, the assessment of other psychometric properties was conducted only for the overall scales (Yoon et al., 2020a, 2020b). For this reason, the recommendations are limited to the overall scales of the 15 included instruments.

6.3 Challenges and Limitations

This subchapter is divided into two sections. Section 6.3.1 briefly discusses the methodological challenges of applying the COSMIN methodology. Section 6.3.2 presents the limitations of this thesis regarding the scope of the three reviews and using the old version of the PRISMA statement.

6.3.1 Challenges of the COSMIN Methodology

Several challenges were encountered in the application of the COSMIN methodology (Prinsen et al., 2018; Terwee, Prinsen, Chiarotto, Westerman, et al., 2018) in all three papers (Yoon et al., 2020a, 2020b, 2021). The first challenge was a lack of literature with meaningful information to assess content validity compared with other psychometric properties. A description of how items were generated in the development of a new instrument was seldom provided in the most of included articles in Paper 1 (Yoon et al., 2020a), which may have been due to word limits restricting specific description of the item-generation process. Second, even though the COSMIN group claims that their methodology is objective and standardised (Prinsen et al., 2018), rating the study quality and psychometric quality for content validity still required a certain degree of subjective judgement from the reviewers. For example, due to the lack of evidence regarding the content validity of most instruments in the studies included in Paper 1 (Yoon et al., 2020a), most of the overall ratings on content validity were determined based only on the reviewers' subjective opinions about the content validity of the instrument itself (i.e., items, response options, and instructions) according to the COSMIN manual on content validity (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018). Last, due to the comprehensiveness and complexity of the COSMIN manuals (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018), the time needed to assess both the study quality and psychometric quality was extensive. The challenge aligns with the claim of Kwok et al. (2021) that authors with graduate-level training in instrument development require at least 25 hours to complete the quality assessment for each instrument.

6.3.2 Limitations

The results of this thesis may have some limitations for the following reasons. First, only instruments originally developed and validated in English were identified due to a lack of language resources (e.g., professional translators). Thus, some results on psychometric properties of CM instruments developed and validated in other languages may have been missed. Second, the systematic literature search for this thesis used the old version of the PRISMA statement (Moher et al., 2009). Even though the PRISMA statement was recently updated and published in 2021, the updated statement (Page et al., 2021) was published after the submission and/or acceptance of the three reviews (Yoon et al., 2020a, 2020b, 2021) for this thesis. The updated PRISMA statement (Page et al., 2021) includes notable changes to help conduct and report a systematic review more transparently than the old version (Moher et al., 2009). For example, the updated PRISMA statement (Page et al., 2021) recommends reporting a detailed screening workflow for identifying eligible studies via both database searching and other methods (e.g., reference checking) in a PRISMA flow chart. However, only the workflow via database searching and the total number of studies identified via reference checking were reported in the PRISMA flow charts of the three reviews (Yoon et al., 2020a, 2020b, 2021), which were recommended by the old version of the PRISMA statement (Moher et al., 2009). Third, Paper 3 (Yoon et al., 2021) evaluated only one aspect of the construct approach for responsiveness by comparing change scores before and after intervention (Mokkink, de Vet, et al., 2018). The other two aspects (i.e., comparison with other outcome instruments and comparison between subgroups) were outside the scope of Paper 3 (Yoon et al., 2021) because of the scale, scope, and complexity of reporting. Fourth, the interpretability of change scores and the feasibility of instruments were beyond the scope of this thesis because these aspects are not considered psychometric properties within the COSMIN taxonomy.

6.4 Implications for Future Research

This subchapter discusses the implications for future research that are needed to overcome the methodological flaws and evidence gaps of the included studies for each psychometric property and for recommendation of the instruments presented in Subchapter 6.2. To discuss the implications of each psychometric property and recommendation, this subchapter is divided into four sections: content validity (6.4.1); construct validity, criterion validity, and reliability (6.4.2); responsiveness (6.4.3); and suggestions for promising and non-recommendable instruments (6.4.4).

6.4.1 Content Validity

Future studies on the development of new CM instruments that aim to generate new items should involve parents or caregivers to identify relevant, comprehensive, and comprehensible items based on their input on CM by using interviews or surveys. Moreover, further content validity studies are needed to assess the relevance, comprehensiveness, and comprehensibility of the included instruments because the currently available evidence on content validity is insufficient to make final recommendations. In particular, the comprehensibility of most of the instruments must be further assessed by gathering input from parents or caregivers. Last, future instrument development and content validity studies should follow the COSMIN manual (Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) in their study design and methodology when generating new instrument items and assessing the content validity of existing instrument items.

6.4.2 Construct Validity, Criterion Validity, and Reliability

Future studies on structural validity should perform factor analyses using CFA or IRT to determine the internal consistency of the nine instruments reported to have indeterminate internal consistency due to a lack of information on their structural validity. For cross-cultural validity, further studies should test measurement invariance across culturally different groups through CFA or DIF analysis. In terms of hypothesis testing for construct validity, future studies should calculate and report the effect sizes, such as Cohen's *d*, rather than *t*-values or *F*-values. Moreover, most of the included studies tended to evaluate convergent validity regarding the associations between two instruments with the same construct of interest rather than discriminative validity regarding the differences in scores between groups; thus, additional studies on discriminative validity are needed to balance the evidence with convergent validity in hypothesis testing. To obtain an overall picture of the reliability domain, further studies should assess all three aspects of reliability: internal consistency, measurement error, and reliability (test–retest, interrater, and intrarater reliability). For test–retest, interrater, and intrarater reliability, the ICC or weighted κ instead of Spearman's ρ or unweighted κ should be calculated and reported.

6.4.3 Responsiveness

Further studies on the responsiveness of parent- or caregiver-report CM instruments should analyse and report the effect sizes to estimate mean differences before and after parental interventions. Moreover, to estimate an unbiased mean difference (effect size) as a measure of responsiveness, more RCT-designed studies using the parent- or caregiver-report CM instruments need to be conducted. All but one subscale (the Physical Abuse subscale of the ICAST-Trial) of the 15 included instruments require further studies on their responsiveness due to a lack of evidence or low-quality evidence. However, the Physical Abuse subscale of the ICAST-Trial could be recommended for use in parenting interventions to reduce the physical abuse of children due to high-quality evidence that the subscale has sufficient responsiveness.

6.4.4 Suggestions for the Promising and Non-recommendable Instruments

The nine promising instruments (in category B) require further validation studies on one or more psychometric properties to confirm whether they can be recommended (i.e., category A). To meet the criterion for category A, the content validity, internal consistency, and/or structural validity of all nine instruments need to be further assessed because additional results from future studies on all three psychometric properties may change the overall quality ratings of the evidence.

To confirm that the six non-recommendable instruments (category C) are indeed not to be recommended, additional validation studies on hypothesis testing and/or internal consistency should be conducted. Further studies on hypothesis testing could change the recommendation of all except one instrument (ICAST-Trial) from not recommendable (category C) to promising (category B). For the ICAST-Trial, both its hypothesis testing and internal consistency should be further evaluated in future psychometric studies. If further studies provide more evidence for sufficient hypothesis testing and/or internal consistency, the six non-recommendable instruments (category C) could be recommended as promising (category B), but they would still require further validation. If these six instruments could be moved from category C to category B, and if further studies on the content validity and internal consistency of the instruments provide sufficient evidence to meet the category A criteria (high-quality evidence for sufficient content validity and at least low-quality evidence for sufficient internal consistency), the instruments can also be recommended as the most suitable instruments (category A).

6.5 Implications for Future Practice

This subchapter begins by discussing the social ecological model for measuring CM (Section 6.5.1) to highlight the following four issues: the positioning of the included CM instruments within the systems of the social ecological model; the potential to use the instruments for culturally different parents or caregivers within the same system where the instruments are positioned; the potential to use the instruments in other systems in addition to the current system where they are positioned; and the contribution of the instruments to measuring attitudes towards CM to more accurately estimate the prevalence of CM at the population level. Next, the implications of the use of the included CM instruments for implementation of public health approaches to preventing CM (Section 6.5.2). Finally, the implications of the COSMIN methodology for future evidence-based assessment practice for monitoring CM prevention are discussed (Section 6.5.3).

6.5.1 Social Ecological Model for Measuring CM

The 15 included instruments measuring maltreating behaviour or attitudes towards CM were designed for use at the microsystem level of the social ecological model (see Figure 6.1) for two reasons: (1) the target population of interest in this thesis was parents or caregivers; and (2) the included studies used the instruments only with their study samples of parents or caregivers who were at risk of perpetrating or who were perpetrating CM (see Online Appendix C in Paper 1; Online Appendix C of Paper 2; Online Supplemental Table S5 in Paper 3). That is, the included studies did not use the included instruments with the general population of parents or caregivers (i.e., exosystem or macrosystem levels) or professionals (i.e., mesosystem level). Therefore, the CM instruments included in this thesis can be used for research and clinical practice for parents who are at risk or have a history of CM perpetration.

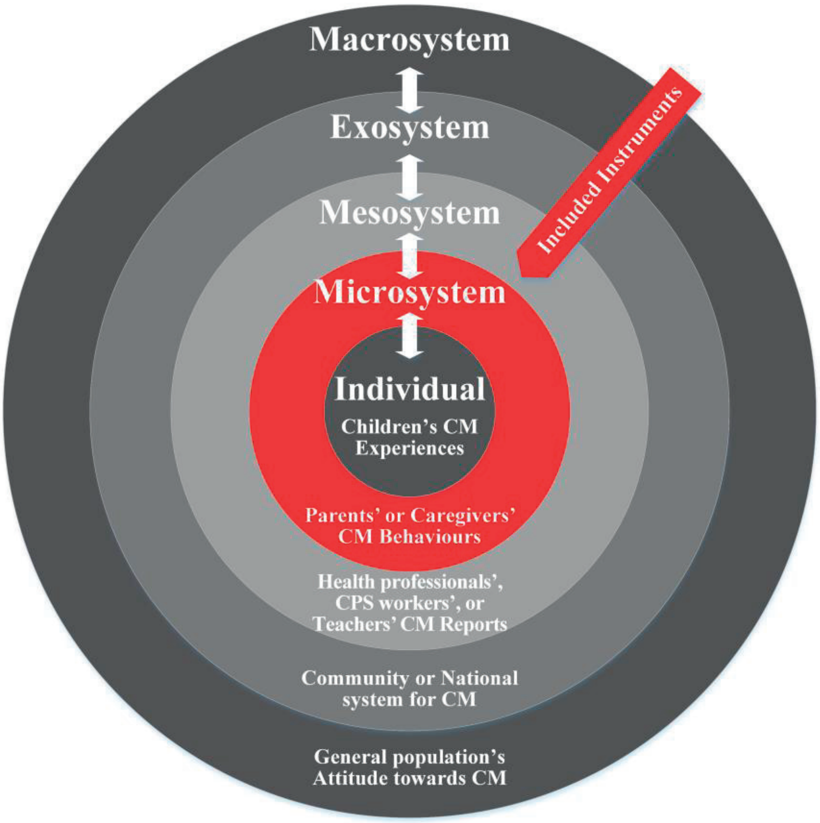

Macrosystem
Exosystem
Mesosystem
Microsystem
Individual
Children's CM Experiences
Parents' or Caregivers' CM Behaviours
Health professionals', CPS workers', or Teachers' CM Reports
Community or National system for CM
General population's Attitude towards CM
Included Instruments

Figure 6.1. *Position of the* Parent- *or Caregiver-Report CM Instruments Included in this Thesis in the Social Ecological Model for Measuring CM*

Within the microsystem, the instruments included in this thesis may apply to culturally different groups, which include different language groups as well as different cultural groups using the same language. However, language and cultural differences in the macrosystem may cause individual parents or caregivers in the microsystem to interpret the same parenting behaviours differently. For instance, 'spanking' may be perceived as CM to parents in New Zealand but as a form of discipline to parents in the U.S. Corporal punishment is illegal (in all settings) in New Zealand, while it is legal if conducted at home in the U.S. (Elgar et al., 2018). This difference between the two English-speaking countries shows how cultural differences may result in different underlying constructs of the same instrument. Thus, applying the same instruments to different cultural groups requires testing the measurement invariance across the different groups despite their use of the same language. In addition, when applying the translated instruments to different language groups, the measurement invariance should also be tested in terms of cross-cultural validity.

Across the systems, the included CM instruments used for parents or caregivers within the microsystem may also be used for either professionals (within the mesosystem) or the entire population (within the exosystem). If the items of the included instruments were appropriately modified to measure suspected CM by asking either professionals or the entire population, the modified items could be applied to either professional-report instruments or population-level questionnaires to measure CM. However, before the modified items are directly applied to professionals or the general population, they should be tested for their content validity to determine whether the modified items are relevant, comprehensive, and comprehensible to professionals or the general population, and they should be tested for their measurement invariance to determine whether the measured scores are not significantly different from those obtained from parents or caregivers. In this respect, the CM instruments included in this thesis may need to be modified for application to professionals or the general population first, and then tested for both the content validity and measurement invariance of the modified items for professionals or the general population. Through modification and further validation, the included CM instruments for parents or caregivers in microsystems may have the potential to be used as CM instruments for professionals in the mesosystem or the general population in the exosystem.

To connect the different systems, the included instruments measuring parents' or caregivers' *attitudes towards CM* within the microsystem can contribute to accurately estimating CM prevalence at the national level (i.e., exosystem). CM occurring within a single country (i.e., the national prevalence in the exosystem) is influenced by its citizens' attitudes towards CM (i.e., public attitude at the macrosystem level). In particular, physical punishment of children tends to be used more frequently in countries where the citizens have more accepting attitudes towards the use of corporal punishment for disciplining children than in countries with less accepting attitudes. In addition, if citizens (i.e., the general population at the macrosystem level) have less accepting attitudes towards CM, then they will more actively report suspected CM cases, resulting in a more accurate estimation of CM prevalence. For this reason, improving the general population's attitudes towards CM by implementing an evidence-based intervention is important; the evidence-based intervention for changing attitudes towards CM can be established based on the selection and use of accurate and reliable instruments for measuring their attitudes towards CM at the population level. Therefore, if the instruments for measuring attitudes towards CM included in this thesis can be used for the general population through modification and further validation, the modified instruments could also contribute to the more accurate estimation of the national prevalence of CM within the exosystem.

6.5.2 Public Health Approaches to Preventing CM

As discussed in the previous subchapter, if the CM instruments included in this thesis can be used to investigate the national prevalence of CM or the effectiveness of CM interventions across countries through content modification and further validation, the modified and validated instruments could contribute to each of the four steps (see Figure 2.1) in public health approaches to preventing CM. To define the CM problem (Step 1), the recommendation of the CM instruments in this thesis can contribute towards the selection of the most suitable instruments for accurately estimating the current status of CM prevalence, which can help identify subgroups of parents at high risk of maltreating behaviours. Collecting the demographic information (e.g., ethnicity and socioeconomic data) of this high-risk subgroup of parents can also help identify risk and protective factors (Step 2). Furthermore, the accurate identification of risk and protective factors of CM can contribute to determining which factors should be considered in the development of a new CM intervention (Step 3). Finally, the results of Paper 3 (Yoon et al., 2021) on responsiveness can support studies on the effectiveness (Step 3) and implementation (Step 4) of CM interventions by use of the recommended, most sensitive instruments in detecting the reduction of parental maltreating behaviours or attitudes towards CM before and after CM interventions.

6.5.3 Evidence-Based Assessment for Monitoring CM Prevention

In terms of parenting interventions for preventing CM, many clinicians tend to use instruments based on the instruments' popularity in most clinical practices rather than the quality of the instrument's psychometric properties (Meinck et al., 2018; Meinck et al., 2016). For example, most of the identified studies on responsiveness in this thesis measured the effectiveness of parenting interventions to prevent CM with the most widely used AAPI-2 or CTSPC. However, the evidence on the responsiveness of these popular instruments was not of sufficient quality to recommend them for use in CM interventions. The frequent use of CM instruments with low-quality evidence can hamper the use of evidence-based interventions (Meinck et al., 2018). Thus, selecting and using non-evidence-based assessment instruments can lead to either the underestimation or the overestimation of an intervention's effectiveness which, in turn, can lead to the use of ineffective interventions.

However, the COSMIN method used in this thesis can contribute to the selection of the best evidence-based assessment instruments to establish evidence-based interventions for CM prevention (Meinck et al., 2018; Meinck et al., 2016). To be selected as an evidence-based instrument, they must have good content validity, other psychometric properties, and interpretability and feasibility (Hunsley & Mash, 2007). The criteria for good content validity and other properties have been suggested to evaluate the psychometric quality of instruments on CM using the COSMIN methodology (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018), which provides standardised criteria for good psychometric properties. Moreover, using both the PRISMA statement (Moher et al., 2009) for the systematic literature search and the COSMIN method (Mokkink, Prinsen, et al., 2018; Terwee, Prinsen, Chiarotto, de Vet, et al., 2018) for the evaluation of psychometric properties can contribute greatly to conducting better systematic reviews to evaluate and recommend child- or professional-report CM instruments as well as survey questionnaires on CM at the population level. Finally, the three reviews included in this thesis contribute greatly to developing evidence-based instruments for monitoring CM prevention. The COSMIN Risk of Bias checklist (Mokkink, de Vet, et al., 2018) presents criteria for research design and statistical methods that should be considered in the development of quality assessment instruments for parental interventions to prevent CM.

6.6 Concluding Remarks

This thesis is the first systematic review to provide a synthesis of validity, reliability, and responsiveness evidence for available parent- or caregiver-report instruments on CM. Fifteen instruments were identified and evaluated, of which the majority had limited and lower-quality evidence concerning psychometric properties. Due to lacking and low-quality evidence, none of the identified instruments can be recommended as the most suitable for use in clinical practice and research. Only nine instruments (AAPI-2, APT, CNS-MMS, CTS-ES, FM-CA, IPPS, P-CAAM, PRCM, and SBS-SV) were recommended as promising based on the available psychometric evidence, but they still require further validation before firm recommendation as the most suitable instrument can be made.

The significance of this review lies in the fact that parent- or caregiver-report CM instruments have been used most frequently within a range of CPS and within research studies to investigate and prevent CM, especially for young children who are the main victims of CM (Meinck et al., 2016). However, the psychometric quality of these instruments remains poor and understudied. To overcome this challenge, future studies aimed at developing new instruments and validating existing instruments should follow the COSMIN guidelines to help researchers and clinicians select the most suitable parent- or caregiver-report instruments on CM.

References

Abedi, A., Prinsen, C. A. C., Shah, I., Buser, Z., & Wang, J. C. (2019, 2019/08/09). Performance properties of health-related measurement instruments in whiplash: systematic review protocol. *Systematic Reviews, 8*(1), 199. https://doi.org/10.1186/s13643-019-1119-0

Achenbach, T. M. (2017, Jan-Feb). Future Directions for Clinical Research, Services, and Training: Evidence-Based Assessment Across Informants, Cultures, and Dimensional Hierarchies. *Journal of Clinical Child & Adolescent Psychology, 46*(1), 159-169. https://doi.org/10.1080/15374416.2016.1220315

Al-Eissa, M. A., AlBuhairan, F. S., Qayad, M., Saleheen, H., Runyan, D., & Almuneef, M. (2015, 2015/04/01/). Determining child maltreatment incidence in Saudi Arabia using the ICAST-CH: A pilot study. *Child Abuse & Neglect, 42*, 174-182. https://doi.org/10.1016/j.chiabu.2014.08.016

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

Altmann, T. K. (2008, 2008/07/01). Attitude: A Concept Analysis. *Nursing Forum, 43*(3), 144-150. https://doi.org/10.1111/j.1744-6198.2008.00106.x

American Psychiatric Association (Ed.). (2013). *Diagnostic and statistical manual of mental disorders* (5 ed.). American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425596.

Anda, R. F., Brown, D. W., Dube, S. R., Bremner, J. D., Felitti, V. J., & Giles, W. H. (2008, 2008/05/01/). Adverse Childhood Experiences and Chronic Obstructive Pulmonary Disease in Adults. *American Journal of Preventive Medicine, 34*(5), 396-403. https://doi.org/10.1016/j.amepre.2008.02.002

Anda, R. F., Butchart, A., Felitti, V. J., & Brown, D. W. (2010, 2010/07/01/). Building a Framework for Global Surveillance of the Public Health Implications of Adverse Childhood Experiences. *American Journal of Preventive Medicine, 39*(1), 93-98. https://doi.org/10.1016/j.amepre.2010.03.015

Aromataris, E., & Munn, Z. (Eds.). (2020). *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/JBIMES-20-01.

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*(3), 1-6. https://doi.org/10.17795/intjsh-31198

Ashton, V. (2001, 2001/03/01/). The relationship between attitudes toward corporal punishment and the perception and reporting of child maltreatment. *Child Abuse & Neglect, 25*(3), 389-399. https://doi.org/10.1016/S0145-2134(00)00258-1

Ateah, C. A., & Durrant, J. E. (2005, 2005/02/01/). Maternal use of physical punishment in response to child misbehavior: implications for child abuse prevention. *Child Abuse & Neglect, 29*(2), 169-185. https://doi.org/10.1016/j.chiabu.2004.10.010

Bae, H.-o., & Kindler, H. (2017, 2017/04/01/). Child maltreatment re-notifications in Germany: Analysis of local case files. *Children and Youth Services Review, 75*, 42-49. https://doi.org/10.1016/j.childyouth.2017.02.012

Banyard, V. L., & Williams, L. M. (2007, Mar). Women's voices on recovery: a multi-method study of the complexity of recovery from child sexual abuse. *Child Abuse & Neglect, 31*(3), 275-290. https://doi.org/10.1016/j.chiabu.2006.02.016

Barnett, D., Manly, J. T., & Cicchetti, D. (1993). Defining Child Maltreatment: the interface between policy and research. In D. Cicchetti & S. L. Tooth (Eds.), *Child abuse, Child development, and social policy* (pp. 7-73). Ablex.

Belsky, J. (1993). Etiology of child maltreatment: A developmental€cological analysis. *Psychological Bulletin, 114*(3), 413-434. https://doi.org/10.1037/0033-2909.114.3.413

Ben-David, V., & Jonson-Reid, M. (2017). Resilience among adult survivors of childhood neglect: A missing piece in the resilience literature. *Children and Youth Services Review, 78*, 93-103. https://doi.org/10.1016/j.childyouth.2017.05.014

Bower-Russa, M. (2005, 2005/08/01). Attitudes Mediate the Association Between Childhood Disciplinary History and Disciplinary Responses. *Child Maltreatment, 10*(3), 272-282. https://doi.org/10.1177/1077559505277531

Brent, D. A., Oquendo, M., Birmaher, B., Greenhill, L., Kolko, D., Stanley, B., Zelazny, J., Brodsky, B., Melhem, N., Ellis, S. P., & Mann, J. J. (2004, 2004/10/01/). Familial Transmission of Mood Disorders: Convergence and Divergence With Transmission of Suicidal Behavior. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(10), 1259-1266. https://doi.org/10.1097/01.chi.0000135619.38392.78

Brockington, I. F., Oates, J., George, S., Turner, D., Vostanis, P., Sullivan, M., Loh, C., & Murdoch, C. (2001, 2001/03/01). A Screening Questionnaire for mother-infant bonding disorders. *Archives of Women's Mental Health, 3*(4), 133-140. https://doi.org/10.1007/s007370170010

Brodsky, B. S., Mann, J. J., Stanley, B., Tin, A., Oquendo, M., Birmaher, B., Greenhill, L., Kolko, D., Zelazny, J., Burke, A. K., Melhem, N. M., & Brent, D. (2008, Apr).

Familial transmission of suicidal behavior: factors mediating the relationship between childhood abuse and offspring suicide attempts. *The Journal of clinical psychiatry, 69*(4), 584-596. https://doi.org/10.4088/jcp.v69n0410

Bronfenbrenner, U. (1992). *Ecological systems theory*. Jessica Kingsley Publishers.

Buss, C., Entringer, S., Moog, N. K., Toepfer, P., Fair, D. A., Simhan, H. N., Heim, C. M., & Wadhwa, P. D. (2017, May). Intergenerational Transmission of Maternal Childhood Maltreatment Exposure: Implications for Fetal Brain Development. *Journal of the American Academy of Child & Adolescent Psychiatry, 56*(5), 373-382. https://doi.org/10.1016/j.jaac.2017.03.001

Cañas, M., Ibabe, I., & De Paúl, J. (2020, 2020/11/01/). Promising observational instruments of parent-child (0–12 years) interaction within the child protection system: A systematic review. *Child Abuse & Neglect, 109*, 104713. https://doi.org/10.1016/j.chiabu.2020.104713

Chavis, A., Hudnut-Beumler, J., Webb, M. W., Neely, J. A., Bickman, L., Dietrich, M. S., & Scholer, S. J. (2013, 2013/12/01/). A brief intervention affects parents' attitudes toward using less physical punishment. *Child Abuse & Neglect, 37*(12), 1192-1201. https://doi.org/10.1016/j.chiabu.2013.06.003

Chen, M., & Chan, K. L. (2015, 2016/01/01). Effects of Parenting Programs on Child Maltreatment Prevention: A Meta-Analysis. *Trauma, Violence, & Abuse, 17*(1), 88-104. https://doi.org/10.1177/1524838014566718

Child Welfare Information Gateway. (2019). *Definitions of Child Abuse and Neglect*. U.S. Department of Health and Human Services, Children's Bureau. https://www.childwelfare.gov/pubPDFs/define.pdf#page=2&view=Defining%20child%20abuse%20or%20neglect%20in%20State%20law

Christian, B., Armstrong, R., Calache, H., Carpenter, L., Gibbs, L., & Gussy, M. (2019, 2019/03/01). A systematic review to assess the methodological quality of studies on measurement properties for caries risk assessment tools for young children. *International Journal of Paediatric Dentistry, 29*(2), 106-116. https://doi.org/10.1111/ipd.12446

Cicchetti, D., & Toth, S. L. (2005). Child Maltreatment. *Annual Review of Clinical Psychology, 1*(1), 409-438. https://doi.org/10.1146/annurev.clinpsy.1.102803.144029

Cluver, L., Meinck, F., Yakubovich, A., Doubt, J., Redfern, A., Ward, C., Salah, N., De Stone, S., Petersen, T., Mpimpilashe, P., Romero, R. H., Ncobo, L., Lachman, J., Tsoanyane, S., Shenderovich, Y., Loening, H., Byrne, J., Sherr, L., Kaplan, L., &

Gardner, F. (2016, 2016/07/13). Reducing child abuse amongst adolescents in low- and middle-income countries: A pre-post trial in South Africa. *BMC Public Health, 16*(1), 567. https://doi.org/10.1186/s12889-016-3262-z

Collishaw, S., Dunn, J., O'Connor, T. G., & Golding, J. (2007, Spring). Maternal childhood abuse and offspring adjustment over time. *Development and Psychopathology, 19*(2), 367-383. https://doi.org/10.1017/s0954579407070186

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, 2017/02/01). Parent-Child Agreement on Parent-to-Child Maltreatment. *Journal of Family Violence, 32*(2), 207-217. https://doi.org/10.1007/s10896-016-9902-3

Cooley, D. T., & Jackson, Y. (2020). Informant Discrepancies in Child Maltreatment Reporting: A Systematic Review. *Child Maltreatment*. https://doi.org/10.1177/1077559520966387

Cordier, R., Speyer, R., Chen, Y.-W., 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*(7), e0132299. https://doi.org/10.1371/journal.pone.0132299

Corso, P. S., Edwards, V. J., Fang, X., & Mercy, J. A. (2008, 2008/06/01). Health-Related Quality of Life Among Adults Who Experienced Maltreatment During Childhood. *American Journal of Public Health, 98*(6), 1094-1100. https://doi.org/10.2105/AJPH.2007.119826

de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: a practical guide*. Cambridge university press.

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006, 2006/10/01/). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*(10), 1033-1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

Del Vecchio, T., Erlanger, A. C. E., & Slep, A. M. S. (2012). Theories of child abuse. In M. A. Fine & F. D. Fincham (Eds.), *Handbook of Family Theories: A Contentbased Approach*. Taylor and Francis/Routledge.

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., Butron Riveros, B. C., Watts, C., & Abrahams, N. (2018). Who perpetrates violence against children? A systematic analysis of age-specific and sex-specific data. *BMJ Paediatrics Open, 2*(1), e000180. https://doi.org/10.1136/bmjpo-2017-000180

Dhingra, K., Boduszek, D., & Sharratt, K. (2015, 2016/09/01). Victimization Profiles, Non-Suicidal Self-Injury, Suicide Attempt, and Post-Traumatic Stress Disorder Symptomology: Application of Latent Class Analysis. *Journal of Interpersonal Violence, 31*(14), 2412-2429. https://doi.org/10.1177/0886260515576967

Diez-Roux, A. V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health, 21*, 171-192. https://doi.org/10.1146/annurev.publhealth.21.1.171

Dvir, Z. (2015, 4). Difference, significant difference and clinically meaningful difference: The meaning of change in rehabilitation. *Journal of Exercise Rehabilitation, 11*(2), 67-73. https://doi.org/10.12965/jer.150199

Elgar, F. J., Donnelly, P. D., Michaelson, V., Gariépy, G., Riehm, K. E., Walsh, S. D., & Pickett, W. (2018). Corporal punishment bans and physical fighting in adolescents: an ecological study of 88 countries. *BMJ Open, 8*(9), e021616. https://doi.org/10.1136/bmjopen-2018-021616

Euser, E. M., van Ijzendoorn, M. H., Prinzie, P., & Bakermans-Kranenburg, M. J. (2010, 2010/02/01). Prevalence of Child Maltreatment in the Netherlands. *Child Maltreatment, 15*(1), 5-17. https://doi.org/10.1177/1077559509345904

Fallon, B., Trocmé, N., Fluke, J., MacLaurin, B., Tonmyr, L., & Yuan, Y.-Y. (2010, 2010/01/01/). Methodological challenges in measuring child maltreatment. *Child Abuse & Neglect, 34*(1), 70-79. https://doi.org/10.1016/j.chiabu.2009.08.008

Fang, X., Fry, D. A., Brown, D. S., Mercy, J. A., Dunne, M. P., Butchart, A. R., Corso, P. S., Maynzyuk, K., Dzhygyr, Y., Chen, Y., McCoy, A., & Swales, D. M. (2015, 2015/04/01/). The burden of child maltreatment in the East Asia and Pacific region. *Child Abuse & Neglect, 42*, 146-162. https://doi.org/10.1016/j.chiabu.2015.02.012

Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., Koss, M. P., & Marks, J. S. (1998, May). 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*(4), 245-258. https://doi.org/10.1016/s0749-3797(98)00017-8

Finkelhor, D., Ormrod, R. K., & Turner, H. A. (2007, 2007/01/01/). Poly-victimization: A neglected component in child victimization. *Child Abuse & Neglect, 31*(1), 7-26. https://doi.org/10.1016/j.chiabu.2006.06.008

Finkelhor, D., Shattuck, A., Turner, H. A., & Hamby, S. L. (2014, 2014/09/01/). The Lifetime Prevalence of Child Sexual Abuse and Sexual Assault Assessed in Late Adolescence. *Journal of Adolescent Health, 55*(3), 329-333. https://doi.org/10.1016/j.jadohealth.2013.12.026

Fisher, B. S. (2008, 2009/02/01). The Effects of Survey Question Wording on Rape Estimates: Evidence From a Quasi-Experimental Design. *Violence Against Women, 15*(2), 133-147. https://doi.org/10.1177/1077801208329391

Fisher, B. S. (2009, Feb). The effects of survey question wording on rape estimates: evidence from a quasi-experimental design. *Violence Against Women, 15*(2), 133-147. https://doi.org/10.1177/1077801208329391

Fluke, J. D., Tonmyr, L., Gray, J., Bettencourt Rodrigues, L., Bolter, F., Cash, S., Jud, A., Meinck, F., Casas Muñoz, A., O'Donnell, M., Pilkington, R., & Weaver, L. (2020, 2020/08/26/). Child maltreatment data: A summary of progress, prospects and challenges. *Child Abuse & Neglect*, 104650. https://doi.org/10.1016/j.chiabu.2020.104650

Foster, R. H., Olson-Dorff, D., Reiland, H. M., & Budzak-Garza, A. (2017, 2017/05/01/). Commitment, confidence, and concerns: Assessing health care professionals' child maltreatment reporting attitudes. *Child Abuse & Neglect, 67*, 54-63. https://doi.org/10.1016/j.chiabu.2017.01.024

Francis, D. O., McPheeters, M. L., Noud, M., Penson, D. F., & Feurer, I. D. (2016, 2016/08/02). Checklist to operationalize measurement characteristics of patient-reported outcome measures. *Systematic Reviews, 5*(1), 129. https://doi.org/10.1186/s13643-016-0307-4

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70*(4), 245-251. https://doi.org/10.1037/h0026258

Fry, D., Fang, X., Elliott, S., Casey, T., Zheng, X., Li, J., Florian, L., & McCluskey, G. (2018, Jan). The relationships between violence in childhood and educational outcomes: A global systematic review and meta-analysis. *Child Abuse Negl, 75*, 6-28. https://doi.org/10.1016/j.chiabu.2017.06.021

Gershoff, E. T. (2013). Spanking and Child Development: We Know Enough Now To Stop Hitting Our Children. *Child development perspectives, 7*(3), 133-137. https://doi.org/10.1111/cdep.12038

Gershoff, E. T., Lee, S. J., & Durrant, J. E. (2017, 2017/09/01/). Promising intervention strategies to reduce parents' use of physical punishment. *Child Abuse & Neglect, 71*, 9-23. https://doi.org/10.1016/j.chiabu.2017.01.017

Gibson, M. (2015). Shame and guilt in child protection social work: new interpretations and opportunities for practice. *Child & Family Social Work, 20*(3), 333-343. https://doi.org/10.1111/cfs.12081

Gilbert, R., Kemp, A., Thoburn, J., Sidebotham, P., Radford, L., Glaser, D., & Macmillan, H. L. (2009, Jan 10). Recognising and responding to child maltreatment. *Lancet, 373*(9658), 167-180. https://doi.org/10.1016/s0140-6736(08)61707-9

Gilbert, R., Widom, C. S., Browne, K., Fergusson, D., Webb, E., & Janson, S. (2009, 2009/01/03/). Burden and consequences of child maltreatment in high-income countries. *The Lancet, 373*(9657), 68-81. https://doi.org/10.1016/S0140-6736(08)61706-7

Greco, A. M., Guilera, G., & Pereda, N. (2017, 2017/10/01/). School staff members experience and knowledge in the reporting of potential child and youth victimization. *Child Abuse & Neglect, 72*, 22-31. https://doi.org/10.1016/j.chiabu.2017.07.004

Gregorich, S. E. (2006). Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups?: Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework. *Medical Care, 44*(11), S78-S94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Guyatt, G., Walter, S., & Norman, G. (1987, 1987/01/01/). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40*(2), 171-178. https://doi.org/10.1016/0021-9681(87)90069-5

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

Hamby, S., Finkelhor, D., Turner, H., & Ormrod, R. (2010, 2010/10/01/). The overlap of witnessing partner violence with child maltreatment and other victimizations in a nationally representative survey of youth. *Child Abuse & Neglect, 34*(10), 734-741. https://doi.org/10.1016/j.chiabu.2010.03.001

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

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*(3), e20154079. https://doi.org/10.1542/peds.2015-4079

Holden, G. W., Brown, A. S., Baldwin, A. S., & Croft Caderao, K. (2014, 2014/05/01/). Research findings can change attitudes about corporal punishment. *Child Abuse & Neglect, 38*(5), 902-908. https://doi.org/10.1016/j.chiabu.2013.10.013

Holden, G. W., & Buck, M. J. (2002). Parental attitudes toward child rearing. In M. H. Bornstein (Ed.), *Handbook of parenting* (2 ed., pp. 537–562). Lawrence Erlbaum.

Hovdestad, W., Campeau, A., Potter, D., & Tonmyr, L. (2015). A systematic review of childhood maltreatment assessments in population-representative surveys since 1990. *PLoS One, 10*(5), e0123366. https://doi.org/10.1371/journal.pone.0123366

Huffhines, L., Tunno, A. M., Cho, B., Hambrick, E. P., Campos, I., Lichty, B., & Jackson, Y. (2016, 2016/08/01/). Case file coding of child maltreatment: Methods, challenges, and innovations in a longitudinal project of youth in foster care. *Children and Youth Services Review, 67*, 254-262. https://doi.org/10.1016/j.childyouth.2016.06.019

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., & Dunne, M. P. (2017, Aug). The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. *Lancet Public Health, 2*(8), e356-e366. https://doi.org/10.1016/s2468-2667(17)30118-4

Hunsley, J., & Mash, E. J. (2007). Evidence-Based Assessment. *Annual Review of Clinical Psychology, 3*(1), 29-51. https://doi.org/10.1146/annurev.clinpsy.3.022806.091419

Institute of Medicine and National Research Council. (2014). Describing the problem. In A. C. Petersen, J. Joseph, & M. Feit (Eds.), *New directions in child abuse and neglect research* (pp. 31–68). National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK195982/

Ireland, T. O., Smith, C. A., & Thornberry, T. P. (2002, 2002/05/01). DEVELOPMENTAL ISSUES IN THE IMPACT OF CHILD MALTREATMENT ON LATER DELINQUENCY AND DRUG USE*. *Criminology, 40*(2), 359-400. https://doi.org/10.1111/j.1745-9125.2002.tb00960.x

Jabraeili, M., Asadollahi, M., Asghari Jafarabadi, M., & Hallaj, M. (2015, 2015/3/1). Attitude toward Child Abuse among Mothers Referring Health Centers of Tabriz. *Journal of Caring Sciences, 4*(1), 75-82. https://doi.org/10.5681/jcs.2015.008

Jackson, S., Thompson, R. A., Christiansen, E. H., Colman, R. A., Wyatt, J., Buckendahl, C. W., Wilcox, B. L., & Peterson, R. (1999, Jan). Predicting abuse-prone parental attitudes and discipline practices in a nationally representative sample. *Child Abuse & Neglect, 23*(1), 15-29. https://doi.org/10.1016/s0145-2134(98)00108-2

Jaffee, S. R., Caspi, A., Moffitt, T. E., & Taylor, A. (2004). Physical Maltreatment Victim to Antisocial Child: Evidence of an Environmentally Mediated Process. *Journal of Abnormal Psychology, 113*(1), 44-55. https://doi.org/10.1037/0021-843X.113.1.44

Janson, S. (2018). Epidemiological Studies of Child Maltreatment: Difficulties and Possibilities. In W. Kiess, C.-G. Bornehag, & C. Gennings (Eds.), *Pediatric Epidemiology* (Vol. 21, pp. 16-29). Karger. https://doi.org/10.1159/000481320

Johnson, C. F. (2002, 2002/10/01). Child maltreatment 2002: Recognition, reporting and risk. *Pediatrics International, 44*(5), 554-560. https://doi.org/10.1046/j.1442-200X.2002.01642.x

Jones, D. A., Trudinger, P., & Crawford, M. (2004, 2004/08/01). Intelligence and achievement of children referred following sexual abuse. *Journal of Paediatrics and Child Health, 40*(8), 455-460. https://doi.org/10.1111/j.1440-1754.2004.00427.x

Jones, L., Bellis, M. A., Wood, S., Hughes, K., McCoy, E., Eckley, L., Bates, G., Mikton, C., Shakespeare, T., & Officer, A. (2012, 2012/09/08/). Prevalence and risk of violence against children with disabilities: a systematic review and meta-analysis of observational studies. *The Lancet, 380*(9845), 899-907. https://doi.org/10.1016/S0140-6736(12)60692-8

Kim, S., Yoo, J. P., Chin, M., Jang, H. J., Kim, H.-S., Lee, S.-G., & Lee, B. J. (2019, 2019/02/01). Factors influencing resubstantiation of child neglect in South Korea: A comparison with other maltreatment cases. *Asian Social Work and Policy Review, 13*(1), 100-116. https://doi.org/10.1111/aswp.12160

Kwok, E. Y. L., Rosenbaum, P., Thomas-Stonell, N., & Cunningham, B. J. (2021, 2021/03/01). Strengths and challenges of the COSMIN tools in outcome measures appraisal: A case example for speech–language therapy. *International Journal of Language & Communication Disorders, 56*(2), 313-329. https://doi.org/10.1111/1460-6984.12603

Lamb, M. E., Orbach, Y., Warren, A. R., Esplin, P. W., & Hershkowitz, I. (2007). Enhancing performance: Factors affecting the informativeness of young witnesses. In R. C. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for events* (Vol. 1, pp. 429–451). Erlbaum.

Laurin, J., Wallace, C., Draca, J., Aterman, S., & Tonmyr, L. (2018). Youth self-report of child maltreatment in representative surveys: A systematic review. *Health Promotion* *and Chronic Disease Prevention in Canada: Research, Policy and Practice, 38*(2), 37–54. https://doi.org/10.24095/hpcdp.38.2.01

Leitzke, B. T., & Pollak, S. D. (2017). Child Maltreatment: Consequences, Mechanisms, and Implications for Parenting. In K. Deater-Deckard & R. Panneton (Eds.), *Parental Stress and Early Child Development: Adaptive and Maladaptive Outcomes* (pp. 209-234). Springer.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ, 339*, b2700. https://doi.org/10.1136/bmj.b2700

Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., Rudan, I., Campbell, H., Cibulskis, R., Li, M., Mathers, C., & Black, R. E. (2012, Jun 9). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet, 379*(9832), 2151-2161. https://doi.org/10.1016/s0140-6736(12)60560-1

Lo, C., Liang, W.-M., Hang, L.-W., Wu, T.-C., Chang, Y.-J., & Chang, C.-H. (2015, 2015/08/20). A psychometric assessment of the St. George’s respiratory questionnaire in patients with COPD using rasch model analysis. *Health and Quality of Life Outcomes, 13*(1), 131. https://doi.org/10.1186/s12955-015-0320-7

Lucas, N. P., Macaskill, P., Irwig, L., & Bogduk, N. (2010, 2010/08/01/). The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology, 63*(8), 854-861. https://doi.org/10.1016/j.jclinepi.2009.10.002

Manly, J. T. (2005, 2005/05/01/). Advances in research definitions of child maltreatment. *Child Abuse & Neglect, 29*(5), 425-439. https://doi.org/10.1016/j.chiabu.2005.04.001

Mash, E. J., & Hunsley, J. (2005, 2005/08/01). Evidence-Based Assessment of Child and Adolescent Disorders: Issues and Challenges. *Journal of Clinical Child & Adolescent Psychology, 34*(3), 362-379. https://doi.org/10.1207/s15374424jccp3403_1

Mathews, B., Pacella, R., Dunne, M. P., Simunovic, M., & Marston, C. (2020). Improving measurement of child abuse and neglect: A systematic review and analysis of national prevalence studies. *PLoS One, 15*(1), e0227884. https://doi.org/10.1371/journal.pone.0227884

McDonald, K. C. (2007, Jan 15). Child abuse: approach and management. *American Family Physician, 75*(2), 221-228. https://www.aafp.org/afp/2007/0115/p221.html

McGloin, J. M., & Widom, C. S. (2001, Fall). Resilience among abused and neglected children grown up. *Development and Psychopathology, 13*(4), 1021-1038. https://doi.org/10.1017/s095457940100414x

Meinck, F., Boyes, M. E., Cluver, L., Ward, C. L., Schmidt, P., DeStone, S., & Dunne, M. P. (2018, 2018/08/01/). 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. https://doi.org/10.1016/j.chiabu.2018.05.022

Meinck, F., Steinert, J., Sethi, D., Gilbert, R., Bellis, M., Alink, L., & Baban, A. (2016). *Measuring and monitoring national prevalence of child maltreatment: a practical handbook*. World Health Organization. https://www.euro.who.int/__data/assets/pdf_file/0003/317505/Measuring-monitoring-national-prevalence-child-maltreatment-practical-handbook.pdf

Milner, J. S. (2000). Social information processing and child physical abuse: Theory and research. In *Nebraska Symposium on Motivation Vol. 46, 1998: Motivation and child maltreatment.* (pp. 39-84). University of Nebraska Press.

Milner, J. S., & Crouch, J. L. (1997). Impact and detection of response distortions on parenting measures used to assess risk for child physical abuse. *Journal of Personality Assessment, 69*(3), 633-650. https://doi.org/10.1207/s15327752jpa6903_15

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The, P. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6*(7), e1000097. https://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, May). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research, 27*(5), 1171-1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs)-User manual (version 1.0).* https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

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. (2010, 2010/07/01/). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737-745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Mokkink, L. B., Terwee, C. B., Prinsen, C. A. C., & de Vet, H. C. W. (2016, 2016/01/01/). Taxonomy of measurement properties: A commentary on Polit (2015). *International Journal of Nursing Studies, 53*, 399-400. https://doi.org/10.1016/j.ijnurstu.2015.08.010

Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2009, Apr). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research, 18*(3), 313-333. https://doi.org/10.1007/s11136-009-9451-9

Molnar, B. E., Beatriz, E. D., & Beardslee, W. R. (2016, 2016/10/01). Community-Level Approaches to Child Maltreatment Prevention. *Trauma, Violence, & Abuse, 17*(4), 387-397. https://doi.org/10.1177/1524838016658879

Moody, G., Cannings-John, R., Hood, K., Kemp, A., & Robling, M. (2018, 2018/10/10). Establishing the international prevalence of self-reported child maltreatment: a systematic review by maltreatment type and gender. *BMC Public Health, 18*(1), 1164. https://doi.org/10.1186/s12889-018-6044-y

Moore, S. E., Scott, J. G., Ferrari, A. J., Mills, R., Dunne, M. P., Erskine, H. E., Devries, K. M., Degenhardt, L., Vos, T., Whiteford, H. A., McCarthy, M., & Norman, R. E. (2015, 2015/10/01/). Burden attributable to child maltreatment in Australia. *Child Abuse & Neglect, 48*, 208-220. https://doi.org/10.1016/j.chiabu.2015.05.006

Morsbach, S. K., & Prinz, R. J. (2006, 2006/03/01). Understanding and Improving the Validity of Self-Report of Parenting. *Clinical Child and Family Psychology Review, 9*(1), 1-21. https://doi.org/10.1007/s10567-006-0001-5

Naaktgeboren, C. A., Bertens, L. C. M., Smeden, M. v., Groot, J. A. H. d., Moons, K. G. M., & Reitsma, J. B. (2013). Value of composite reference standards in diagnostic research. *BMJ, 347*, f5605. https://doi.org/10.1136/bmj.f5605

Negriff, S., Schneiderman, J. U., & Trickett, P. K. (2016, 2017/02/01). Concordance Between Self-Reported Childhood Maltreatment Versus Case Record Reviews for Child Welfare–Affiliated Adolescents: Prevalence Rates and Associations With Outcomes. *Child Maltreatment, 22*(1), 34-44. https://doi.org/10.1177/1077559516674596

Norman, R. E., Byambaa, M., De, R., Butchart, A., Scott, J., & Vos, T. (2012). The long-term health consequences of child physical abuse, emotional abuse, and neglect: a systematic review and meta-analysis. *PLOS Medicine, 9*(11), e1001349. https://doi.org/10.1371/journal.pmed.1001349

O'Donnell, M., Scott, D., & Stanley, F. (2008). Child abuse and neglect — is it time for a public health approach? *Australian and New Zealand Journal of Public Health, 32*(4), 325-330. https://doi.org/10.1111/j.1753-6405.2008.00249.x

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine, 18*(3), e1003583. https://doi.org/10.1371/journal.pmed.1003583

Pallant, J. F., Haines, H. M., Hildingsson, I., Cross, M., & Rubertsson, C. (2014, 2014/03/15). Psychometric evaluation and refinement of the Prenatal Attachment Inventory. *Journal of Reproductive and Infant Psychology, 32*(2), 112-125. https://doi.org/10.1080/02646838.2013.871627

Parsons, A. M., Heyman, R. E., Mitnick, D. M., & Smith Slep, A. M. (2020). Chapter 8 - Intimate partner violence and child maltreatment: Definitions, prevalence, research, and theory through a cross-cultural lens. In W. K. Halford & F. van de Vijver (Eds.), *Cross-Cultural Family Research and Practice* (pp. 249-285). Academic Press. https://doi.org/10.1016/B978-0-12-815493-9.00008-9

Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011, Dec). 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 1--eliciting concepts for a new PRO instrument. *Value in Health, 14*(8), 967-977. https://doi.org/10.1016/j.jval.2011.06.014

Pechtel, P., & Pizzagalli, D. A. (2011, 2011/03/01). Effects of early life stress on cognitive and affective function: an integrated review of human literature. *Psychopharmacology, 214*(1), 55-70. https://doi.org/10.1007/s00213-010-2009-2

Peden, M., Oyebite, K., Ozanne-Smith, J., Hyder, A., Branche, C., Rahman, A., Rivara, F., & Bartolomeos, K. (Eds.). (2008). *World report on child injury prevention*. World Health Organization. https://www.ncbi.nlm.nih.gov/books/NBK310641/.

Pelletier, H. L., & Knox, M. (2017, 2017/09/01). Incorporating Child Maltreatment Training into Medical School Curricula. *Journal of Child & Adolescent Trauma, 10*(3), 267-274. https://doi.org/10.1007/s40653-016-0096-x

Plant, D. T., Barker, E. D., Waters, C. S., Pawlby, S., & Pariante, C. M. (2013, Mar). Intergenerational transmission of maltreatment and psychopathology: the role of antenatal depression. *Psychological Medicine, 43*(3), 519-528. https://doi.org/10.1017/s0033291712001298

Polit, D. F. (2015, 2015/11/01/). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies, 52*(11), 1746-1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002

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

Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P. R., & Terwee, C. B. (2016, 2016/09/13). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" – a practical guideline. *Trials, 17*(1), 449. https://doi.org/10.1186/s13063-016-1555-2

Putnam-Hornstein, E., Webster, D., Needell, B., & Magruder, J. (2011). A Public Health Approach to Child Maltreatment Surveillance: Evidence from a Data Linkage Project in the United States. *Child Abuse Review, 20*(4), 256-273. https://doi.org/10.1002/car.1191

Repetti, R. L., Taylor, S. E., & Seeman, T. E. (2002). Risky families: Family social environments and the mental and physical health of offspring. *Psychological Bulletin, 128*(2), 330-366. https://doi.org/10.1037/0033-2909.128.2.330

Ricci, L., Lanfranchi, J.-B., Lemetayer, F., Rotonda, C., Guillemin, F., Coste, J., & Spitz, E. (2018, 2019/01/01). Qualitative Methods Used to Generate Questionnaire Items: A Systematic Review. *Qualitative Health Research, 29*(1), 149-156. https://doi.org/10.1177/1049732318783186

Rijlaarsdam, J., Stevens, G. W., Jansen, P. W., Ringoot, A. P., Jaddoe, V. W., Hofman, A., Ayer, L., Verhulst, F. C., Hudziak, J. J., & Tiemeier, H. (2014, May). Maternal Childhood Maltreatment and Offspring Emotional and Behavioral Problems: Maternal and Paternal Mechanisms of Risk Transmission. *Child Maltreat, 19*(2), 67-78. https://doi.org/10.1177/1077559514527639

Ripoll-Núñez, K. J., & Rohner, R. P. (2006, 2006/08/01). Corporal Punishment in Cross-Cultural Perspective: Directions for a Research Agenda. *Cross-Cultural Research, 40*(3), 220-249. https://doi.org/10.1177/1069397105284395

Roberts, A. L., Lyall, K., Rich-Edwards, J. W., Ascherio, A., & Weisskopf, M. G. (2013, May). Association of maternal exposure to childhood abuse with elevated risk for autism in offspring. *JAMA Psychiatry, 70*(5), 508-515. https://doi.org/10.1001/jamapsychiatry.2013.447

Rodriguez, C. M., Silvia, P. J., & Gaskin, R. E. (2019). Predicting maternal and paternal parent-child aggression risk: Longitudinal multimethod investigation using social information processing theory. *Psychology of Violence, 9*(3), 370-382. https://doi.org/10.1037/vio0000115

Rosenkoetter, U., & Tate, R. L. (2018). Assessing Features of Psychometric Assessment Instruments: A Comparison of the COSMIN Checklist with Other Critical Appraisal Tools. *Brain Impairment, 19*(1), 103-118. https://doi.org/10.1017/BrImp.2017.29

Ryan, J. P., Jacob, B. A., Gross, M., Perron, B. E., Moore, A., & Ferguson, S. (2018, 2018/11/01). Early Exposure to Child Maltreatment and Academic Outcomes. *Child Maltreatment, 23*(4), 365-375. https://doi.org/10.1177/1077559518786815

Saini, S. M., Hoffmann, C. R., Pantelis, C., Everall, I. P., & Bousman, C. A. (2019, 2019/02/01/). Systematic review and critical appraisal of child abuse measurement instruments. *Psychiatry Research, 272*, 106-113. https://doi.org/10.1016/j.psychres.2018.12.068

Scarborough, A. A., Lloyd, E. C., & Barth, R. P. (2009). Maltreated Infants and Toddlers: Predictors of Developmental Delay. *Journal of Developmental & Behavioral Pediatrics, 30*(6), 489-498. https://doi.org/10.1097/DBP.0b013e3181c35df6

Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011, 2011/03/01/). What makes a measurement instrument valid and reliable? *Injury, 42*(3), 236-240. https://doi.org/10.1016/j.injury.2010.11.042

Scott, D., Lonne, B., & Higgins, D. (2016). Public Health Models for Preventing Child Maltreatment:Applications From the Field of Injury Prevention. *Trauma, Violence, & Abuse, 17*(4), 408-419. https://doi.org/10.1177/1524838016658877

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*. Administrationfor Children and Families. https://www.acf.hhs.gov/media/8885

Slep, A. M., Heyman, R. E., & Foran, H. M. (2015, Mar). Child maltreatment in DSM-5 and ICD-11. *Family Process, 54*(1), 17-32. https://doi.org/10.1111/famp.12131

Sprangers, M. A. G., & Aaronson, N. K. (1992, 1992/07/01/). 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*(7), 743-760. https://doi.org/10.1016/0895-4356(92)90052-O

Stith, S. M., Liu, T., Davies, L. C., Boykin, E. L., Alder, M. C., Harris, J. M., Som, A., McPherson, M., & Dees, J. E. M. E. G. (2009, 2009/01/01/). Risk factors in child maltreatment: A meta-analytic review of the literature. *Aggression and Violent Behavior, 14*(1), 13-29. https://doi.org/10.1016/j.avb.2006.03.006

Stoltenborgh, M., Bakermans-Kranenburg, M. J., Alink, L. R. A., & van Ijzendoorn, M. H. (2015, 2015/01/01). The Prevalence of Child Maltreatment across the Globe: Review of a Series of Meta-Analyses. *Child Abuse Review, 24*(1), 37-50. https://doi.org/10.1002/car.2353

Stoltenborgh, M., van Ijzendoorn, M. H., Euser, E. M., & Bakermans-Kranenburg, M. J. (2011, 2011/05/01). A Global Perspective on Child Sexual Abuse: Meta-Analysis of Prevalence Around the World. *Child Maltreatment, 16*(2), 79-101. https://doi.org/10.1177/1077559511403920

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

Sumner, S. A., Mercy, A. A., Saul, J., Motsa-Nzuza, N., Kwesigabo, G., Buluma, R., Marcelin, L. H., Lina, H., Shawa, M., Moloney-Kitts, M., Kilbane, T., Sommarin, C., Ligiero, D. P., Brookmeyer, K., Chiang, L., Lea, V., Lee, J., Kress, H., & Hillis, S. D. (2015, Jun 5). Prevalence of sexual violence against children and use of social services - seven countries, 2007-2013. *MMWR Morb Mortal Wkly Rep, 64*(21), 565-569. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4584766/

Tang, W., Hu, J., Zhang, H., Wu, P., & He, H. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai archives of psychiatry, 27*(1), 62-67. https://doi.org/10.11919/j.issn.1002-0829.215010

Taylor, S. E., Lerner, J. S., Sage, R. M., Lehman, B. J., & Seeman, T. E. (2004, 2004/12/01). Early Environment, Emotions, Responses to Stress, and Health. *Journal of Personality, 72*(6), 1365-1394. https://doi.org/10.1111/j.1467-6494.2004.00300.x

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., Lai, J.-S., Choi, S. W., Hays, R. D., Reeve, B. B., Reise, S. P., Pilkonis, P. A., &

Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science, 51*(2), 148-180. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2844669/

Terwee, C. B., de Vet, H. C. W., Prinsen, C. A. C., & Mokkink, L. B. (2016). *Comment on "Checklist to operationalize measurement characteristics of patient-reported outcome measures"*. VU University Medical Center, Department of Epidemiology and Biostatistics. https://www.cosmin.nl/wp-content/uploads/Letter-comment-on-Francis.pdf

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, May). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research, 27*(5), 1159-1170. https://doi.org/10.1007/s11136-018-1829-0

Terwee, C. B., Prinsen, C. A. C., Ricci Garotti, M. G., Suman, A., de Vet, H. C. W., & Mokkink, L. B. (2016, 2016/04/01). The quality of systematic reviews of health-related outcome measurement instruments. *Quality of Life Research, 25*(4), 767-779. https://doi.org/10.1007/s11136-015-1122-4

Tessier, N. G., O'Higgins, A., & Flynn, R. J. (2018, Jan). Neglect, educational success, and young people in out-of-home care: Cross-sectional and longitudinal analyses. *Child Abuse & Neglect, 75*, 115-129. https://doi.org/10.1016/j.chiabu.2017.06.005

Thacker, S. B., & Berkelman, R. L. (1988). Public health surveillance in the United States. *Epidemiologic Reviews, 10*, 164-190. https://doi.org/10.1093/oxfordjournals.epirev.a036021

Thornberry, T. P., & Henry, K. L. (2013, May). Intergenerational continuity in maltreatment. *Journal of Abnormal Child Psychology, 41*(4), 555-569. https://doi.org/10.1007/s10802-012-9697-5

U.S. Department of Health and Human Services. (2018). *Child Abuse Prevention and Treatment Act*. Washington, D.C.

U.S. Department of Health Human Services. (2020). *Child maltreatment 2018*. U.S. DHHS Administration for Children and Families. https://www.acf.hhs.gov/sites/default/files/documents/cb/cm2018.pdf

U.S. Department of Health Human Services. (2021). *Child maltreatment 2019*. U.S. DHHS Administration for Children and Families. https://www.acf.hhs.gov/sites/default/files/documents/cb/cm2019.pdf

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. United Nations, Department of Economic and Social Affairs. https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

Vachon, D. D., Krueger, R. F., Rogosch, F. A., & Cicchetti, D. (2015). Assessment of the Harmful Psychiatric and Behavioral Effects of Different Forms of Child Maltreatment. *JAMA Psychiatry, 72*(11), 1135-1142. https://doi.org/10.1001/jamapsychiatry.2015.1792

van der Put, C. E., Assink, M., Gubbels, J., & Boekhout van Solinge, N. F. (2018, 2018/06/01). Identifying Effective Components of Child Maltreatment Interventions: A Meta-analysis. *Clinical Child and Family Psychology Review, 21*(2), 171-202. https://doi.org/10.1007/s10567-017-0250-5

Verweij, L. M., Terwee, C. B., Proper, K. I., Hulshof, C. T. J., & van Mechelen, W. (2013). Measurement error of waist circumference: gaps in knowledge. *Public Health Nutrition, 16*(2), 281-288. https://doi.org/10.1017/S1368980012002741

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*(6), 2055-2064. https://doi.org/10.1542/peds.2005-2204

Voisine, S., & Baker, A. J. L. (2012, 2012/07/01). Do Universal Parenting Programs Discourage Parents from Using Corporal Punishment: A Program Review. *Families in Society, 93*(3), 212-218. https://doi.org/10.1606/1044-3894.4217

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M. (2011, 2011/10/18). QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Annals of Internal Medicine, 155*(8), 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wiering, B., de Boer, D., & Delnoij, D. (2017, 2017/02/01). Patient involvement in the development of patient-reported outcome measures: a scoping review. *Health Expectations, 20*(1), 11-23. https://doi.org/10.1111/hex.12442

Wittkowski, A., Vatter, S., Muhinyi, A., Garrett, C., & Henderson, M. (2020). Measuring bonding or attachment in the parent-infant-relationship: A systematic review of parent-report assessment measures, their psychometric properties and clinical utility. *Clinical psychology review, 82*, 101906-101906. https://doi.org/10.1016/j.cpr.2020.101906

World Health Organization. (1999). *Report of the Consultation on Child Abuse Prevention*. Author. https://apps.who.int/iris/handle/10665/65900

World Health Organization. (2002). *Child abuse and neglect facts.* Author. https://www.who.int/violence_injury_prevention/violence/world_report/factsheets/en/childabusefacts.pdf

World Health Organization. (2005). *Violence Prevention Alliance: building global commitment for violence prevention*. Author. https://apps.who.int/iris/bitstream/handle/10665/43200/924159313X_eng.pdf

World Health Organization. (2006). *Preventing child maltreatment: a guide to taking action and generating evidence*. Author. https://apps.who.int/iris/bitstream/handle/10665/43499/9241594365_eng.pdf?sequence=1

World Health Organization. (2018). *International Classification of Diseases, 11th edition (ICD-11)*. Author. https://icd.who.int/en

World Health Organization. (2020). *Global status report on preventing violence against children 2020*. Author. https://apps.who.int/iris/handle/10665/332394

Yehuda, R., & Lehrner, A. (2018, 2018/10/01). Intergenerational transmission of trauma effects: putative role of epigenetic mechanisms. *World Psychiatry, 17*(3), 243-257. https://doi.org/10.1002/wps.20568

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020a). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 1: Content Validity. *Trauma, Violence, & Abuse*. https://doi.org/10.1177/1524838019898456

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020b). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 2: Internal Consistency, Reliability, Measurement Error, Structural Validity, Hypothesis Testing, Cross-Cultural Validity, and Criterion Validity. *Trauma, Violence, & Abuse*. https://doi.org/10.1177/1524838020915591

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2021). A Systematic Review Evaluating Responsiveness of Parent- or Caregiver-Reported Child Maltreatment Instruments to Parenting Interventions. *Trauma, Violence, & Abuse*. Manuscript submitted for publication

PART Ⅱ

ARTICLES

Article 1

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 1: Content Validity. *Trauma, Violence, & Abuse.* Advanced online publication. https://doi.org/10.1177/1524838019898456

TRAUMA, VIOLENCE, & ABUSE
1-19
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1524838019898456
journals.sagepub.com/home/tva
SAGE

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 Health 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

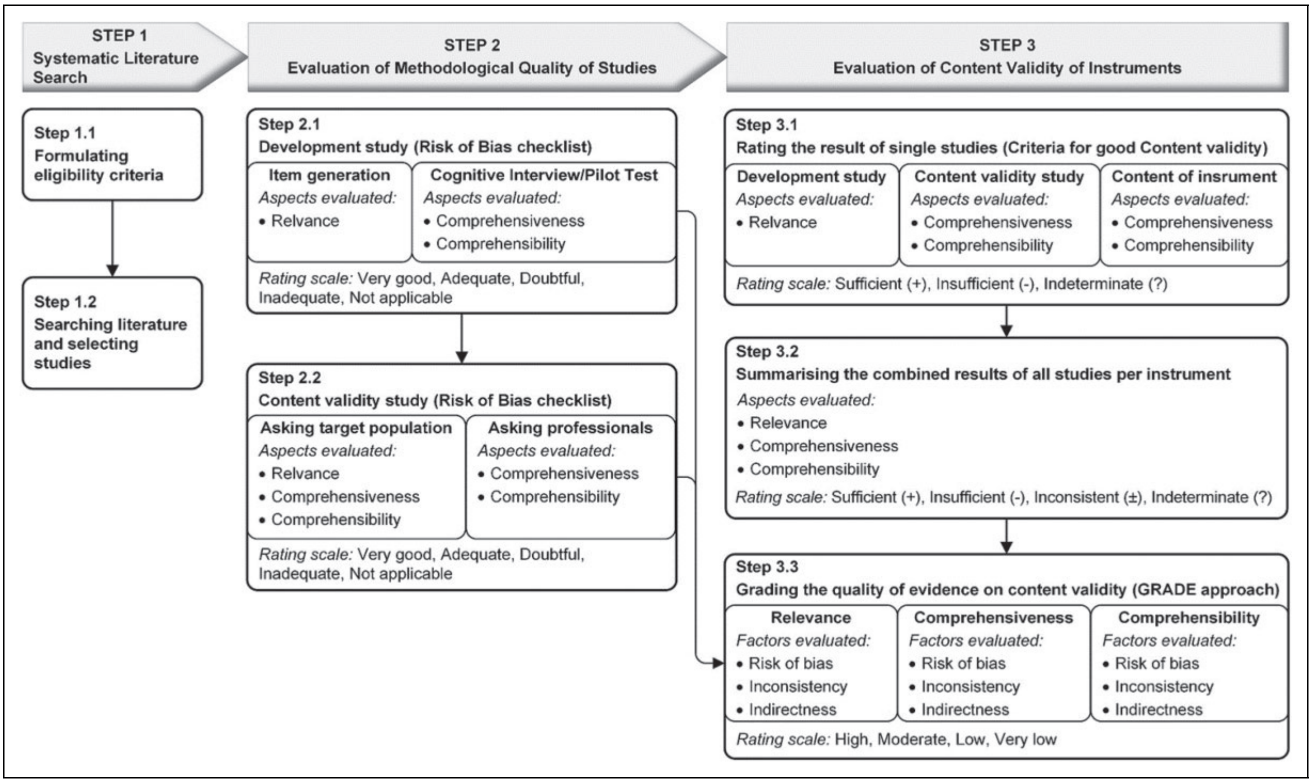

STEP 1
Systematic Literature Search
STEP 2
Evaluation of Methodological Quality of Studies
STEP 3
Evaluation of Content Validity of Instruments
Step 1.1
Formulating eligibility criteria
Step 1.2
Searching literature and selecting studies
Step 2.1
Development study (Risk of Bias checklist)
Item generation
Aspects evaluated:
• Relvance
Cognitive Interview/Pilot Test
Aspects evaluated:
• Comprehensiveness
• Comprehensibility
Rating scale: Very good, Adequate, Doubtful, Inadequate, Not applicable
Step 2.2
Content validity study (Risk of Bias checklist)
Asking target population
Aspects evaluated:
• Relvance
• Comprehensiveness
• Comprehensibility
Asking professionals
Aspects evaluated:
• Comprehensiveness
• Comprehensibility
Rating scale: Very good, Adequate, Doubtful, Inadequate, Not applicable
Step 3.1
Rating the result of single studies (Criteria for good Content validity)
Development study
Aspects evaluated:
• Relvance
Content validity study
Aspects evaluated:
• Comprehensiveness
• Comprehensibility
Content of insrument
Aspects evaluated:
• Comprehensiveness
• Comprehensibility
Rating scale: Sufficient (+), Insufficient (-), Indeterminate (?)
Step 3.2
Summarising the combined results of all studies per instrument
Aspects evaluated:
• Relevance
• Comprehensiveness
• Comprehensibility
Rating scale: Sufficient (+), Insufficient (-), Inconsistent (±), Indeterminate (?)
Step 3.3
Grading the quality of evidence on content validity (GRADE approach)
Relevance
Factors evaluated:
• Risk of bias
• Inconsistency
• Indirectness
Comprehensiveness
Factors evaluated:
• Risk of bias
• Inconsistency
• Indirectness
Comprehensibility
Factors evaluated:
• Risk of bias
• Inconsistency
• Indirectness
Rating scale: High, Moderate, Low, Very low

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).

$$\text{Total score for methodological quality (\%)} \\ = \frac{(\text{total score obtained} - \text{min score possible})}{(\text{max score possible} - \text{min score possible})} \times 100.$$

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

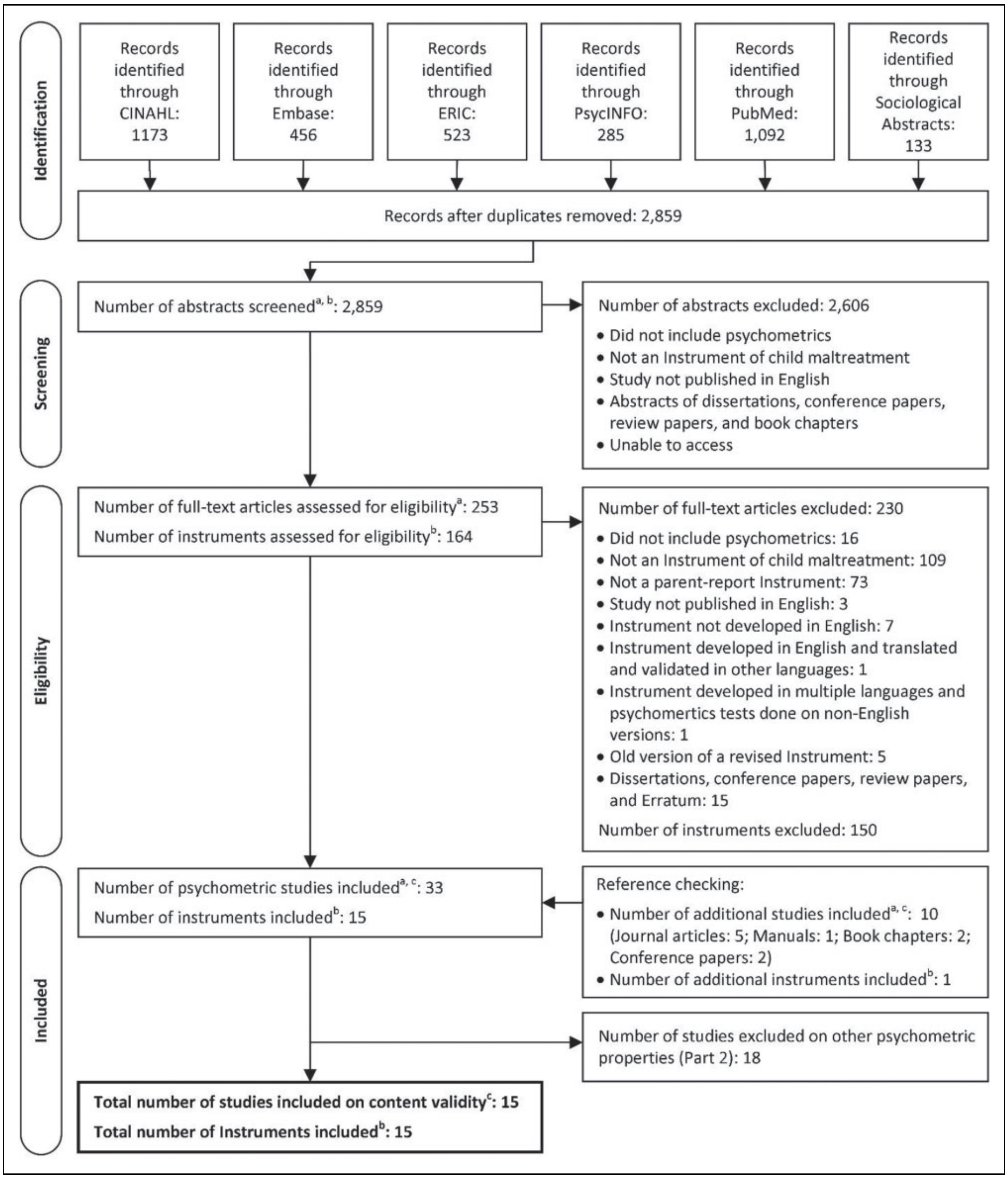

Identification
Records identified through CINAHL: 1173
Records identified through Embase: 456
Records identified through ERIC: 523
Records identified through PsycINFO: 285
Records identified through PubMed: 1,092
Records identified through Sociological Abstracts: 133
Records after duplicates removed: 2,859
Screening
Number of abstracts screened[a, b]: 2,859
Number of abstracts excluded: 2,606
• Did not include psychometrics
• Not an Instrument of child maltreatment
• Study not published in English
• Abstracts of dissertations, conference papers, review papers, and book chapters
• Unable to access
Eligibility
Number of full-text articles assessed for eligibility[a]: 253
Number of instruments assessed for eligibility[b]: 164
Number of full-text articles excluded: 230
• Did not include psychometrics: 16
• Not an Instrument of child maltreatment: 109
• Not a parent-report Instrument: 73
• Study not published in English: 3
• Instrument not developed in English: 7
• Instrument developed in English and translated and validated in other languages: 1
• Instrument developed in multiple languages and psychomertics tests done on non-English versions: 1
• Old version of a revised Instrument: 5
• Dissertations, conference papers, review papers, and Erratum: 15
Number of instruments excluded: 150
Included
Number of psychometric studies included[a, c]: 33
Number of instruments included[b]: 15
Reference checking:
• Number of additional studies included[a, c]: 10 (Journal articles: 5; Manuals: 1; Book chapters: 2; Conference papers: 2)
• Number of additional instruments included[b]: 1
Number of studies excluded on other psychometric properties (Part 2): 18
Total number of studies included on content validity[c]: 15
Total number of Instruments included[b]: 15

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

(continued)

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
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 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

(continued)

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%)	NR	NR
IPPS	Gordon et al. (1979)	Inadequate (3.5%)	Inadequate (7.1%)	Inadequate (4.8%)	Inadequate (12.5%)	NR	NR	Doubtful (33.3%)	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.

		Relevance[a]			Comprehensiveness[a]			Comprehensibility[a]		
Instrument	Reference	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; ? = indeterminate rating; – = insufficient 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.

	Relevance		Comprehensiveness		Comprehensibility	
Instrument	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 COSMIN 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.

Authors' Note

The authors confirm that this work has not been published elsewhere nor is it currently under consideration for publication elsewhere.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

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

Supplemental Material

The supplemental material for this article is available online.

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.

Appendices

Appendix A. *Database Search Strategies.*

Database	Search Terms (Subject heading and Free text words)	Number of records
CINAHL	(((MH "Child Abuse+") OR (MH "Domestic Violence+") OR (MH "Family Conflict") OR (MH "Aggression+") OR (MH "Punishment")) AND ((MH "Parents+") OR (MH "Parenting") OR (MH "Father-Infant Relations")OR (MH "Father-Child Relations") OR (MH "Fathers+") OR (MH "Mother-Child Relations") OR (MH "Mother-Infant Relations") OR (MH "Mothers+") OR (MH "Family+") OR (MH "Caregivers") OR (MH "Child Rearing+")) AND ((MH "Psychometrics") OR (MH "Measurement Issues and Assessments") OR (MH "Validity") OR (MH "Predictive Validity") OR (MH "Reliability and Validity") OR (MH "Internal Validity") OR (MH "Face Validity") OR (MH "External Validity") OR (MH "Discriminant Validity") OR (MH "Criterion-Related Validity") OR (MH "Consensual Validity") OR (MH "Concurrent Validity") OR (MH "Qualitative Validity") OR (MH "Construct Validity") OR (MH "Content Validity") OR (MH "Questionnaire Validation") OR (MH "Validation Studies") OR (MH "Test-Retest Reliability") OR (MH "Sensitivity and Specificity") OR (MH "Reproducibility of Results") OR (MH "Reliability") OR (MH "Intrarater Reliability") OR (MH "Interrater Reliability") OR (MH "Measurement Error") OR (MH "Bias (Research)") OR (MH "Selection Bias") OR (MH "Sampling Bias") OR (MH "Precision") OR (MH "Sample Size Determination") OR (MH "Repeated Measures") OR (Psychometric* or reliability or validit* or reproducibility or bias))) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) Limiters - Published Date: 20181001-20191031)	1173
Embase	((child abuse/ OR child neglect/ OR emotional abuse/ OR physical abuse/ OR battering/ OR domestic violence/ OR physical violence/ OR family conflict/ OR victim/ OR aggression/ OR punishment/) AND (parent/ OR father/ OR father child relation/ OR mother/ OR mother child relation/ OR family/ OR caregiver/ OR child rearing/) AND (psychometry/ or validity/ or reliability/ or measurement error/ or measurement precision/ or measurement repeatability/ or error/ or statistical bias/ or test retest reliability/ or intrarater reliability/ or interrater reliability/ or accuracy/ or criterion validity/ or internal validity/ or face validity/ or external validity/ or discriminant validity/ or concurrent validity/ or qualitative validity/ or construct validity/ or content validity/)) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) limit to yr="2019 -Current")	456
ERIC	((Child abuse/ OR Child neglect/ OR violence/ OR family violence/) AND (parenting styles/ OR parents/ OR child rearing/ OR father attitudes/ OR fathers/ OR mother attitudes/ OR mothers/ OR family attitudes/ OR caregiver attitudes/ OR caregiver child relationship/ OR caregiver role/ OR family environment/) AND (Psychometrics/ OR Validity/ OR Reliability/ OR Error of Measurement/ OR Bias/ OR Interrater Reliability/ OR Accuracy/ OR Predictive Validity/ OR Construct Validity/ OR Content Validity/)) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) limit to yr="Last year")	523

(continued)

Appendix A. *(continued)*

Database	Search Terms (Subject heading and Free text words)	Number of records
PsycINFO	((child abuse/ OR child neglect/ OR violence/ OR domestic violence/ OR physical abuse/ OR family conflict/ OR victimization/ OR aggressive behaviOR/ OR aggressiveness/ OR punishment/) AND (parent child communication/ OR parent child relations/ OR parenting/ OR parenting style/ OR parents/ OR father child communication/ OR father child relations/ OR fathers/ OR mother child communication/ OR mother child relations/ OR mothers/ OR family/ OR caregivers/) AND (Psychometrics/ OR Statistical Validity/ OR Test Validity/ OR Statistical Reliability/ OR Test Reliability/ OR Error of Measurement/ OR Errors/ OR Response Bias/ OR Interrater Reliability/ OR Repeated Measures/)) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) limit to yr= “2019 -Current”)	285
PubMed	(("Child Abuse"[Mesh] OR "Physical Abuse"[Mesh] OR "Domestic Violence"[Mesh] OR "Violence"[Mesh] OR "Family Conflict"[Mesh] OR "Aggression"[Mesh] OR "Punishment"[Mesh]) AND ("Parents"[Mesh] OR "Parent-Child Relations"[Mesh] OR "Parenting"[Mesh] OR "Fathers"[Mesh] OR "Father-Child Relations"[Mesh] OR "Mothers"[Mesh] OR "Mother-Child Relations"[Mesh] OR "Family"[Mesh] OR "Caregivers"[Mesh] OR "Child Rearing"[Mesh]) AND ("Psychometrics"[Mesh] OR "Reproducibility of Results"[Mesh] OR "Validation Studies as Topic"[Mesh] OR "Validation Studies" [Publication Type] OR "Bias"[Mesh] OR "Observer Variation"[Mesh] OR "Selection Bias"[Mesh] OR "Diagnostic Errors"[Mesh] OR "Dimensional Measurement Accuracy"[Mesh] OR “Predictive Value of Tests"[Mesh] OR "Discriminant Analysis"[Mesh])) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) Filters: Publication date from 2018/10/05 to 2019/10/05)	1092
Sociological Abstracts	(MAINSUBJECT.EXACT("Child Neglect") OR MAINSUBJECT.EXACT("Child Abuse") OR (MAINSUBJECT.EXACT("Violence") OR MAINSUBJECT.EXACT("Family Violence")) OR MAINSUBJECT.EXACT("Family Conflict") OR MAINSUBJECT.EXACT("Victimization") OR MAINSUBJECT.EXACT("Victims") OR MAINSUBJECT.EXACT("Aggression") OR (MAINSUBJECT.EXACT("Punishment") OR MAINSUBJECT.EXACT("Corporal Punishment")) OR MAINSUBJECT.EXACT("Emotional Abuse")) AND (MAINSUBJECT.EXACT("Parent Child Relations") OR MAINSUBJECT.EXACT("Parental Influence") OR MAINSUBJECT.EXACT("Parents") OR MAINSUBJECT.EXACT("Parental Attitudes") OR MAINSUBJECT.EXACT("Parenthood")) OR MAINSUBJECT.EXACT("Childrearing Practices") OR MAINSUBJECT.EXACT("Fathers") OR MAINSUBJECT.EXACT("Mothers") OR (MAINSUBJECT.EXACT("Family") OR MAINSUBJECT.EXACT("Family Relations") OR MAINSUBJECT.EXACT("Family Conflict") OR MAINSUBJECT.EXACT("Family Violence")) OR MAINSUBJECT.EXACT("Caregivers")) AND (MAINSUBJECT.EXACT("Psychometric Analysis") OR MAINSUBJECT.EXACT("Validity") OR MAINSUBJECT.EXACT("Reliability") OR MAINSUBJECT.EXACT("Error of Measurement") OR MAINSUBJECT.EXACT("Errors") OR MAINSUBJECT.EXACT("Test Bias") OR MAINSUBJECT.EXACT("Statistical Bias") OR MAINSUBJECT.EXACT("Bias") OR MAINSUBJECT.EXACT("Accuracy") OR MAINSUBJECT.EXACT("Agreement") OR MAINSUBJECT.EXACT("Research Design Error") OR MAINSUBJECT.EXACT("Specificity") OR MAINSUBJECT.EXACT("Sampling"))	133

Notes. All searches performed on the 29th of January 2018 with an update on the 5th of October 2019.

Appendix B. *Overview of Child Maltreatment Instrument: Reasons for Exclusion.*

No	Instrument[a] *(alphabetical order)*	Abbreviation	Reason for exclusion
1	Adolescent Clinical Sexual Behavior Inventory (William N. Friedrich, Lysne, Sim, & Shamos, 2004)	ACSBI	Not a measure of child maltreatment
2	Adolescent Sexual Behavior Inventory- Self Report (Wherry, Berres, Sim, & Friedrich, 2009)	ACSBI-S	Not a measure of child maltreatment
3	Adult Attachment Interviews (Hesse, 2008)	AAIs	Not a parent-report measure
4	Adult-Adolescent Parenting Inventory (Bavolek, 1984)	AAPI	Old version of a revised measure
5	Adverse Childhood Experiences Questionnaire (Felitti et al., 1998)	ACEs	Not a parent-report measure
6	Alabama Parenting Questionnaire (Shelton, Frick, & Wootton, 1996)	APQ	Not a measure of child maltreatment
7	Assessing Environments (Berger, Knutson, Mehm, & Perkins, 1988)	AEIII	Not a parent-report measure
8	Assessment of parental awareness of the shaken baby syndrome[b] (Mann, Rai, Sharif, & Vavasseur, 2015)	N/A	No psychometric data found
9	Body Image Victimization Experiences Scale (Duarte & Pinto-Gouveia, 2017)	BIVES	Not a measure of child maltreatment
10	Brief Child Abuse Potential Inventory (Ondersma, Chaffin, Mullins, & LeBreton, 2005)	BCAP	Not a measure of child maltreatment
11	Brigid Collins Risk Screener (Weberling, Forgays, Crain-Thoreson, & Hyman, 2003)	BCRS	Not a measure of child maltreatment
12	California Family Risk Assessment (W. L. Johnson, 2011)	CFRA	Not a parent-report measure
13	Caregiver–Child Social/Emotional and Relationship Rating Scale (McCall, Groark, & Fish, 2010)	CCSERRS	Not a measure of child maltreatment
14	Child Abuse Inventory at Emergency Rooms (Sittig et al., 2016)	CHAINER	Not a parent-report measure
15	Child Abuse Potential Inventory (Milner, 1986)	CAP	Not a measure of child maltreatment
16	Child Abuse Risk Assessment Scale (Chan, 2012)	CARAS	Not developed in English
17	Child and Adolescent Trauma Screen (Sachser et al., 2017)	CATS	Not a measure of child maltreatment
18	Child Behavior Checklist (Achenbach & Rescorla, 2000)	CBCL	Not a measure of child maltreatment
19	Child emotional maltreatment module[b] (A. M. Slep, Heyman, & Snarr, 2011)	N/A	No psychometric data found
20	Child maltreatment assessment (Salum et al., 2016)	N/A	Not developed in English
21	Child Maltreatment Measure[b] (Tajima, Herrenkohl, Huang, & Whitney, 2004)	N/A	No psychometric data found
22	Child Protective Services Review Document (Fanshel, Finch, & Grundy, 1994)	CPSRD	Not a parent-report measure
23	Child Reflective Functioning Scale (Ensink et al., 2015)	CRF	Not a measure of child maltreatment
24	Child Sexual Behavior Inventory (W. N. Friedrich et al., 2001)	CSBI	Not a measure of child maltreatment
25	Child Well-Being Scales (Gaudin, Polansky, & Kilpatrick, 1992)	CWBS	Not a parent-report measure
26	Childhood Experience of Care and Abuse (Brown, Craig, Harris, Handley, & Harvey, 2007)	CECA	Not a parent-report measure
27	Childhood Experience of Care and Abuse Questionnaire (N. Smith, Lam, Bifulco, & Checkley, 2002)	CECA.Q	Not a parent-report measure
28	Childhood Experiences of Violence Questionnaire (Walsh, MacMillan, Trocme, Jamieson, & Boyle, 2008)	CEVQ	Not a parent-report measure
29	Childhood Trauma Interview (Fink, Bernstein, Handelsman, Foote, & Lovejoy, 1995)	CTI	Not a parent-report measure
30	Childhood Trauma Questionnaire (Bernstein, Ahluvalia, Pogge, & Handelsman, 1997)	CTQ	Not a parent-report measure
31	Childhood Trauma Questionnaire Short Form (Forde, Baron, Scher, & Stein, 2012)	CTQ-SF	Not a parent-report measure
32	Child-Parent Relationship Scale (Driscoll & Pianta, 2011)	CPRS	Not a measure of child maltreatment
33	Child–Parent Relationship Scale–Short Form (Pianta, 1992)	CPRS-SF	Not a measure of child maltreatment
34	Children Intimate Relationships, and Conflictual Life Events Interview (Marshall, Feinberg, Jones, & Chote, 2017)	CIRCLE	Not a parent-report measure
35	Children's Impact of Traumatic Events Scale-Revised (Chaffin & Shultz, 2001)	CITES-R	Not a measure of child maltreatment
36	Christchurch Trauma Assessment (Nelson, Lynskey, Heath, & Martin, 2010)	N/A	Not a parent-report measure
37	Cleveland Child Abuse Potential Scale (Ezzo & Young, 2012)	C-CAPS	Not a parent-report measure
38	Comprehensive Childhood Maltreatment Inventory (Riddle & Aponte, 1999)	CCMI	Not a parent-report measure
39	Conflict Tactic Scale 2 (Straus et al., 2003)	CTS 2	Not a measure of child maltreatment

(Continued)

Appendix B. *(continued)*

No	Instrument[a] *(alphabetical order)*	Abbreviation	Reason for exclusion
40	Conflict Tactics Scales (Straus et al., 2003)	CTS	Not a measure of child maltreatment
41	Defense Style Questionnaire (Bond & Wesley, 1996)	DSQ	Not a parent-report measure
42	Disciplinary Methods Interview[b] (Thompson, 2017)	N/A	Not a measure of child maltreatment
43	Discipline Survey (Socolar, Savage, Devellis, & Evans, 2004)	N/A	Not a measure of child maltreatment
44	Dunedin Family Services Indicator (Muir et al., 1989)	DFSI	Not a parent-report measure
45	Dyadic Parent-Child Interaction Coding System-II (Eyberg, Bessmer, Newcomb, Edwards, & Robinson, 1994)	DPICS-II	Not a parent-report measure
46	Egna Minnen Beträffande Uppfostran (My Memories of Upbringing) (Castro, de Pablo, Gomez, Arrindell, & Toro, 1997)	EMBU	Not developed in English
47	Egna Minnen Betrffånde Uppfostran for Children (Castro et al., 1997; Markus, Lindhout, Boer, Hoogendijk, & Arrindell, 2003)	EMBU-C	Not a parent-report measure
48	Emotional and Physical Abuse Questionnaire (Kemper, Carlin, & Buntain-Ricklefs, 1994)	EPAB	Not a parent-report measure
49	Environmental Harshness, Health, and Life History Strategy Indicators[b] (Chua, Lukaszewski, Grant, & Sng, 2017)	N/A	Not a measure of child maltreatment
50	Exposure to Community Violence (Richters & Martinez, 1993)	ETV	Not a measure of child maltreatment
51	Exposure to violence questionnaire[b] (Kuo, Mohler, Raudenbush, & Earls, 2000)	N/A	Not a measure of child maltreatment
52	Familial Experiences Questionnaire (Wheelock, Lohr, & Silk, 1997)	FEQ	Not a parent-report measure
53	Family Affective Attitude Rating Scale (Waller, Gardner, Dishion, Shaw, & Wilson, 2012)	FAARS	Not a measure of child maltreatment
54	Family Aggression Screening Tool (Cecil, McCrory, Viding, Holden, & Barker, 2016)	FAST	Not a parent-report measure
55	Family Background Questionnaire-Brief (Melchert & Kalemeera, 2009)	FBQ-B	Not a parent-report measure
56	Family Behaviors Screen (Simmons, Craun, Farrar, & Ray, 2017)	FBS	Not a measure of child maltreatment
57	Family Betrayal Questionnaire (Delker, Smith, Rosenthal, Bernstein, & Freyd, 2017)	FBQ	Not a measure of child maltreatment
58	Family Law Detection of Overall Risk Screen (McIntosh, Wells, & Lee, 2016)	FL-DOORS	Not a measure of child maltreatment
59	Family Maltreatment Diagnostic Criteria (Heyman & Smith Slep, 2009)	N/A	Not a parent-report measure
60	Family Risk of Abuse and Neglect (Lennings, Brummert Lennings, Bussey, & Taylor, 2014)	FRAAN	Not a measure of child maltreatment
61	Family Therapy Alliance Scale (L. N. Johnson, Ketring, & Anderson, 2013)	FTAS	Not a measure of child maltreatment
62	Family Unpredictability Scale (Ross & Hill, 2000)	FUS	Not a measure of child maltreatment
63	Go/No-go Association Task Physical Discipline (Sturge-Apple, Rogge, Peltz, Suor, & Skibo, 2015)	GNAT-Physical Discipline	Not a measure of child maltreatment
64	Home Observation Measure of the Environment (Caldwell & Bradley, 2003)	HOME	Not a parent-report measure
65	Home Safety Screening (Scribano, Stevens, Marshall, Gleason, & Kelleher, 2011)	N/A	Not a measure of child maltreatment
66	Identification of Parents At Risk for Child Abuse and Neglect (van der Put et al., 2017)	IPARAN	Not developed in English
67	Index of Child Care Environment (Anme et al., 2013)	ICCE	Not developed in English
68	Invalidating Childhood Environments Scale (Mountford, Corstorphine, Tomlinson, & Waller, 2007)	ICES	Not a measure of child maltreatment
69	Inventory on Beliefs and Attitudes Towards Domestic Violence (Hutchinson & Doran, 2017)	N/A	Not a measure of child maltreatment
70	ISPCAN Child Abuse Screening Tool Children's Version (Zolotor et al., 2009)	ICAST-C	Not a parent-report measure
71	ISPCAN Child Abuse Screening Tool Parents' Version (Runyan et al., 2009)	ICAST-P	Developed in multiple languages
72	ISPCAN Child Abuse Screening Tools Retrospective Version (Dunne et al., 2009)	ICAST-R	Not a parent-report measure
73	Japanese version of Conflict Tactics Scale[b] (Baba et al., 2017)	CTS1: Japanese version	Developed in English but translated and validated in other languages
74	Juvenile Victimization Questionnaire (Finkelhor, Hamby, Ormrod, & Turner, 2005)	JVQ	Not a parent-report measure

(Continued)

Appendix B. *(continued)*

No	Instrument[a] *(alphabetical order)*	Abbreviation	Reason for exclusion
75	Maternal Characteristics Scale (Polansky, Gaudin, & Kilpatrick, 1992)	MCS	Not a measure of child maltreatment
76	Maternal discipline and appropriateness[b] (Padilla-Walker, 2008)	N/A	Not a parent-report measure
77	Maternal Responsiveness Questionnaire (Leerkes & Qu, 2017)	MRQ	Not a measure of child maltreatment
78	Maternal Self-report Support Questionnaire (D. W. Smith et al., 2010)	MSSQ	Not a measure of child maltreatment
79	Maternal Support Questionnaire–Child Report (D. W. Smith et al., 2017)	MSQ-CR	Not a measure of child maltreatment
80	Meaning of the Child Interview (Grey & Farnfield, 2017)	MotC	Not a measure of child maltreatment
81	Measure of Parenting Style (Parker et al., 1997)	MOPS	Not a parent-report measure
82	Measure Trauma Associated with Child Sexual Abuse (Choudhary, Satapathy, & Sagar, 2018)	MSCSA	Not a measure of child maltreatment
83	Measures of Community-Relevant Outcomes for Violence Prevention Programs[b] (Hausman et al., 2013)	N/A	Not a measure of child maltreatment
84	Medical History Questionnaire[b] (Famularo, Fenton, & Kinscherff, 1992)	N/A	Not a measure of child maltreatment
85	Minnesota Multiphasic Personality Inventory-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kreammer, 1989)	MMPI-2	Not a measure of child maltreatment
86	Multidimensional Assessment of Parenting Scale (Parent & Forehand, 2017)	MAPS	Not a measure of child maltreatment
87	Multidimensional Inventory for Assessment of Parental Functioning (Reis, Orme, Barbera-Stein, & Herz, 1987)	N/A	Not a measure of child maltreatment
88	Multidimensional Neglectful Behavior Scale: Adolescent and Adult Recall Version (Dubowitz et al., 2011)	MNBS-A	Not a parent-report measure
89	Multidimensional Neglectful Behavior Scale-Child Report (Beyazit & Ayhan, 2018)	MNBS-CR	Not a parent-report measure
90	National Council on Crime and Delinquency Indicators (Wood, 1997)	N/A	Not a parent-report measure
91	Needs-Based Assessment of Parental (Guardian) Support (Bolen, Lamb, & Gradante, 2002)	NAPS	Not a measure of child maltreatment
92	Neglect Scale (Harrington, Zuravin, DePanfilis, Ting, & Dubowitz, 2002)	N/A	Not a parent-report measure
93	Parent Cognition Scale[b] (Snarr, Slep, & Grande, 2009)	N/A	Not a measure of child maltreatment
94	Parent discipline style[b] (Mezzich et al., 2007)	N/A	Not a measure of child maltreatment
95	Parent Perception Inventory (Glaser, Horne, & Myers, 1995)	PPI	Not a measure of child maltreatment
96	Parent Perception Inventory-Child version (Bruce et al., 2006)	PPIC	Not a measure of child maltreatment
97	Parent Problem Checklist (Stallman, Morawska, & Sanders, 2009)	PPC	Not a measure of child maltreatment
98	Parent Qualities Measure (Crick, 2006; Stallman et al., 2009)	PQM	Not a measure of child maltreatment
99	Parent Threat Inventory (Crick, 2006; Scher, Stein, Ingram, Malcarne, & McQuaid, 2002)	PTI	Not a parent-report measure
100	Parental Acceptance-Rejection Questionnaire (Rohner & Khaleque, 2005)	PARQ	Not a parent-report measure
101	Parental Anger Inventory (Scher et al., 2002; Sedlar & Hansen, 2001)	PAI	Not a measure of child maltreatment
102	Parental Authority Questionnaire (Buri, 1991)	PAQ	Not a measure of child maltreatment
103	Parental Emotion Regulation Inventory (Lorber, Del Vecchio, Feder, & Smith Slep, 2017; Sedlar & Hansen, 2001)	PERI	Not a measure of child maltreatment
104	Parental Empathy Measure (Kilpatrick, 2005; Lorber et al., 2017)	PEM	Not a measure of child maltreatment
105	Parent-Child Activities Interview (Kilpatrick, 2005; Lefever et al., 2008)	PCA	Not a parent-report measure
106	PARENT-INFANT RELATIONSHIP GLOBAL ASSESSMENT SCALE (Lefever et al., 2008; THREE, 2005)	PIR-GAS	Not a measure of child maltreatment
107	Parenting Anxious Kids Ratings Scale-Parent Report (Flessner, Murphy, Brennan, & D'Auria, 2017; THREE, 2005)	PAKRS-PR	Not a measure of child maltreatment
108	Parenting Behavior Rating Scales (Flessner et al., 2017; G. A. King, Rogers, Walters, & Oldershaw, 1994)	N/A	Not a parent-report measure
109	Parenting Daily Diary (G. A. King et al., 1994; Peterson, Tremblay, Ewigman, & Popkey, 2002)	N/A	Not a parent-report measure
110	Parenting Practices Questionnaire-Corporal Punishment (Avinun, Davidov, Mankuta, Knafo-Noam, & Knafo-Noam, 2018)	PPQ-CP	Not a measure of child maltreatment

(Continued)

Appendix B. *(continued)*

No	Instrument[a] *(alphabetical order)*	Abbreviation	Reason for exclusion
111	Parenting Scale (Peterson et al., 2002; Salari, Terreros, & Sarkadi, 2012)	PS	Not a measure of child maltreatment
112	Parenting Support Needs Assessment (Murry & Lewin, 2014; Salari et al., 2012)	PSNA	Not a measure of child maltreatment
113	Plotkin Child Vignettes (Plotkin, 1983)	PCV	Not a measure of child maltreatment
114	Post-Divorce Parental Conflict Scale (Morris & West, 2000; Murry & Lewin, 2014)	PPCS	Not a measure of child maltreatment
115	Preschool Symptom Self-Report (Martini, Strayhorn, & Puig-Antich, 1990)	PRESS	Not a measure of child maltreatment
116	Production of Discipline Alternatives (Rodriguez, Wittig, & Christl, 2019)	PDA	Not a parent-report measure
117	Protective Factors Survey (Counts, Buffington, Chang-Rios, Rasmussen, & Preacher, 2010; Martini et al., 1990)	PFS	Not a measure of child maltreatment
118	Psychological Maltreatment Rating Scales (Brassard, Hart, & Hardy, 1993; Counts et al., 2010)	PMRS	Not a parent-report measure
119	Psychological Neglect (Brassard et al., 1993; Christ, Kwak, & Lu, 2017)	N/A	Not a parent-report measure
120	Psychologically Violent Parental Practices Inventory (Christ et al., 2017; Gagne, Pouliot-Lapointe, & St-Louis, 2007)	PVPPI	Not developed in English
121	Questionnaire for evaluating maltreatment and neglect (Calheiros, Patrício, Graça, & Magalhães, 2018)	N/A	Not developed in English
122	Reflective Parenting Assessment (Ensink, Leroux, Normandin, Biberdzic, & Fonagy, 2017; Gagne et al., 2007)	RPA	Not a measure of child maltreatment
123	Responsiveness Index (Ensink et al., 2017; Yates, Hull, & Huebner, 1983)	N/A	Not a parent-report measure
124	Revised Child Anxiety and Depression Scale Parent Version (Ebesutani, Tottenham, & Chorpita, 2015; Yates et al., 1983)	RCADS-P	Not a measure of child maltreatment
125	Risk Scale[b] (Ebesutani et al., 2015; Grietens, Geeraert, & Hellinckx, 2004)	N/A	Not a parent-report measure
126	Rorschach Inkblot Method (Choca, 2013; Grietens et al., 2004)	RIM	Not a measure of child maltreatment
127	Scale of Negative Family Interactions (Choca, 2013; Simonelli, Mullis, & Rohde, 2005)	SNFI	Not a parent-report measure
128	Screen for Adolescent Violence Exposure for children version (Flowers, Lanclos, & Kelley, 2002; Simonelli et al., 2005)	KID-SAVE	Not a parent-report measure
129	Sexual Abuse Indicators (Flowers et al., 2002; Terrell et al., 2008)	SAI	Not a parent-report measure
130	Sexual Behavior Problems Questionnaire[b] (Hall, Mathews, & Pearce, 1998; Terrell et al., 2008)	N/A	Not a parent-report measure
131	Sexual Events Questionnaire (Finkelhor, 1979; Hall et al., 1998)	SEQ	Not a parent-report measure
132	Sexual Experiences Survey (Finkelhor, 1979; Koss & Gidycz, 1985)	SES	Not a parent-report measure
133	Shaken Baby Syndrome Awareness Assessment (Koss & Gidycz, 1985; Russell & Britner, 2006)	SBS	Old version of a revised measure
134	Sixteen Personality Factor Questionnaire (Francis, Hughes, & Hitz, 1992; Russell & Britner, 2006)	16-PF	Not a measure of child maltreatment
135	Social Factors and Children Violence Questionnaire (Francis et al., 1992; Oni & Adetoro, 2014)	SPCVQ	No psychometric data found
136	Standardized Observation Codes (Cerezo, Keesler, Dunn, & Wahler, 1986; Oni & Adetoro, 2014)	SOC III	Not a measure of child maltreatment
137	Structured Problem Analysis of Raising Kids (Cerezo et al., 1986; Staal, van den Brink, Hermanns, Schrijvers, & van Stel, 2011)	SPARK	Not a measure of child maltreatment
138	Supervisory Neglect (Coohey, 2003; Staal et al., 2011)	N/A	Not a parent-report measure
139	Symptoms of Trauma Scale (Coohey, 2003; Ford et al., 2017)	SOTS	Not a measure of child maltreatment
140	Trauma Experiences Checklist (Cristofaro et al., 2013; Ford et al., 2017)	TEC	Not a measure of child maltreatment
141	Trauma history questionnaire (Cristofaro et al., 2013; Hooper, Stockton, Krupnick, & Green, 2011)	THQ	Not a parent-report measure
142	Trauma Symptom Checklist for Children (Briere et al., 2001; Hooper et al., 2011)	TSCC	Not a measure of child maltreatment
143	Trauma Symptom Checklist for Young Children (Briere et al., 2001)	TSCYC	Not a measure of child maltreatment

(Continued)

Appendix B. *(continued)*

No	Instrument[a] *(alphabetical order)*	Abbreviation	Reason for exclusion
144	U.S. military's Family Advocacy Program Severity Index (Briere et al., 2001; A. M. Slep & Heyman, 2004)	USAF-FAP Severity Index	Not a parent-report measure
145	Violent Experiences Questionnaire-Revised (A. R. King & Russell, 2017; A. M. Slep & Heyman, 2004)	VEQ-R	Not a parent-report measure
146	Weekly Problems Scales (A. R. King & Russell, 2017; Sawyer, Tsao, Hansen, & Flood, 2006)	WPS	Not a measure of child maltreatment
147	When Bad Things Happen Scale (Fletcher, 1995; Sawyer et al., 2006)	WBTH	Not a measure of child maltreatment
148	Young Parenting Inventory (Young, Klosko, & Weishaar, 2003)	YPI	Not a parent-report measure
149	Young Parenting Inventory-Revised (Louis, Wood, & Lockwood, 2018)	YPI-R2	Not a parent-report measure
150	Young Schema Questionnaire-Short form 3 (Young, 2005)	YSQ-S3	Not a parent-report measure

Notes. N/A = Not Applicable (No Abbreviation).

[a] References of the excluded instruments in this review are available from the first author upon request.

[b] Unofficial title retrieved from publication content as an instrument published without a title or abbreviation.

Appendix C. *Descriptions of the Development and Content Validity Studies on Included Instruments.*

Source[a] *(alphabetical order)*	Instrument	Purpose of study	Study population[b]	Age[c] (range [R] and/or Mean [MN] and/or Standard Deviation [SD])
Bavolek et al. (1979)	**Adult Adolescent Parenting Inventory-2 (AAPI-2)**	To develop and validate the AAPI (as an original version of the AAPI-2)	N = 9 (Stage: Construct development): (I) Professionals in child maltreatment N = 3,000 (Stage: Pilot Testing): (II) Adolescents attending high schools (grade 10-12)	(I) R = NR, MN = NR, SD = NR; (II) R= NR, MN = NR, SD = NR
Gordon et al. (1979)	**Intensity of Parental Punishment Scale (IPPS)**	To develop and validate the IPPS	N = 417: (I) n = 301: Parents of 5- to 10-year-old children; (II) n = 50: Upper-middle-class parents of 7- to 12-year old children; (III) n = 26: Mothers of 6- to 9-year-old children; (IV) n = 40: Mothers of 6- to 14-year-old children	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR; (III) R= NR, MN = NR, SD = NR; (IV) R = NR, MN = NR, SD = NR
Heyman et al. (2019)	**Family Maltreatment-Child Abuse criteria (FM-CA)**	To develop and validate the FM-CA	N = 126: U.S. Air Force service members and their spouses (F = 41; M = 85)	R = NR, MN = NR, SD = NR
Holden and Zambarano (1992)	**Parental Response to Child Misbehavior questionnaire (PRCM)**	To exam parental responses to children's misbehavior in maternal reported use of physical punishment by using the CPSS and the PRCM	N = 132: Mothers of 12- to 48-month-old children (F = 132; M = 0)	R = 20-44y, MN = 31.4y, SD = 4.5y
Lang and Connell (2017)	**Child Trauma Screen-Exposure Score (CTS-ES)**	To develop and validate the CTS-ES	N = 923 (Stage: CTS-ES Development): (I) Parents of children receiving care at outpatient behavioral health clinics N = 69 (Stage: CTS-ES Validation): (II) Parents of children receiving care at outpatient behavioral health clinics	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR
Loeber et al. (1998)	**Child Neglect Scales-Maternal Monitoring and Supervision scale (CNS-MMS)**	To examine delinquency, substance use, early sexual behavior, and mental health problems of urban boys by using diverse instruments including the SIS (as an original version of the CNS-MMS)	N = 1507: (I) n = 503: parents with boys in the first grade in Pittsburgh public schools; (II) n = 508: parents with boys in the fourth grade in Pittsburgh public schools (III) n = 506: parents with boys in the seventh grade in Pittsburgh	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR; (III) R = NR, MN = NR, SD = NR
Meinch et al. (2018)	**ISPCAN Child Abuse Screening Tool for use in Trials (ICAST-Trial)**	To develop and validate the ICAST-Trial	N = 115 (Stage: Pilot study) (I) Parents of adolescents participated in a parenting program to prevent child abuse (F = 112; M = 3) N = 552 (Stage: Validation of ICAST-Trial) (II) Parents of adolescents participated in a parenting program to prevent child abuse (F = 523; M = 29)	(I) R = NR, MN = 48y, SD = 13.6y; (II) R = NR, MN = 49.4y, SD = 14.69y
Runyan et al. (2009)	**ISPCAN Child Abuse Screening Tool for use in Trials (ICAST-Trial)**	To develop and validate the ICAST-P (as an original version of the ICAST-Trial)	N = 51 (Stage: Item development): (I) Professionals in child maltreatment N = 697 (Stage: Pilot Testing): (II) Parents with children under the age of 18 in six different countries	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR

(continued)

Appendix C. *(continued)*

Source[a] *(alphabetical order)*	Instrument	Purpose of study	Study population[b]	Age[c] (range [R] and/or Mean [MN] and/or Standard Deviation [SD])
Russell and Britner (2006)	**Shaken Baby Syndrome awareness assessment-Short Version (SBS-SV)**	To develop and evaluate the psychometric properties of the SBS (as an original version of the SBS-SV)	N = 288 (Stage: Pilot study) (I) Undergraduate psychology students (F = 207; M = 81) N = 264 (Stage: Validation of SBS) (II) Caregivers and non-caregivers over the age of 18 (F = 191; M = 73)	(I) R = 17-31y, MN = 19y, SD = NR; (II) R = 18-78y, MN = 32y, SD = NR
Straus et al. (1995)	**Mother-Child Neglect Scale (MCNS)**	To describe the development and validation of the MNBS (as an original version of the MCNS)	N = 359: Adolescences and adults (F = 236, M = 123)	R = NR, MN = NR, SD = NR
Straus et al. (1995)	**Mother-Child Neglect Scale-Short Form (MCNS-SF)**	To describe the development and validation of the MNBS-SF (as an original version of the MCNS-SF)	N = 359: Adolescences and adults (F = 236, M = 123)	R = NR, MN = NR, SD = NR
Straus et al. (1998)	**Conflict Tactics Scales: Parent-Child version (CTSPC)**	To develop and test the reliability and validity of CTSPC	N = 1,000: Parents of children under 18 years old participated in an U.S. national survey (F = 660; M = 340)	R = NR, MN = 36.8y, SD = NR
Stewart et al. (2015)	**Child Neglect Questionnaire (CNQ)**	To develop and evaluate psychometric properties of the CNQ	N = 172: (I) n = 76: Parents of children having fathers with Substance Use Disorder (SUD); (II) n = 96: Parents of children having fathers without SUD	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR
Twentyman et al. (1981)	**Parent Opinion Questionnaire (POQ)**	To develop and validate the POQ	N = 30 (Stage: Item development): (I) n = 23: Child protective case workers (II) n = 7: Health nurses N = 15 (Stage: Cross validation): (III) Child protective case workers	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR; (III) R = NR, MN = NR, SD = NR

Zaidi et al. (1989)	**Analog Parenting Task (APT)**	To determine whether there was an association between punitive childhood histories by the AEIII and abusive parenting by the APT.	N = 86 (Stage: preliminary study) (I) n= 49: university students experienced severe physical punishment in childhood (F = 19; M = 30); (II) n= 37: university students experienced mild physical punishment in childhood (F = 26; M = 11)
			N = 338 (Stage: main study) (III) n = 169: Mothers of children referred for child psychiatry service (F = 169; M = 0); (IV) n = 169: Fathers of children referred for child psychiatry service (F = 0; M = 169)
				(I) R = 18-24y, MN = 19.4y, SD = NR;
				(II): R = 17-23y, MN = 19.0y, SD = NR;
				(III) R = 22-51y, MN = 34.2y, SD = NR;
				(IV) R = 22-57y, MN = 36.8y, SD = NR

Notes. AAPI = Adult Adolescent Parenting Inventory; AEIII = Assessing Environments III; CAP = Child Abuse Potential inventory; CPSS = Computer-Presented Social Situations; ICAST-P = ISPCAN (International Society for the Prevention of Child Abuse and Neglect) Child Abuse Screening Tool-Parent version; MNBS = Multidimensional Neglectful Behavior Scale; MNBS-SF = Multidimensional Neglectful Behavior Scale-Short Form; SIS = Supervision and Involvement Scale; SBS = Shaken Baby Syndrome awareness assessment.

[a] References of the development and content validity studies on included instruments can be found in the reference section of this review.

[b] N = total sample size; n = subgroups; M = male; F = female.

[c] R = range; MN = mean; Med = median; NR = not reported; SD = standard deviation; NR = Not Reported.

Article 2

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 2: Internal Consistency, Reliability, Measurement Error, Structural Validity, Hypothesis Testing, Cross-Cultural Validity, and Criterion Validity. *Trauma, Violence, & Abuse.* Advanced online publication. https://doi.org/10.1177/1524838020915591

Review Manuscript

TRAUMA, VIOLENCE, & ABUSE
1-20
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1524838020915591
journals.sagepub.com/home/tva
SAGE

A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 2: Internal Consistency, Reliability, Measurement Error, Structural Validity, Hypothesis Testing, Cross-Cultural Validity, and Criterion Validity

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

Abstract

Aims: Child maltreatment (CM) is global public health issue with devastating lifelong consequences. Global organizations have endeavored to eliminate CM; however, there is lack of consensus on what instruments are most suitable for the investigation and prevention of CM. This systematic review aimed to appraise the psychometric properties (other than content validity) of all current parent- or caregiver-reported CM instruments and recommend the most suitable for use. **Method:** A systematic search of the CINAHL, Embase, ERIC, PsycINFO, PubMed, and Sociological Abstracts databases was performed. The evaluation of psychometric properties was conducted according to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines for systematic reviews of patient-report outcome measures. Responsiveness was beyond the scope of this systematic review, and content validity has been reported on in a companion paper (Part 1). Only instruments developed and published in English were included. **Results:** Twenty-five studies reported on selected psychometric properties of 15 identified instruments. The methodological quality of the studies was overall adequate. The psychometric properties of the instruments were generally indeterminate or not reported due to incomplete or missing psychometric data; high-quality evidence on the psychometric properties was limited. **Conclusions:** No instruments could be recommended as most suitable for use in clinic and research. Nine instruments were identified as promising based on current psychometric data but would need further psychometric evidence for them to be recommended.

Keywords
assessment, caregiver-reported measures, child abuse, child neglect, COSMIN, measurement properties, parent-reported measures

Child maltreatment (CM) is a major public health issue. More than half of the world's children (1 billion children aged 2–17 years) are exposed to CM (Hillis et al., 2016). Approximately 155,000 children younger than 15 years die worldwide annually as a result of CM (Gilbert et al., 2009), which is the second leading cause of childhood death (Johnson, 2002). Furthermore, early exposure to CM has resulted in short-term and long-term devastating consequences from childhood to adulthood, such as behavioral problems, poor academic performance in childhood (Boden et al., 2007; Godinet et al., 2014), mental health problems, and experiencing poverty in adulthood (Currie & Spatz Widom, 2010; Kisely et al., 2018; Sugaya et al., 2012).

[1] Department of Special Needs Education, Faculty of Educational Sciences, 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] Department of Social Work, Education and Community Wellbeing, Faculty of Health and Life Sciences, Northumbria University, Newcastle, United Kingdom
[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, Oslo 0371, Norway.
Email: sangwon.yoon@isp.uio.no

Due to the worldwide high prevalence and serious consequences of CM, the United Nations (UN) and World Health Organization (WHO) have urged that member states not only enact laws for the abolition of CM but also take action to investigate and prevent CM in each country (Hillis et al., 2016). In 1989, the UN (1989) presented the Convention on the Rights of the Child to protect children against all forms of abuse and neglect; the Convention was ratified by 196 member nations. Ten years later, the WHO (1999) published the Report of the Consultation on Child Abuse Prevention to provide global guidelines for investigation and prevention of CM based on international expert consensus. Recently, the UN (2015) has launched a new commitment to end CM as part of their 2030 Agenda for Sustainable Development Goals; all member states will evaluate their progress from 2016 to 2030 toward this goal for elimination of CM.

The task of monitoring progress toward elimination of CM is complicated by the trend that the prevalence of CM tends to underestimate the true incidence because information about the CM prevalence mostly relies on professional reports (from child protection workers, doctors, and teachers, who are mandated to report CM) rather than parent/carer or child reports (Shanahan et al., 2018). As CM usually occurs in private places, such as homes, in the absence of witnesses and is mostly perpetrated by parents (Institute of Medicine and National Research Council, 2014), actual incidences of CM are difficult to be accurately reported by individuals other than parents/carers or children. For this reason, parent/carer or child reports are the only way to determine the true incidence of CM that is committed, instead of relying on professional reports (Miller-Perrin & Perrin, 2013).

A recent meta-analysis on the prevalence of caregiver-perpetrated CM has shown that prevalence rates based on child reports is far lower than when based on caregiver reports (Devries et al., 2018) due to recall bias (i.e., difficulty remembering past events; Greenhoot, 2011; Milner & Crouch, 1997). In addition, even though caregiver reports on their own perpetration of CM appear not to underestimate, the accuracy of caregiver reports is still a subject for debate due to social desirability bias (i.e., the tendency to respond in a socially desirable way; Della Femina et al., 1990; Milner & Crouch, 1997). Thus, identifying high-quality parent or caregiver report instruments is essential to accurately estimate prevalence of CM.

The choice of high-quality instruments is strongly determined by having robust psychometric properties such as validity and reliability (Karanicolas et al., 2009). The best way to select the most reliable and valid instruments is to systematically review the literature on its psychometric properties (Scholtes et al., 2011). Good systematic reviews of psychometric properties of instruments should evaluate the quality of the studies on psychometric properties of an instrument, evaluate the quality of psychometric properties of an instrument, and synthesize the findings from all the psychometric studies using consensus-based standards and methods (Terwee et al., 2016). Recently, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) group has published guidelines for conducting systematic reviews on psychometric properties of patient-reported outcome instruments (Prinsen et al., 2018; Terwee et al., 2018). The COSMIN guidelines include the following practical tools: a taxonomy defining each psychometric property (Mokkink et al., 2010b), a checklist to assess methodological quality of psychometric studies (Mokkink, de Vet et al., 2018), criteria to assess each result of single study on a psychometric property (Prinsen et al., 2018; Terwee et al., 2018), and a rating system summarizing all results of studies on each psychometric property and grading quality of all evidence used for the assessments of both the methodological and the psychometric quality (Prinsen et al., 2018; Terwee et al., 2018).

The COSMIN taxonomy provides consensus-based terminology and definitions on nine psychometric properties, which forms the following three domains (Mokkink et al., 2010b): (1) validity (the extent to which an instrument measures the construct it is intended to measure), (2) reliability (the extent to which scores for patients who have not changed are the same for repeated measurements), and (3) responsiveness (the ability to detect clinically important change over time in the construct measured). The following psychometric properties are part of the validity domain (Mokkink et al., 2010b): (1) content validity (extent to which the content of an instrument adequately reflects the construct measured), (2) criterion validity (extent to which the scores adequately reflect a gold standard), and (3) construct validity (extent to which the scores are consistent with hypotheses based on the assumption that an instrument validly measures the construct measured). Construct validity is subdivided into the following three psychometric properties: (3.1) structural validity (extent to which the scores adequately reflect the dimensionality of the construct measured), (3.2) hypothesis testing (extent to which the scores are consistent with hypotheses on differences between relevant groups and relations to scores of other instruments), and (3.3) cross-cultural validity (extent to which a translated or culturally adapted version of an instrument adequately reflects the performance of the items of the original instrument). The following three psychometric properties comprise the reliability domain (Mokkink et al., 2010b): internal consistency (degree of the interrelatedness of items), reliability (the proportion of total score variance which is due to true differences among respondents), and measurement error (systematic and random error of a respondent's score that is not due to true changes in the construct being measured). Responsiveness is a separate domain (Mokkink et al., 2010b).

The most significant advantage of the COSMIN guidelines over other methods is that they were designed to assess the quality of *all* domains of psychometric properties comprehensively, while other methods were designed for evaluating limited aspects of psychometric properties only. For example, the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) checklist (Whiting et al., 2011) mainly focuses on the single measurement property of criterion validity (Christian et al., 2019), whereas the Quality Appraisal of Reliability

Studies (QAREL) checklist (Lucas et al., 2010) was designed for evaluating reliability only (Abedi et al., 2019). Furthermore, compared with the COSMIN guidelines, both the QUADAS-2 and QAREL checklists have more criteria that rely on subjective interpretation of psychometric reporting to determine the quality of psychometric studies (Abedi et al., 2019; Christian et al., 2019).

Another point of difference is that the COSMIN system deviates from earlier appraisal methods in that construct validity can be evaluated through hypothesis testing, structural validity, and cross-cultural validation. Hypothesis testing involves determining the presence and magnitude of relationships between items of instruments following the traditional multitrait-multimethod (MTMM) approach (Campbell & Fiske, 1959). In turn, structural validity should be evaluated by determining the relationships between the hypothesized and observed factor structure by conducting modern confirmatory factor analysis (CFA; Prinsen et al., 2018). According to the COSMIN guidelines, evidence on structural validity should be considered more important than hypothesis testing when recommending instruments in terms of construct validity (Prinsen et al., 2018), as CFA is a more robust approach than the MTMM in evaluating construct validity. The reasons are 2-fold: first, CFA is more accurate in determining measurement error than the MTMM (Gaither, 1993); and second, Campbell and Fiske's method (1959) were based on a subjective interpretation of rules of thumb criteria of the MTMM correlations, which lacked clear standards to differentiate satisfactory and unacceptable results (Shen, 2017). An additional advantage of using the COSMIN guidelines is that both traditional (classic test theory) and contemporary psychometric theories (item response theory) can be employed to evaluate the quality of psychometric properties of an instrument (Prinsen et al., 2018). However, although the COSMIN guidelines are comprehensive, precise, and balanced, it is complex and requires in-depth knowledge of psychometrics and quality rating criteria for conducting systematic reviews of the psychometric properties of an instrument (Christian et al., 2019; Dobbs et al., 2019).

To date, two systematic reviews have evaluated the psychometric characteristics of CM instruments: Kim et al. (2016) and Saini et al. (2019). Kim et al. (2016) conducted a systematic review to evaluate the methodological quality of studies reporting on the development of CM instruments using the 14 criteria of the QUADAS (Whiting et al., 2003), which is an assessment tool for methodological quality of psychometric studies. However, the authors did not evaluate the psychometric quality of the included instruments. Another systematic review by Saini et al. (2019) evaluated both the study quality and psychometric quality of the CM instruments. However, the authors mainly identified and evaluated child self-report and clinician-report interview instruments, excluding parent- or caregiver-reported CM instruments. Moreover, the authors did not use the latest, thoroughly revised COSMIN guidelines (Prinsen et al., 2018; Terwee et al., 2018), but instead used a previous version of the COSMIN checklist (Mokkink et al., 2010a) and criteria (Terwee et al., 2007) for quality assessment of included studies and instruments. The previous version of checklist and criteria does not have specific and comprehensive standards for assessing content validity, even though it is the most important psychometric property, nor do the guidelines have a standardized method to synthesize psychometric data (Prinsen et al., 2018; Terwee et al., 2018). To overcome these weaknesses of the previous version, the COSMIN guidelines (Prinsen et al., 2018; Terwee et al., 2018) were completely revised in recent years. The COSMIN guidelines recommend evaluating content validity of an instrument first because if it is unclear what construct(s) the instrument is actually measuring, the evaluation of the other psychometric properties is meaningless (Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018). In other words, if reviews find high-quality evidence that an instrument has insufficient content validity, the other psychometric properties of the instrument do not need to be further evaluated. Accordingly, the content validity of the parent- or caregiver-reported CM instruments was evaluated first in a companion paper (Part 1; Yoon et al., 2020). As no high-quality evidence of insufficient content validity was found, this present review (Part 2) continued to evaluate the other psychometric properties of the included parent- or caregiver-reported CM instruments. To date, no systematic review on the psychometric properties of parent- or caregiver-reported CM instruments has been published.

Study Aim

The aim of this systematic review (Part 2) was to evaluate psychometric properties (other than content validity) of all current parent- or caregiver-reported CM instruments and to recommend the most suitable parent- or caregiver-reported CM instruments using the COSMIN guidelines (Prinsen et al., 2018). Content validity has been evaluated and reported on in a companion paper (Part 1; Yoon et al., 2020).

Method

This systematic review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement (Moher et al., 2009) and the COSMIN guidelines (Prinsen et al., 2018). This review was conducted in four sequential steps (see Figure 1):

- Step 1: *Systematic literature search* formulating eligibility criteria (Step 1.1) and searching the literature and selecting studies (Step 1.2);
- Step 2: *Evaluation of the methodological quality of studies* on psychometric properties of instruments using the COSMIN Risk of Bias checklist;
- Step 3: *Evaluation of the psychometric properties of instruments* rating the result of single studies against the criteria for good psychometric properties (Step 3.1), summarizing all results of studies per instrument (Step 3.2), and grading the quality of evidence on psychometric properties (Step 3.3); and

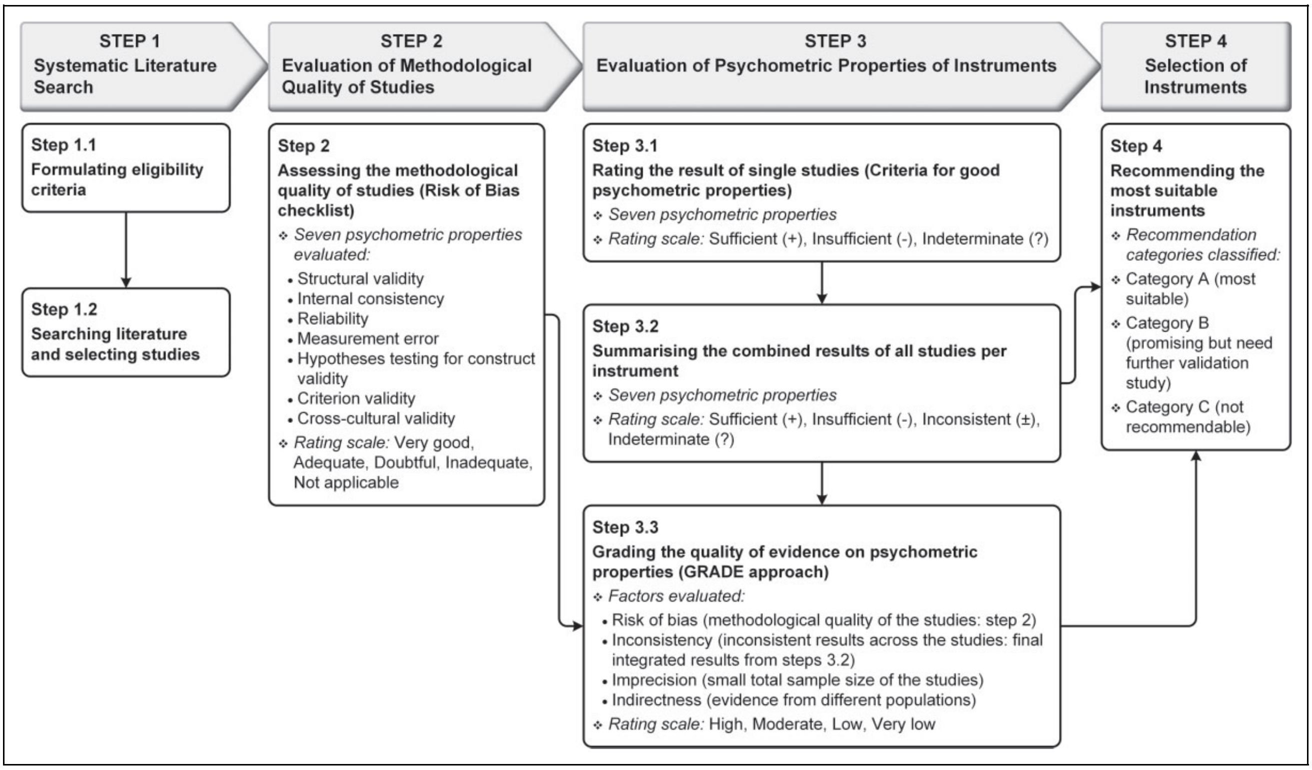

STEP 1
Systematic Literature Search
STEP 2
Evaluation of Methodological Quality of Studies
STEP 3
Evaluation of Psychometric Properties of Instruments
STEP 4
Selection of Instruments
Step 1.1
Formulating eligibility criteria
Step 1.2
Searching literature and selecting studies
Step 2
Assessing the methodological quality of studies (Risk of Bias checklist)
❖ Seven psychometric properties evaluated:
• Structural validity
• Internal consistency
• Reliability
• Measurement error
• Hypotheses testing for construct validity
• Criterion validity
• Cross-cultural validity
❖ Rating scale: Very good, Adequate, Doubtful, Inadequate, Not applicable
Step 3.1
Rating the result of single studies (Criteria for good psychometric properties)
❖ Seven psychometric properties
❖ Rating scale: Sufficient (+), Insufficient (-), Indeterminate (?)
Step 3.2
Summarising the combined results of all studies per instrument
❖ Seven psychometric properties
❖ Rating scale: Sufficient (+), Insufficient (-), Inconsistent (±), Indeterminate (?)
Step 3.3
Grading the quality of evidence on psychometric properties (GRADE approach)
❖ Factors evaluated:
• Risk of bias (methodological quality of the studies: step 2)
• Inconsistency (inconsistent results across the studies: final integrated results from steps 3.2)
• Imprecision (small total sample size of the studies)
• Indirectness (evidence from different populations)
❖ Rating scale: High, Moderate, Low, Very low
Step 4
Recommending the most suitable instruments
❖ Recommendation categories classified:
❖ Category A (most suitable)
❖ Category B (promising but need further validation study)
❖ Category C (not recommendable)

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. *Note*. Responsiveness was outside the scope of this review; Content validity was evaluated in a companion paper (Part 1; Yoon et al., 2020).

- Step 4: *Selection of instruments* recommending the most suitable instruments.

Each of these steps will be further described in the sections that follow.

Step 1: Systematic Literature Search

Systematic literature search for this review was performed in two substeps: formulating eligibility criteria (Step 1.1) and searching literature and selecting studies (Step 1.2). These two steps are in agreement with the PRISMA statement (Moher et al., 2009).

Eligibility criteria (Step 1.1). To be included for this review, instruments needed to meet the following four eligibility criteria: (1) parent or caregiver report instruments; (2) instruments were developed and published in English; (3) instruments assessed parents' or caregivers' attitude toward CM or perpetration of CM; (4) to ensure that an instrument reflects an overarching construct of CM, at least one subscale or a minimum of 30% of all items within an instrument measured one or more of the four main types of CM, including physical abuse (acts causing actual or potential physical harm to a child), emotional abuse (acts having adverse impact on the child's emotional development), sexual abuse (acts using a child for sexual gratification), neglect (failure providing for the development of a child in health, education, emotional development, nutrition, shelter, and safe living conditions; Krug et al., 2002; WHO, 1999).

The following two additional selection criteria were used for psychometric studies: (1) Journal articles and manuals were published in English; (2) reported psychometric data of at least one of the following eight psychometric properties as defined in the COSMIN taxonomy (Mokkink et al., 2010b): structural validity, internal consistency, reliability, measurement error, hypotheses testing for construct validity, criterion validity, cross-cultural validity, and content validity. Responsiveness was beyond the scope of the present review, and content validity was assessed in a companion paper (Part 1; Yoon et al., 2020).

Literature search and study selection (Step 1.2). Systematic literature searches were conducted in six electronic databases: CINAHL, Embase, ERIC, PsycINFO, PubMed, and Sociological Abstracts. All database searches were conducted in January 2018 with an updated search conducted in October 2019. Subject headings and free text words were used to search databases and to retrieve all journal articles up until October 2019 (see Supplementary Appendix A).

Abstracts identified by database searches were screened to retrieve eligible instruments and full-text articles on any psychometric property by two independent reviewers. One reviewer screened all abstracts while the other reviewer screened a randomly selection of half of all abstracts. All full texts of eligible abstracts were extracted and screened independently by two reviewers. Any differences between two reviewers were resolved through consensus with a third reviewer. The interrater agreement was assessed by calculating weighted κ (Cohen & Humphreys, 1968) and interpreted as very good (0.81–1.00), good (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), and poor (0.00–0.20; Altman, 1991).

Next, reference lists of all included full texts were hand searched to identify additional eligible instruments and studies. Websites of two major publishers of measurements in social science (Pearson and Western Psychological Services) were also searched to identify potential instruments and manuals. Both searches for reference lists and websites were conducted by one reviewer and the identified additional instruments and studies were checked by the other reviewer. When instruments were not published or available for free, the developers of the instruments were contacted to obtain the original instruments.

Step 2: Evaluation of Methodological Quality of Studies

The methodological quality of the studies on the psychometric properties of the included instruments was rated using the COSMIN Risk of Bias checklist (Mokkink, de Vet et al., 2018), which is a standardized tool for evaluating study quality of psychometric studies. The checklist contains 3–38 items for each psychometric property (Mokkink, de Vet et al., 2018). The checklist items rate the quality of study design and the robustness of statistical analyses conducted in studies on any of the seven psychometric properties evaluated in this article (Mokkink, de Vet et al., 2018). Evaluation of reliability included all three aspects (Mokkink et al., 2010b): test–retest reliability (the degree of total score variance in repeated measurement on the same patients over time), interrater reliability (the degree of total score variance in repeated measurement on the same occasions by different raters), and intrarater reliability (the degree of total score variance in repeated measurement on different occasions by the same rater). Cross-cultural validity was evaluated for measurement invariance of an instrument across culturally different groups (e.g., nationality, gender, and age) within English-speaking populations only (Mokkink, de Vet et al., 2018), due to including only instruments developed and published in English in this review. Furthermore, evaluation of criterion validity involved exploring associations between an instrument and a gold standard, as well as between an original long version and the shortened version thereof (Mokkink, Prinsen, et al., 2018). Lastly, hypothesis testing for construct validity was evaluated by appraising the associations between two instruments to determine whether they are measuring a similar construct of interest (i.e., convergent validity) and to compare differences in scores between subgroups of the target population (i.e., discriminative validity; Mokkink, de Vet et al., 2018).

When rating the methodological quality of the included studies on psychometric properties, each checklist item was ranked on a 4-point rating scale: 1 = *inadequate*, 2 = *doubtful*, 3 = *adequate*, and 4 = *very good* (Mokkink, de Vet et al., 2018). A total rating for each psychometric property was obtained by calculating the ratio between "the obtained total score minus the minimum score possible' and 'the maximum score possible minus the minimum score possible" (Cordier et al., 2015). This approach was adopted instead of a worst score counts method (i.e., reporting total ratings obtained by taking the lowest rating among any of the checklist items) recommended by COSMIN guideline (Mokkink, Prinsen, et al., 2018), as determining the total ratings entirely based on the lowest rating single item tends to impede the detection of subtle differences in methodological quality between studies (Speyer et al., 2014). Therefore, the total score of methodological quality ratings per psychometric property was presented as a percentage of the ratings: inadequate (0%–25%), doubtful (25.1%–50%), adequate (50.1%–75%), and very good (75.1%–100%). Two reviewers rated the methodological quality independently, and any discrepancies were resolved by consensus. The interrater agreement between two reviewers was determined by calculating the weighted κ (Cohen & Humphreys, 1968).

After evaluating methodological quality of the included psychometric studies, the following data were extracted from the included studies and instruments (Mokkink, Prinsen, et al., 2018): (1) study characteristics (i.e., study purpose, assessed psychometric properties, and study population); (2) instrument characteristics (i.e., instrument names, construct to be measured, target population, purpose of use, number of [sub] scales and items, and response options and recall period); and (3) study results on seven psychometric properties (internal consistency, reliability, measurement error, structural validity, hypothesis testing, cross-cultural validity, and criterion validity). One reviewer extracted all relevant data from included studies, and the other reviewer checked the extracted data for accuracy and completeness.

Step 3: Evaluation of Psychometric Properties of Instruments

The psychometric properties of instruments were assessed for each of seven psychometric properties in three consecutive steps: Step 3.1 rating the result of single studies, Step 3.2 summarizing the results of all studies per instrument, and Step 3.3 grading the quality of evidence on psychometric properties. All ratings were conducted by two reviewers independently where after consensus ratings were determined by discussion between reviewers.

Rating the result of single studies (Step 3.1). Rating the results of single studies was conducted for each psychometric property separately. The results of each psychometric property in each individual study were rated as sufficient (above the quality criteria threshold: +), insufficient (below the quality criteria threshold: −), or indeterminate (less robust data that do not meet the quality criteria:?), using the predefined criteria for good psychometric properties (Mokkink, Prinsen, et al., 2018; see Supplementary Appendix B).

Summarizing the results of all studies per instrument (Step 3.2). All results on each psychometric property from available studies per instrument were qualitatively summarized into overall ratings of the psychometric property per instrument (Prinsen et al., 2018). An overall sufficient (+), insufficient (−) inconsistent (±), or indeterminate (?) rating was given for each psychometric property per instrument, with a 75% agreement rule used (Mokkink, Prinsen, et al., 2018): that is, for an overall sufficient (+) or insufficient (−) rating on a psychometric property, 75% or more of the studies reporting the psychometric property must be sufficient (+) or insufficient (−); otherwise, for an overall inconsistent (±) rating, less than 75% of studies showed the same rating; and for overall indeterminate (?) rating, all studies must be indeterminate (?).

Grading the quality of evidence on psychometric properties (Step 3.3). The quality of the evidence (i.e., the total body of evidence used for overall ratings on each psychometric property of an instrument) was graded as high, moderate, low, or very low using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Prinsen et al., 2018; see Supplementary Appendix C). The GRADE approach considers the initial quality of evidence used for overall ratings to be high, but the evidence quality is subsequently downgraded by one or more levels (to moderate, low, or very low) if there are serious (one level down: −1), very serious (two levels down: −2), or extremely serious (three levels down: −3) concerns. The following four factors were considered in determining the ratings: (a) risk of bias (limitations in the methodological quality of studies: Step 2), (b) inconsistency (unexplained heterogeneity in results of studies: Step 3.2), (c) indirectness (evidence from different populations than the targeted population in the review), and (d) imprecision (a low total number of samples included in the studies; Mokkink, Prinsen, et al., 2018). For example, for downgrading one level (from *high* to *moderate*), only one factor is allowed to have a serious concern (−1); for two levels (from *high* to *low*), either only one factor with a very serious concern (−2) or two factors with serious concerns (−1) is allowed; for three levels (from *high* to *very low*), one factor with an extremely serious concern (−3), one factor with very serious concern (−2), and one factor with serious (−1) to extremely serious concerns (−3), or more than three factors with serious (−1) to extremely serious concerns (−3) is allowed. Quality of evidence was not graded when the overall rating was indeterminate (?) as this indicates lack of robust evidence (Prinsen et al., 2018). Further details on grading quality of evidence can be found in the COSMIN usual manual for systematic reviews of instruments (Mokkink, Prinsen, et al., 2018).

Step 4: Selection of Instruments

The selection of instruments and recommendation of suitable instruments for future use was based on combining overall rating results of each psychometric property (Step 3.2) and grading results of evidence quality for each property (Step 3.3; Prinsen et al., 2018). The recommendation was based on both findings of content validity (Part 1) and other psychometric properties (Part 2) of included instruments. Each instrument was classified into three recommendation categories (Mokkink, Prinsen, et al., 2018): (A) most suitable (i.e., instruments with high-quality evidence for sufficient content validity—in any aspects of relevance, comprehensiveness, and comprehensibility—and at least low-quality evidence for sufficient internal consistency); (B) promising but need further validation studies (i.e., instruments categorized not in A or C); and (C) not recommendable (i.e., instruments with high-quality evidence for an insufficient psychometric property).

To determine suitable instruments, content validity and internal consistency were considered as decisive psychometric properties rather than other properties because if it is unclear what an instrument is actually measuring and how different items in the instrument are related with construct to be measured, the evaluation of the other psychometric properties is meaningless. Furthermore, this review did not consider interpretability (the degree to which clinical meaning can be assigned to an instrument's quantitative scores or change in scores) and feasibility (ease of use such as length, completion time, and access fee of an instrument) to recommend the most suitable CM instruments because neither interpretability nor feasibility is considered psychometric properties (Prinsen et al., 2018).

Results

Systematic Literature Search

A total of 2,859 abstracts (removing duplicates) were retrieved from six databases: 1,173 records from CINAHL; 456 records from Embase; 523 records from ERIC; 285 records from PsycINFO; 1,092 records from PubMed; and 133 records from Sociological Abstracts. Figure 2 presents the flow chart of the studies and instruments identified during the searching literature and selecting studies (Step 1.2) according to the PRISMA (Moher et al., 2009). In total, 253 full-text articles and 164 instruments were assessed for eligibility, of which 23 articles and 14 instruments met all inclusion criteria: a list of the 150 excluded instruments and reasons for exclusion are provided in Supplementary Appendix D. Reference checking of the included 23 full-text articles identified two additional studies (one article and one manual) and one additional instrument met all inclusion criteria. As a result, 25 studies reporting and analyzing psychometric properties of 15 parent or carer report CM instruments were included in this review. The interreviewer agreement for study selection between two reviewers was very good (Altman, 1991): weighted κ for abstract selection = 0.87 (95% confidence interval [CI] = [0.83, 0.90]); weighted κ for article selection = 0.86 (95% CI [0.77, 0.94]).

Characteristics of Included Studies and Instruments

General characteristics of the psychometric studies of included CM instruments are presented in Supplementary Appendix E. Table 1 summarizes the characteristics of the included 15

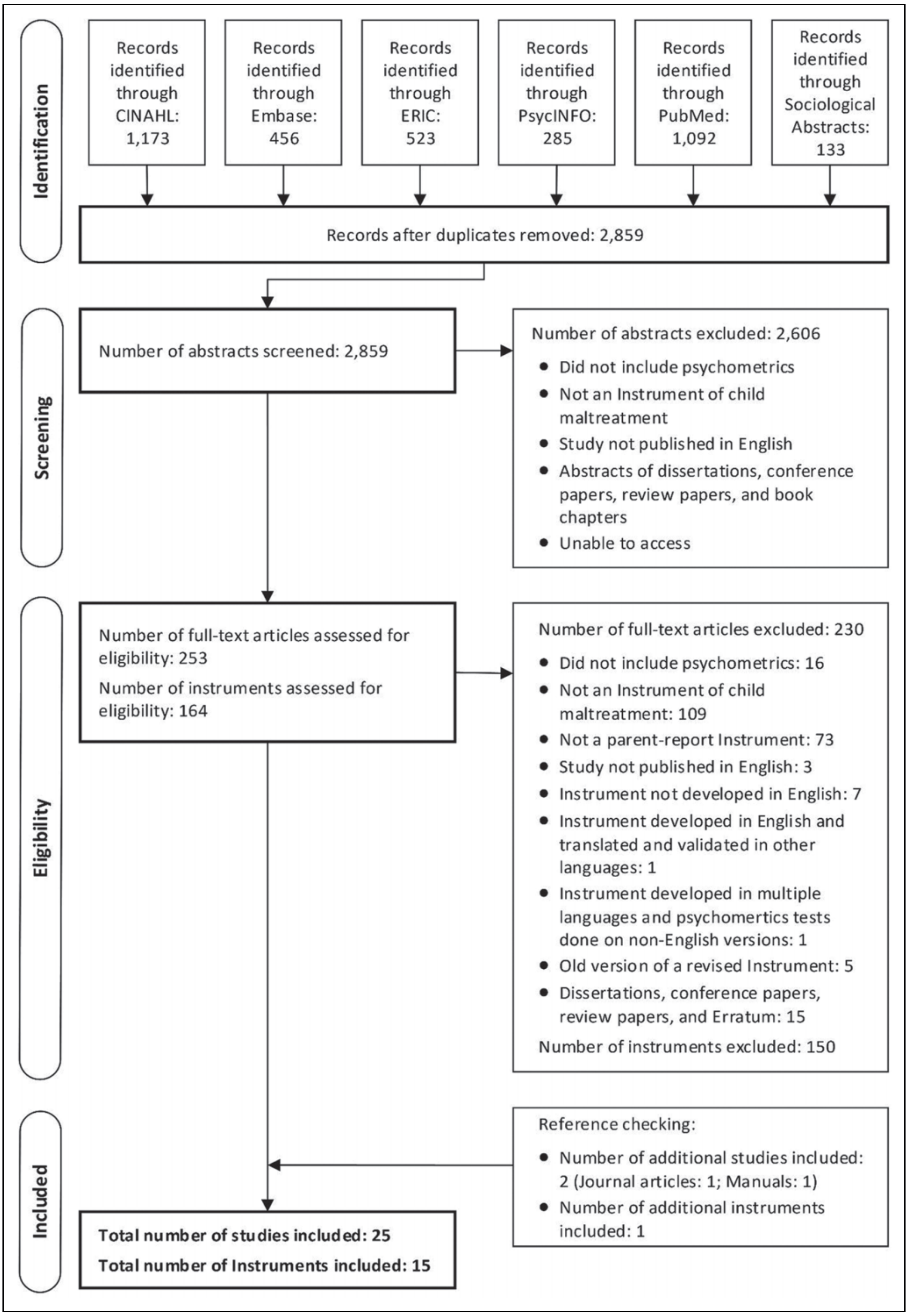

Identification
Records identified through CINAHL: 1,173
Records identified through Embase: 456
Records identified through ERIC: 523
Records identified through PsycINFO: 285
Records identified through PubMed: 1,092
Records identified through Sociological Abstracts: 133
Records after duplicates removed: 2,859
Screening
Number of abstracts screened: 2,859
Number of abstracts excluded: 2,606
● Did not include psychometrics
● Not an Instrument of child maltreatment
● Study not published in English
● Abstracts of dissertations, conference papers, review papers, and book chapters
● Unable to access
Eligibility
Number of full-text articles assessed for eligibility: 253
Number of instruments assessed for eligibility: 164
Number of full-text articles excluded: 230
● Did not include psychometrics: 16
● Not an Instrument of child maltreatment: 109
● Not a parent-report Instrument: 73
● Study not published in English: 3
● Instrument not developed in English: 7
● Instrument developed in English and translated and validated in other languages: 1
● Instrument developed in multiple languages and psychomertics tests done on non-English versions: 1
● Old version of a revised Instrument: 5
● Dissertations, conference papers, review papers, and Erratum: 15
Number of instruments excluded: 150
Included
Reference checking:
● Number of additional studies included: 2 (Journal articles: 1; Manuals: 1)
● Number of additional instruments included: 1
Total number of studies included: 25
Total number of Instruments included: 15

Figure 2. Flow diagram of the reviewing procedure based on Preferred Reporting Items for Systematic reviews and Meta-Analyses (Moher et al., 2009).

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

Instrument (References)	Construct	(Sub)scales	Target Population	Purpose of Use	Number of Items	Range of Score	Response Options	Recall Period
AAPI-2 (Bavolek & Keene, 1999; Conners et al., 2006; Lawson et al., 2017; Rodriguez et al., 2011; Russa & Rodriguez, 2010)	Abusive and neglecting parenting practices	Five (sub)scales: 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	Identification of maltreating parents/carers; Evaluation of intervention	40	0–50 (Raw total scores per subscale are converted into standard scores: range 0–10)	5-point ordinal scale (*strongly disagree* = 1 to *strongly disagree* = 5)	Not specified
APT (Rodriguez et al., 2011; Russa & Rodriguez, 2010)	Attitude toward physical discipline	Two (sub)scales: Physical discipline; Escalation of physical discipline	Prospective parent populations	Identification of maltreating parents/carers	26	0–26	10 nominal scale (from nonphysical discipline tactics to physical discipline tactics)	Not specified
CNQ (Stewart et al., 2015)	Child neglect	Four (sub)scales: Physical neglect; Emotional neglect; Educational neglect; Supervision neglect	Parents with older children	Identification of maltreating parents/carers	46	46–184	4-point ordinal scale (*always* = 1 to *never* = 4)	Past 6 months
CNS-MMS (Kirisci et al., 2001)	Child neglect	One (sub)scales: Child neglect	Mothers	Evaluation of intervention	11	11–33	3-point ordinal scale (*hardly ever* = 1 to *often* = 3)	Past 6 months
CTS-ES (Lang & Connell, 2017)	Potentially traumatic event (including childhood physical abuse, sexual abuse, and domestic or community violence)	One (sub)scale: Potentially traumatic event	Caregivers	Identification of children maltreated by parents/carers	4	0–4	Dichotomous scale (*no* = 0 or *yes* = 1)	Not specified
CTSPC (Compier-de Block et al., 2017; Grasso et al., 2016; Kobulsky et al., 2017; Lorber & Slep, 2017; O'Dor et al., 2017; Rodriguez, 2010; Straus et al., 1998)	Physical and psychological child abuse	Three (sub)scales: Nonviolent discipline; Psychological aggression; Physical assault	Parents	Identification of maltreating parents/carers; Evaluation of intervention	22	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)	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*)	Past 1 year
FM-CA (Heyman et al., 2019)	Clinically significant child abuse and neglect	Two (sub)scales: Physical child abuse; Psychological child abuse	Parents	Identification of maltreating parents/carers; Evaluation of intervention	27	0–63	Dichotomous scale for physical child abuse subscale (*I did* = 0 or *I never did* = 1); 6-point ordinal scale for psychological child abuse subscale (*never* = 0 to *more than once a day* = 5)	Past 1 year
ICAST-Trial (Meinck et al., 2018)	Child abuse and neglect	Four (sub)scales: Physical abuse; Emotional abuse; Contact sexual abuse; Neglect	Caregivers	Evaluation of intervention	14	0–112	9-point ordinal scale (*never* = 0 to *more than 8 times* = 8)	Past 1 month
IPPS (Gordon et al., 1979)	Intensity of parent behavioral responses to hypothetical child misbehavior situations	Five (sub)scales: School misbehavior; Disobedience after a recent reminder; Public disobedience; Crying; Destructiveness	Parents	Identification of maltreating parents/carers; Evaluation of intervention	33	33–231	7-point ordinal scale (*no reaction* = 1 to *very strong punishment* = 7)	Not specified
MCNS (Lounds et al., 2004)	Maternal neglectful behavior towards their children	Four (sub)scales: Emotional neglect; Cognitive neglect; Supervisory neglect; Physical needs neglect	Mothers	Identification of maltreating parents/carers	20	20–80	4-point ordinal scale (*strongly disagree* = 1 to *strongly agree* = 4)	Past 1 year
MCNS-SF (Lounds et al., 2004)	Maternal neglectful behavior towards their children	Two (sub)scales: Emotional neglect; Cognitive neglect; Supervisory neglect; Physical needs neglect	Mothers	Identification of maltreating parents/carers	8	4–32	4-point ordinal scale (*strongly disagree* = 1 to *strongly agree* = 4)	Past 1 year

(continued)

Table 1. (continued)

Instrument (References)	Construct	(Sub)scales	Target Population	Purpose of Use	Number of Items	Range of Score	Response Options	Recall Period
P-CAAM (Rodriguez et al., 2011)	Acceptance of parent-child aggression	Two (sub)scales: Physical discipline; Physical abuse	Current and prospective parent populations	Evaluation of intervention	8 video clips: 90 sec each	0–NR	Clips builds towards "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
POQ (Azar & Rohrbeck, 1986; Haskett et al., 2006; Mammen et al., 2003)	Parental expectations of child behavior	Six (sub)scales: Self-care; Family responsibility and care of siblings; Help and affection to parents; Leaving children alone; Proper behavior and feelings; Punishment	Parents	Identification of maltreating parents/ carers	60	0–60	Dichotomous scale (*disagree* = 0 or *agree* = 1)	Not specified
PRCM (Vittrup et al., 2006)	Discipline techniques in response to children's misbehaviors	One (sub)scale: Discipline techniques	Parents with young children	Identification of maltreating parents/ carers; Evaluation of intervention	12	0–72	6-point ordinal scale (never = 0–9 ≥ times per week = 6)	Past one week
SBS-SV (Russell, 2010)	Shaken baby syndrome awareness	Three (sub)scales: Soothing techniques; Discipline techniques; Potential for injury	Parents and caregivers of young children	Evaluation of intervention	36	36–216	6-point ordinal scale (*strongly disagree* = 1 to *strongly agree* = 6)	Not specified

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.

instruments. All but three instruments were multidimensional, having some subscales to measure a range of different facets of CM, while the remaining instruments were a unidimensional scale. The majority of the instruments (14/15) were designed for current parent or carer respondents, except one instrument that was designed for prospective parents (i.e., before or during pregnancy) to reduce the risk of future CM. Ten instruments had a purpose of use for identifying maltreating parents/carers and/or evaluating intervention programs; four instruments for evaluating intervention programs; and one for identifying abused children by parents/carers.

Methodological Quality of the Included Studies

The methodological quality of the 25 included studies (24 articles and 1 manual) was assessed using the COSMIN Risk of Bias checklist (Mokkink, de Vet et al., 2018). Some studies measured more than one psychometric property and included more than one instrument: the studies were rated multiple times for each psychometric property and instrument, respectively. For all 29 studies (including four duplicates), an overview of all methodological quality ratings is displayed in Table 2. Most studies reported on hypotheses testing for construct validity (25/29) and internal consistency (21/29). Only a small number of studies included psychometric data on structural validity (10 studies), reliability (5 studies), cross-cultural validity (1 study), and criterion validity (1 study). No information was retrieved on measurement error in any study. The interreviewer agreement for quality assessment of included studies between both reviewers was very good: weighted $\kappa = 0.86$ (95% CI [0.83, 0.90]).

Psychometric Properties and Quality of Evidence of the Instruments (Step 3)

Table 3 summarizes ratings for each psychometric property for single studies, respectively (Step 3.1). All data on a psychometric property extracted from the 25 included studies were evaluated against the criteria for good psychometric properties for the seven psychometric properties reported in this article (Prinsen et al., 2018). A summary of rating criteria is presented in detail in Supplementary Appendix B.

Table 4 presents the overall ratings (Step 3.2) and the quality of evidence (Step 3.3) for each psychometric property per instrument; the results of all included studies on each psychometric property per instrument and their quality ratings are summarized in Supplementary Appendix F. None of the instruments reported overall ratings for all seven psychometric properties, given that measurement error was not reported (NR) for any of the 15 instruments. Furthermore, grades for quality of evidence were reported in only 21% (22 of 105 possible ratings) of all overall ratings on psychometric quality for all 15 instruments, while all other quality of evidence was rated as NR due to no psychometric data reported or not evaluated due to less robust psychometric data reported (i.e., indeterminate overall ratings).

Recommendations for the Most Suitable Instruments to Measure CM (Step 4)

Table 5 provides the recommendations for the use of parent or carer report instruments to measure CM in the future. None of instruments were rated as the most suitable; nine instruments (AAPI-2, APT, CNS-MMS, CTS-ES, FM-CA, IPPS, P-CAAM, PRCM, and SBS-SV) were considered the most promising but would still need further validation studies; six instruments (CNQ, CTSPC, ICAST-Trial, MCNS, MCNS-SF, and POQ), however, were not recommendable.

Discussion

The purpose of this systematic review was to evaluate the quality of psychometric properties (other than content validity and responsiveness) of all current parent/caregiver report instruments on CM by parents or caregivers and recommend the most suitable of these instruments using the COSMIN guidelines. This review identified 15 instruments and 25 studies on psychometric properties of these instruments. In general, the methodological quality of included studies was adequate. However, most of the identified instruments (12/15) reported on only three or less psychometric properties of the seven properties under review. Furthermore, there are limited high-quality evidence to suggest that any of the psychometric properties are inherently sufficient or insufficient. Therefore, most CM instruments (9/15) have the potential to be used in research and in clinical practice, but their psychometric quality should undergo further evaluation.

Methodological Quality of the Included Studies

For structural validity, all but six instruments (AAPI-2, CNQ, CNS-MMS, CTSPC, ICAST-Trial, and IPPS) did not report any psychometric data or reported doubtful study quality. The doubtful study quality is due to using a less preferred factor analysis method, such as the exploratory factor analysis (EFA). The EFA can be used to identify a factor structure of new instruments without any prior hypothesis of the structure, while structural validity is to test a hypothesized factor structure of existing instruments (Mokkink, Prinsen, et al., 2018). To test the hypothesized factor structure, confirmative factor analysis (CFA) or item response theory (IRT) analysis was preferred in the COSMIN Risk of Bias checklist (Mokkink, de Vet et al., 2018). While having the same overall purpose for testing how well the data fit a predetermined factor structure (de Vet et al., 2011), the specific concerns of each analysis differ. That is, CFA focuses on total summed scores or responses because it assumes each item is equally weighted in terms of difficulty, whereas IRT analysis is concerned with individual responses to items under the assumption individual items may have different difficulty level (Lo et al., 2015). However, neither of these two analyses had been conducted for the factor structure of 10 instruments (APT, CTS-ES, FM-CA, IPPS, MCNS, MCNS-SF, P-CAAM, POQ, PRCM, and SBS-SV).

Table 2. Methodological Quality Assessment of Studies on Psychometric Properties of the Included Instruments.

		Psychometric Property: Methodological Quality per Study[a]					
Instrument	Reference	Structural Validity	Internal Consistency	Cross-Cultural Validity	Reliability	Criterion Validity	Hypotheses Testing
AAPI-2	Bavolek and Keene (1999)	Very good (88.9%)	Very good (100.0%)	NR	NR	NR	Adequate (55.6%)
	Conners et al. (2006)	Very good (100.0%)	Very good (77.8%)	NR	NR	NR	Very good (81.3%)
	Lawson et al. (2017)	Adequate (66.7%)	Very good (100.0%)	NR	NR	NR	Adequate (66.7%)
	Rodriguez et al. (2011)	NR	Adequate (66.7%)	NR	NR	NR	Very good (100.0%)
	Russa and Rodriguez (2010)	NR	NR	NR	NR	NR	Very good (100.0%)
APT	Rodriguez et al. (2011)	NR	Very good (100.0%)	NR	NR	NR	Very good (83.3%)
	Russa and Rodriguez (2010)	NR	Very good (77.8%)	NR	NR	NR	Very good (90.0%)
CNQ	Stewart et al. (2015)	Adequate (75.0%)	Doubtful (33.3%)	NR	NR	NR	Very good (91.2%)
CNS-MMS	Kirisci et al. (2001)	Very good (100.0%)	Very good (100%)	NR	NR	NR	Very good (100.0%)
CTS-ES	Lang and Connell (2017)	NR	NR	NR	NR	NR	Very good (91.7%)
CTSPC	Compier-de Block et al. (2017)	NR	Very good (88.9%)	NR	Very good (77.8%)	NR	Adequate (55.6%)
	Cotter et al. (2018)	Very good (77.8%)	Adequate (55.6%)	NR	NR	NR	Very good (83.3%)
	Grasso et al. (2016)	NR	Very good (100.0%)	NR	NR	NR	NR
	Kobulsky et al. (2017)	NR	NR	NR	Very good (100.0%)	NR	NR
	Lorber and Slep (2017)	Very good (100.0%)	Adequate (58.3%)	NR	NR	NR	NR
	O'Dor et al. (2017)	NR	Very good (100.0%)	NR	NR	NR	Very good (100.0%)
	Rodriguez (2010)	NR	NR	NR	NR	NR	Very good (91.7%)
	Straus et al. (1998)	NR	Adequate (66.7%)	NR	NR	NR	Adequate (66.7%)
FM-CA	Heyman et al. (2019)	NR	NR	NR	NR	NR	Doubtful (41.7%)
ICAST-Trial	Meinck et al. (2018)	Very good (100.0%)	Very good (100.0%)	NR	NR	NR	Very good (91.7%)
IPPS	Gordon et al. (1979)	Adequate (55.6%)	Very good (77.8%)	Inadequate (25.0%)	Doubtful (26.7%)	NR	Adequate (54.1%)
MCNS	Lounds et al. (2004)	NR	Very good (100.0%)	NR	Adequate (73.3%)	NR	Very good (83.3%)
MCNS-SF	Lounds et al. (2004)	NR	Very good (77.8%)	NR	NR	Very good (100.0%)	Very good (83.3%)
P-CAAM	Rodriguez et al. (2011)	NR	Adequate (66.7%)	NR	NR	NR	Very good (89.2%)
POQ	Azar and Rohrbeck (1986)	NR	NR	NR	Doubtful (33.3%)	NR	Very good (77.8%)
	Haskett et al. (2006)	Doubtful (33.3%)	Very good (77.8%)	NR	NR	NR	Very good (82.8%)
	Mammen et al. (2003)	NR	NR	NR	NR	NR	Very good (77.3%)
PRCM	Vittrup et al. (2006)	NR	NR	NR	NR	NR	Very good (77.8%)
SBS-SV	Russell (2010)	NR	Very good (100.0%)	NR	NR	NR	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] Responsiveness was beyond the scope of this review; Measurement error is not displayed since it was not reported in any study; The methodological quality was rated using the consensus-based standards for the selection of health measurement instruments checklist (Mokkink, de Vet et al., 2018): very good, adequate, doubtful, and inadequate. 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 (due to no psychometric data reported).

Table 3. Quality of the Psychometric Properties per Study.

		Psychometric Property: Quality of Psychometric Properties per Study[a]					
Instrument	Reference	Structural Validity	Internal Consistency	Cross-Cultural Validity	Reliability	Criterion Validity	Hypotheses Testing
AAPI-2	Bavolek and Keene (1999)	?	?	NR	NR	NR	±
	Conners et al. (2006)	−	?	NR	NR	NR	−
	Lawson et al. (2017)	±	?	NR	NR	NR	−
	Rodriguez et al. (2011)	NR	?	NR	NR	NR	±
	Russa and Rodriguez (2010)	NR	NR	NR	NR	NR	−
APT	Rodriguez et al. (2011)	NR	?	NR	NR	NR	−
	Russa and Rodriguez (2010)	NR	?	NR	NR	NR	±
CNQ	Stewart et al. (2015)	+	+	NR	NR	NR	−
CNS−MMS	Kirisci et al. (2001)	+	+	NR	NR	NR	−
CTS-ES	Lang and Connell (2017)	NR	NR	NR	NR	NR	±
CTSPC	Compier-de Block et al. (2017)	NR	?	NR	−	NR	+
	Cotter et al. (2018)	?	?	NR	NR	NR	−
	Grasso et al. (2016)	NR	?	NR	NR	NR	NR
	Kobulsky et al. (2017)	NR	NR	NR	?	NR	NR
	Lorber and Slep (2017)	?	?	NR	NR	NR	NR
	O'Dor et al. (2017)	NR	?	NR	NR	NR	−
	Rodriguez (2010)	NR	NR	NR	NR	NR	−
	Straus et al. (1998)	NR	?	NR	NR	NR	−
FM-CA	Heyman et al. (2019)	NR	NR	NR	NR	NR	?
ICAST-Trial	Meinck et al. (2018)	+	−	NR	NR	NR	−
IPPS	Gordon et al. (1979)	?	?	?	?	NR	±
MCNS	Lounds et al. (2004)	NR	?	NR	?	NR	−
MCNS-SF	Lounds et al. (2004)	NR	?	NR	NR	+	−
P-CAAM	Rodriguez et al. (2011)	NR	?	NR	NR	NR	±
POQ	Azar and Rohrbeck (1986)	NR	NR	NR	?	NR	+
	Haskett et al. (2006)	?	?	NR	NR	NR	−
	Mammen et al. (2003)	NR	NR	NR	NR	NR	−
PRCM	Vittrup et al. (2006)	NR	NR	NR	NR	NR	+
SBS-SV	Russell (2010)	NR	?	NR	NR	NR	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] Responsiveness was beyond the scope of this review; Measurement error is not displayed since it was not reported in any study; The psychometric properties was rated using the criteria for good psychometric properties (Prinsen et al., 2018); + = sufficient; ? = indeterminate (due to less robust psychometric data); − = insufficient; ± = inconsistent (in case of rating one more results per psychometric property within a study, if < 75% of ratings displayed the same scoring); NR = not reported (due to no psychometric data); Data and ratings on each psychometric property per study are available in the Supplementary Appendix F.

None of the instruments reported on all three psychometric properties within the domain of reliability (Mokkink et al., 2010b). Only four instruments (CTSPC, IPPS, MCNS, and POQ) reported reliability, while all but three instruments (CTS-ES, FM-CA, and PRCM) reported internal consistency. Even though measurement error is clinically very relevant information, none of the instruments reported measurement error. This is an important limitation to note as instruments with low error are able to detect clinically important changes sensitively and help clinicians to decide when to adjust treatment plans or to terminate treatment if the intervention has shown to have successfully addressed the underlying problem (Dvir, 2015; Guyatt et al., 1987). Consequently, the lack of reporting on all three of these psychometric properties makes it difficult to grasp overall reliability for all instruments comprehensibly.

Only one instrument (MCNS-SF) reported criterion validity between the shortened and an original (long) version; the MCNS-SF received a very good score for study quality. As there is no universally accepted gold standard to measure CM (Bailhache et al., 2013), this aspect of criterion validity could not be reported on in this review. In addition, cross-cultural validity for different demographic groups was reported for only one instrument (IPPS), with an inadequate score for study quality due to not reporting information on what kinds of factor analysis was used, despite comparing factor structures between mother and father respondents. Among culturally different groups using the same language, the same question may

Table 4. Overall Quality of Psychometric Properties and Evidence Quality per Instrument.

	Psychometric Property: Quality of Psychometric Properties and Quality of Evidence per Instrument											
	Structural Validity		Internal Consistency		Cross-Cultural Validity		Reliability		Criterion Validity		Hypotheses Testing	
Instrument	Overall Rating[a]	Quality of Evidence[b]	Overall rating[a]	Quality of Evidence[b]	Overall Rating[a]	Quality of Evidence[b]	Overall Rating[a]	Quality of Evidence[b]	Overall Rating[a]	Quality of Evidence[b]	Overall Rating[a]	Quality of Evidence[b]
AAPI-2	±	Moderate	?	NE	NR	NR	NR	NR	NR	NR	–	Moderate
APT	NR	NR	?	NE	NR	NR	NR	NR	NR	NR	±	Very Low
CNQ	+	Moderate	+	Low	NR	NR	NR	NR	NR	NR	–	High
CNS-MMS	+	High	+	High	NR	NR	NR	NR	NR	NR	–	Moderate
CTS-ES	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	±	Low
CTSPC	?	NE	?	NE	NR	NR	–	Moderate	NR	NR	–	High
FM-CA	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	?	NE
ICAST-Trial	+	High	–	High	NR	NR	NR	NR	NR	NR	–	High
IPPS	?	NE	?	NE	?	NE	?	NE	NR	NR	±	Low
MCNS	NR	NR	?	NE	NR	NR	?	NE	NR	NR	–	High
MCNS-SF	NR	NR	?	NE	NR	NR	NR	NR	+	High	–	High
P-CAAM	NR	NR	?	NE	NR	NR	NR	NR	NR	NR	±	Low
POQ	?	NE	?	NE	NR	NR	?	NE	NR	NR	–	High
PRCM	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	+	High
SBS-SV	NR	NR	?	NE	NR	NR	NR	NR	NR	NR	NR	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 overall quality of psychometric properties was rated using the criteria for good psychometric properties (Mokkink, Prinsen, et al., 2018); + = sufficient rating; ? = indeterminate rating (due to less robust psychometric data); – = insufficient rating; ± = inconsistent rating; NR = not reported (due to no psychometric data); Data and ratings on each psychometric property per instrument are available in the Supplementary Appendix F. [b] The quality of evidence (confidence level for the overall quality rating of each psychometric property) was rated using a modified GRADE approach (Mokkink, Prinsen, 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, NR = not reported (due to not reported overall rating of psychometric properties); NE = not evaluated (due to indeterminate overall rating); If the evidence quality is very low, we should be concerned about using the overall ratings alone to recommend good instruments; Reasons for each grading on quality of evidence are available in the Supplementary Appendix F.

be interpreted differently. For example, "spanking" (as the most common form of corporal punishment) may be perceived as child abuse to parents in New Zealand but as discipline to American parents because corporal punishment is illegal (in all settings) in New Zealand but is legal if done at home in American (Elgar et al., 2018). This difference in interpretations between countries that speak the same language but show cultural differences may result in different underlying factor structures of the same instrument. For this reason, applying the same instruments to culturally different groups also requires testing measurement invariance across the different groups, even if they speak the same language.

Hypothesis testing for construct validity was reported for all instruments with ratings of either adequate or very good quality, except for the following two instruments: FM-CA received doubtful rating, and SBS-SV was NR. Seven instruments (APT, CNS-MMS, CTS-ES, FM-CA, ICAST-Trial, MSCNS, and MCNS-SF) reported on convergent validity only, calculating correlations between the scores of the seven instruments and a comparator CM instrument. One instrument (PRCM) reported on discriminative validity only, analyzing statistical differences in scores between parents who perpetrated CM and parents who did not. For six instruments (AAPI-2, CNQ, CTSPC, IPPS, P-CAAM, and POQ), both convergent and discriminative validity were reported. Except these six instruments, the imbalance between convergent and discriminative validity of the remaining instruments, therefore, has limited evidence for construct validity.

Psychometric Properties of the Instruments

The evidence on structural validity is a prerequisite for interpreting the evidence on internal consistency (i.e., the interrelatedness of items in each scale or subscale; Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018). For example, if results on structural validity show that a scale has four factors, internal consistency of each of those four subscales is more relevant than that of the total scale. As such, evidence on structural validity directly affected the overall ratings of internal consistency. Of the 12 instruments reporting evidence on internal consistency, only two instruments (CNQ and CNS-MMS) displayed sufficient internal consistency, CNQ with moderate

Table 5. Recommendations on Suitable Instruments for Their Future Use Adapted From Prinsen et al. (2018).

Category	Description on Category (Criteria)	Instruments	
A: Most suitable	Instruments that have the potential to be recommended for use in respect of the construct and population of interest (*instruments with high-quality evidence for sufficient content validity in any aspects of and at least low-quality evidence for sufficient internal consistency*)	None	
B: Promising but need further validation study	Instruments that may have the potential to be recommended for use, but further validation studies are needed (*instrument categorised not in A or C*)	• AAPI-2 • APT • CNS-MMS • CTS-ES • FM-CA	• IPPS • P-CAAM • PRCM • SBS-SV
C: Not recommendable	Instruments that should not be recommended for use (*instruments with high-quality evidence for an insufficient psychometric property*)	• CNQ • CTSPC • ICAST-Trial	• MCNS • MCNS-SF • POQ

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.

evidence (due to only one adequate study available) for sufficient structural validity and high Cronbach's α values and CNS-MMS with high evidence (due to very good study quality, consistent results, adequate sample sizes, and same populations between studies) for sufficient structural validity and a high Cronbach's α. Conversely, five instruments (APT, MCNS, MCNS-SF, P-CAAM, and SBS-SV) did not report any data on structural validity; three instruments (CTSPC, IPPS, and POQ) reported indeterminate structural validity due to using a less robust factor analysis (EFA) or presenting only incomplete information on the structure of the instruments; one instrument (AAPI-2) reported conflicting results on the factor structure between studies. As these nine instruments (AAPI-2, APT, CTSPC, IPPS, MCNS, MCNS-SF, P-CAAM, POQ, and SBS-SV) demonstrated poor structural validity by not meeting the criteria of "at least low evidence for sufficient structural validity," their internal consistency was therefore rated as indeterminate. Although one instrument (ICAST-Trial) reported high evidence for sufficient structural validity, internal consistency of the instrument was rated as insufficient due to a low Cronbach's α.

Of four instruments reporting the evidence on reliability (test–retest, interrater, and intrarater reliability), three instruments (IPPS, MCNS, and POQ) gained indeterminate overall ratings because of reporting other reliability statistics (e.g., Spearman's correlation coefficients and κ) than the preferred reliability statistics in the COSMIN criteria for good psychometric properties (Prinsen et al., 2018). The COSMIN criteria prefer the intraclass correlation coefficient (ICC) or the weighted κ as appropriate reliability statistics because in contrast to the Spearman's ρ coefficient, the ICC takes into account systematic error caused by different conditions and learning effects in repeated measurements for continuous scales (Scholtes et al., 2011); the weighted κ takes into account the degree of disagreement between two raters for categorical scales whereas the unweighted κ does not (Tang et al., 2015). Although one instrument (CTSPC) reported ICC, reliability of the instrument was rated as insufficient (due to the ICC below the criterion for good reliability) with moderate evidence quality (due to some evidence from different population such as children).

Evidence on criterion validity of the shorten version of MCNS (MCNS-SF) was sufficient because the correlation with the original long version (MCNS) was over 0.70, which is the criterion for good criterion validity. In addition, evidence on cross-cultural validity was evaluated for only one instrument (IPPS), with an indeterminate overall rating, due to incomplete information on the measurement invariance of the instruments between two different groups. For good cross-cultural validity of an instrument, evidence on measurement invariance between culturally different groups (i.e., age, gender, language) should be found in factor structures at the scale level by performing CFA (Gregorich, 2006) or in item difficulty at item level by performing differential item functioning (DIF) analysis (Teresi et al., 2009). However, none of the instruments included in this review reported clear evidence on the measurement invariance between the different groups by using CFA or DIF analysis.

Evidence on hypothesis testing for construct validity was evaluated for all instruments except the SBS-SV. More than half of the instruments (8 of 15) reported insufficient hypothesis testing with high or moderate evidence quality: six instruments (CNQ, CTSPC, ICAST-Trial, MCNS, MCNS-SF, and POQ) had high-quality evidence while other two instruments (AAPI-2 and CNS-MMS) had moderate evidence (due to some evidence from different population such as university students who are not parents or caregivers). Conversely, only one instrument (PRCM) reported sufficient hypothesis testing with high-quality evidence. Four instruments (APT, CTS-ES, IPPS, and P-CAAM) reported conflicting results between studies on hypothesis testing, with low or very low evidence quality; only one instrument (FM-CA) reported indeterminate hypothesis testing due to using inappropriate statistical methods for comparison between FM-CA and a comparator CM instrument (i.e., calculating interrater agreement between two different measures rather than correlation). Furthermore, most hypothesis testing of instruments presented and considered only a *t*-value or *F*-value to confirm the statistical significance of the difference in scores between two groups (e.g., parents who perpetrated CM and parents who did not). However, these two statistics depend on sample size and do not account for the direction or magnitude of difference (Coe, 2002). To avoid this weakness of both statistics, this review converted the *t*-value or *F*-value to an effect size estimate (i.e., Cohen's *d*) showing the direction and magnitude of differences between two groups regardless of sample sizes (Friedman, 1968; Thalheimer & Cook, 2002); an effect size of 0.5 or higher was used as a criterion for sufficient hypothesis testing on group differences. For this reason, some of the hypotheses, which were originally confirmed based on the *t*-value or *F*-value in the studies on hypothesis testing of the instruments, were rejected (insufficient rating) in our review based on the converted Cohen's *d*.

Recommendation of the Instruments (Step 4)

None of the included instruments have the potential to be recommended as the most suitable (category A) due to no high-quality evidence for sufficient content validity in a companion paper (Part 1; Yoon et al., 2020) and no at least low-quality evidence for sufficient internal consistency in this article (Part 2), while six instruments (CNQ, CTSPC, ICAST-Trial, MCNS, MCNS-SF, and POQ) should not be recommended at all (category C) due to high-quality evidence for insufficient hypotheses testing or internal consistency. As having no high-quality evidence for an insufficient psychometric property, nine instruments (AAPI-2, APT, CNS-MMS, CTS-ES, FM-CA, IPPS, P-CAAM, PRCM, and SBS-SV) may have potential to be recommended but need further validation studies (category B).

For each of the nine promising instruments, further validation studies on one or more properties are needed to determine whether the nine promising instruments could be recommendable (i.e., category A). As a criterion for category A, content validity, internal consistency, and/or structural validity (not the criterion but as a prerequisite for internal consistency) of all nine instruments should be further evaluated as a priority. In a companion paper (Part 1; Yoon et al., 2020), no high-quality evidence for content validity of any promising instruments (except FM-CA) was found due to missing data or lack of robust evidence in the content validity studies. For this reason, future studies on content validity may provide additional information and result in changed overall quality ratings of evidence for content validity. In addition, the internal consistency of most instruments (except CNS-MMS) was scored as NR due to no information of their internal consistency or indeterminate (?) due to no information of their structural validity. As such, the CTS-ES and PRCM require urgently further studies on their content validity, structural validity, and internal consistency due to no high-quality evidence on these psychometric properties; the AAPI-2, APT, CTS-ES, IPPS, P-CAAM, PRCM, and SBS-SV require further studies on their content validity and structural validity due to no high evidence for content validity and indeterminate internal consistency caused by unclarity around the unidimensionality of a scale or subscale (i.e., indeterminate or conflicting structural validity); the CNS-MMS requires further content validity studies due to no high evidence for content validity and high evidence for sufficient internal consistency; and the FM-CA requires further studies on its structural validity and internal consistency due to no evidence for these psychometric properties.

To confirm whether the six instruments (CNQ, CTSPC, ICAST-Trial, MCNS, MCNS-SF, and POQ) should indeed not be recommended, further validation studies on hypotheses testing and/or internal consistency need to be conducted. All six instruments were categorized into "not recommendable" (category C) due to high-quality evidence for insufficient hypotheses testing, while ICAST-Trial had high evidence for insufficient internal consistency—another reason for not being recommended. However, most hypotheses testing focused on comparisons between different instruments (convergent validity) rather than differences between groups (discriminative validity): that is, the ratio between the amount of hypotheses on convergent validity and discriminative validity is 5–1 in the CNQ; 7–5 in the CTSPC; 1–0 in the ICAST-Trial; 3–0 in the MCNS; 3–0 in the MCNS-SF; and 14–4 in the POQ. As the vast majority of evidence were based on convergent validity, hypotheses testing of the six instruments showed mostly one side of hypotheses testing without data on discriminative validity. To capture the overall picture of hypotheses testing, further discriminative validity studies of the six instruments are needed. These additional studies may change the assessment of the five of the six instruments (except ICAST-Trial) from not recommendable (category C) to promising (category B). In the case of ICAST-Trial, further studies on both hypotheses testing and internal consistency are needed.

Limitations

This systematic review has some limitations. First of all, only instruments validated in English and studies published in English were included. Thus, some findings on psychometric properties of CM instruments published in other languages may have been excluded. Secondly, this review did not report on all of nine psychometric properties of the COSMIN taxonomy (Mokkink et al., 2010b); responsiveness was not considered for this review because evaluation of responsiveness would require to review all studies that have used the identified instruments as an outcome measure and would require a different search strategy altogether. Lastly, interpretability and feasibility were outside the scope of this article because they are not considered to be psychometric property according to the COSMIN taxonomy, even though these two instrument characteristics should be considered when recommending the most suitable instruments (Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018).

From a feasibility perspective, ideally instruments should have the least amount of items required to fully capture the construct under investigation to reduce the response time, particularly when it comes to investigating sensitive issues such as CM.

Implication for Future Research

For researchers who want to comprehensively understand the overall psychometric properties of all current parent- or carer-reported CM instruments, this systematic review highlights the need for further validation studies of the instruments. Regarding structural validity, future factor analyses using CFA or IRT are needed for nine instruments (AAPI-2, APT, CTSPC, IPPS, MCNS, MCNS-SF, P-CAAM, POQ, and SBS-SV) to determine the quality of internal consistency of these nine instruments. To gain a comprehensive picture of reliability, all three elements of reliability should be assessed: internal consistency for CTS-ES, FM-CA, and PRCM; reliability (test–retest, interrater, and intrarater) for AAPI-2, APT, CNQ, CNS-MMS, CTS-ES, FM-CA, ICAST-Trial, MCNS-SF, P-CAAM, PRCM, and SBS-SV; and measurement error for all 15 instruments. In particular, ICC or weighted κ are required to be calculated and reported in future studies for test–retest, interrater, and intrarater reliability, rather than Spearman's ρ or κ. With respect to cross-cultural validity, all 15 instruments (including IPPS with indeterminate cross-cultural validity) are needed to test measurement invariance across culturally different groups by performing CFA (Gregorich, 2006) or DIF analysis (Teresi et al., 2009). More hypothesis testing for construct validity should be conducted to determine convergent validity of the FM-CA, PRCM, and SBS-SV, and discriminative validity of the APT, CNS-MMS, CTS-ES, FM-CA, ICAST-Trial, MCNS, MCNS-SF, and SBS-SV. In particular, discriminative validity regarding differences in scores between groups should be based on the calculation of effect sizes such as Cohen's *d* rather than *t*-values or *F*-values.

Apart from the suggestion of further validation studies on the psychometric properties of the identified instruments, the current results in this review support the need of future instrument development research of new parent/carer report instruments on CM as none of the included instruments on CM in this review could be identified or recommended as best instrument; and suggest some implications for the future development of a good instruments on CM. For good content validity as the most important psychometric property (Terwee et al., 2018), the items of a new instrument should be identified by an interview or survey with parents/carers to reflect respondents' perspective on CM. This interview or survey with respondents was rarely done in the development studies for the existing 15 instruments on CM according to the findings of review in a companion paper (Part 1; Yoon et al., 2020), thus having a negative impact on the content validity. Next, for good internal consistency as the second most important property, robust factor analysis such as CFA or IRT should be conducted to identify a clear factor structure (good structural validity) as a prerequisite for internal consistency according to the Risk of Bias checklist (Mokkink, de Vet et al., 2018). Thirdly, for good psychometric properties in general, appropriate statistics for each psychometric property need to be calculated and reported on, in accordance with the criteria for good psychometric properties (Prinsen et al., 2018). Lastly, for high-quality evidence on each psychometric property, new parent/carer report instruments on CM should be developed against the standards set out in the COSMIN Risk of Bias checklist (Mokkink, de Vet et al., 2018): that is, appropriate study design and robust statistical analysis would ensure good methodological quality (no concern regarding risk of bias), consistent results across the psychometric studies (no concern regarding inconsistency), precision of the evidence by using appropriate sample size (no concern regarding imprecision), and direct evidence from targeted population such as parents or caregivers (no concern regarding indirectness) in terms of evidence quality according to the GRADE approach (Prinsen et al., 2018).

Conclusion

This systematic review evaluated the psychometric properties of 15 parent- or caregiver-reported CM instruments using the COSMIN guidelines. Evidence concerning psychometric properties was limited and mostly of lower quality. Based on current available psychometric evidence, none of the included instruments met the requirements to be recommended as most suitable instrument. Only nine instruments (AAPI-2, APT, CNS-MMS, CTS-ES, FM-CA, IPPS, P-CAAM, PRCM, and SBS-SV) were recommended as promising but would still need further validation before any possible recommendations as most suitable instrument may be made.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

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

Supplemental Material

Supplemental material for this article is available online.

References

Abedi, A., Prinsen, C. A. C., Shah, I., Buser, Z., & Wang, J. C. (2019). Performance properties of health-related measurement instruments in whiplash: Systematic review protocol. *Systematic Reviews*, *8*(1), 199–199. https://doi.org/10.1186/s13643-019-1119-0

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

Azar, S. T., & Rohrbeck, C. A. (1986). Child abuse and unrealistic expectations: Further validation of the parent opinion questionnaire. *Journal of Consulting and Clinical Psychology*, *54*(6), 867–868. https://doi.org/10.1037/0022-006X.54.6.867

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*(1), 202. https://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.

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*(10), 1101–1114. https://doi.org/10.1016/j.chiabu.2007.03.022

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Christian, B., Armstrong, R., Calache, H., Carpenter, L., Gibbs, L., & Gussy, M. (2019). A systematic review to assess the methodological quality of studies on measurement properties for caries risk assessment tools for young children. *International Journal of Paediatric Dentistry*, *29*(2), 106–116. https://doi.org/10.1111/ipd.12446

Coe, R. (2002, September 12–14). *It's the effect size, stupid: What "effect size" is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association*, University of Exeter, UK. http://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J., & Humphreys, L. H. (1968). Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Compier-de Block, L. H. C. G., Alink, L. R. A., Linting, M., van den Berg, L. J., 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*(2), 207–217. https://doi.org/10.1007/s10896-016-9902-3

Conners, N. A., Whiteside-Mansell, L., Deere, D., Ledet, T., & Edwards, M. C. (2006). Measuring the potential for child maltreatment: The reliability and validity of the Adult Adolescent Parenting Inventory–2. *Child Abuse & Neglect*, *30*(1), 39–53. https://doi.org/10.1016/j.chiabu.2005.08.011

Cordier, R., Speyer, R., Chen, Y. W., 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*(7), e0132299. https://doi.org/10.1371/journal.pone.0132299

Cotter, A., Proctor, K. B., & Brestan-Knight, E. (2018). Assessing child physical abuse: An examination of the factor structure and validity of the Parent-Child Conflict Tactics Scale (CTSPC). *Children and Youth Services Review*, *88*, 467–475. https://doi.org/10.1016/j.childyouth.2018.03.044

Currie, J., & Spatz Widom, C. (2010). Long-term consequences of child abuse and neglect on adult economic well-being. *Child Maltreatment*, *15*(2), 111–120. https://doi.org/10.1177/1077559509355316

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge University Press.

Della Femina, D., Yeager, C. A., & Lewis, D. O. (1990). Child abuse: Adolescent records vs. adult recall. *Child Abuse & Neglect*, *14*(2), 227–231. https://doi.org/10.1016/0145-2134(90)90033-P

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., Butron Riveros, B. C., Watts, C., & Abrahams, N. (2018). Who perpetrates violence against children? A systematic analysis of age-specific and sex-specific data. *BMJ Paediatrics Open*, *2*(1), e000180. https://doi.org/10.1136/bmjpo-2017-000180

Dobbs, T. D., Gibson, J. A. G., Hughes, S., Thind, A., Patel, B., Hutchings, H. A., & Whitaker, I. S. (2019). Patient-reported outcome measures for soft-tissue facial reconstruction: A systematic review and evaluation of the quality of their measurement properties. *Plastic and Reconstructive Surgery*, *143*(1), 255–268. https://doi.org/10.1097/prs.0000000000005112

Dvir, Z. (2015). Difference, significant difference and clinically meaningful difference: The meaning of change in rehabilitation. *Journal of Exercise Rehabilitation*, *11*(2), 67–73. https://doi.org/10.12965/jer.150199

Elgar, F. J., Donnelly, P. D., Michaelson, V., Gariépy, G., Riehm, K. E., Walsh, S. D., & Pickett, W. (2018). Corporal punishment bans and physical fighting in adolescents: An ecological study of 88 countries. *BMJ Open*, *8*(9), e021616. https://doi.org/10.1136/bmjopen-2018-021616

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, *70*(4), 245–251. https://doi.org/10.1037/h0026258

Gaither, C. A. (1993). Evaluating the construct validity of work commitment measures: A confirmatory factor model. *Evaluation & the Health Professions*, *16*(4), 417–433. https://doi.org/10.1177/016327879301600405

Gilbert, R., Widom, C. S., Browne, K., Fergusson, D., Webb, E., & Janson, S. (2009). Burden and consequences of child maltreatment in high-income countries. *The Lancet*, *373*(9657), 68–81. https://doi.org/10.1016/S0140-6736(08)61706-7

Godinet, M. T., Li, F., & Berg, T. (2014). Early childhood maltreatment and trajectories of behavioral problems: Exploring gender and racial differences. *Child Abuse & Neglect*, *38*(3), 544–556. https://doi.org/10.1016/j.chiabu.2013.07.018

Gordon, D. A., Jones, R. H., & Nowicki, S. (1979). A measure of intensity of parental punishment. *Journal of Personality Assessment*, *43*(5), 485–496. https://doi.org/10.1207/s15327752jpa4305_9

Grasso, D. J., Henry, D., Kestler, J., Nieto, R., Wakschlag, L. S., & Briggs-Gowan, M. J. (2016). Harsh parenting as a potential mediator of the association between intimate partner violence and child disruptive behavior in families with young children. *Journal of Interpersonal Violence*, *31*(11), 2102–2126. https://doi.org/10.1177/0886260515572472

Greenhoot, A. F. (2011). Retrospective methods in developmental science. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 196–210). The Guilford Press.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11 Suppl 3), S78–S94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, *40*(2), 171–178. https://doi.org/10.1016/0021-9681(87)90069-5

Haskett, M. E., Scott, S. S., Willoughby, M., Ahern, L., & Nears, K. (2006). The parent opinion questionnaire and child vignettes for use with abusive parents: Assessment of psychometric properties. *Journal of Family Violence*, *21*(2), 137–151. https://doi.org/10.1007/s10896-005-9010-2

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*. Advanced online publication. 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*(3), e20154079. https://doi.org/10.1542/peds.2015-4079

Institute of Medicine and National Research Council. (2014). Describing the problem. In A. C. Petersen, J. Joseph, & M. Feit (Eds.), *New directions in child abuse and neglect research* (pp. 31–68). National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK195982/

Johnson, C. F. (2002). Child maltreatment 2002: Recognition, reporting and risk. *Pediatrics International*, *44*(5), 554–560. https://doi.org/10.1046/j.1442-200X.2002.01642.x

Karanicolas, P. J., Bhandari, M., Kreder, H., Moroni, A., Richardson, M., Walter, S. D., Norman, G. R., & Guyatt, G. H., & Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group. (2009). Evaluating agreement: Conducting a reliability study. *Journal of Bone and Joint Surgery*, *91*(Suppl 3), 99–106. https://doi.org/10.2106/JBJS.H.01624

Kim, H., Choi, H., & Park, H. (2016). A systematic review of child abuse screening instruments. *Child Health Nursing Research*, *22*(4), 265–278. https://doi.org/10.4094/chnr.2016.22.4.265

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*(4), 241–255. https://doi.org/10.1023/a:1013662132189

Kisely, S., Abajobir, A. A., Mills, R., Strathearn, L., Clavarino, A., & Najman, J. M. (2018). Child maltreatment and mental health problems in adulthood: birth cohort study. *The British Journal of Psychiatry*, *213*(6), 698–703. https://doi.org/10.1192/bjp.2018.207

Kobulsky, J. M., Kepple, N. J., Holmes, M. R., & Hussey, D. L. (2017). Concordance of parent- and child-reported physical abuse following child protective services investigation. *Child Maltreatment*, *22*(1), 24–33. https://doi.org/10.1177/1077559516673156

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*(3), 390–398. https://doi.org/10.1037/tra0000235

Lawson, M. A., Alameda-Lawson, T., & Byrnes, E. (2017). Analyzing the validity of the adult-adolescent parenting inventory for low-income populations. *Research on Social Work Practice*, *27*(4), 441–455. https://doi.org/10.1177/1049731514567154

Lo, C., Liang, W. M., Hang, L. W., Wu, T. C., Chang, Y. J., & Chang, C. H. (2015). A psychometric assessment of the St. George's respiratory questionnaire in patients with COPD using Rasch model analysis. *Health and Quality of Life Outcomes*, *13*(1), 131. https://doi.org/10.1186/s12955-015-0320-7

Lorber, M. F., & Slep, A. M. (2017). The reliability paradox of the parent-child conflict tactics corporal punishment subscale. *Journal of Family Psychology*, *32*(1), 145–150. https://doi.org/10.1037/fam0000307

Lounds, J. J., Borkowski, J. G., & Whitman, T. L. (2004). Reliability and validity of the mother-child neglect scale. *Child Maltreatment*, *9*(4), 371–381. https://doi.org/10.1177/1077559504269536

Lucas, N. P., Macaskill, P., Irwig, L., & Bogduk, N. (2010). The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology*, *63*(8), 854–861. https://doi.org/10.1016/j.jclinepi.2009.10.002

Mammen, O., Kolko, D., & Pilkonis, P. (2003). Parental cognitions and satisfaction: relationships to aggressive parental behavior in child physical abuse. *Child Maltreatment*, *8*(4), 288–301. https://doi.org/10.1177/1077559503257112

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

Miller-Perrin, C. L., & Perrin, R. D. (2013). *Child maltreatment: An introduction* (3rd ed.). Sage.

Milner, J. S., & Crouch, J. L. (1997). Impact and detection of response distortions on parenting measures used to assess risk for child physical abuse. *Journal of Personality Assessment*, *69*(3), 633–650. https://doi.org/:10.1207/s15327752jpa6903_15

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*(7), e1000097. https://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*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs)-User manual (version 1.0). https://www.cosmin.nl/

wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

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*(4), 539–549. https://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*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

O'Dor, S. L., Grasso, D. J., Forbes, D., Bates, J. E., McCarthy, K. J., Wakschlag, L. S., & Briggs-Gowan, M. J. (2017). The Family Socialization Interview-Revised (FSI-R): A comprehensive assessment of parental disciplinary behaviors. *Prevention Science*, *18*(3), 292–304. https://doi.org/10.1007/s11121-016-0707-7

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

Rodriguez, C. M. (2010). Parent-child aggression: association with child abuse potential and parenting styles. *Violence and Victims*, *25*(6), 728–741. https://doi.org/10.1891/0886-6708.25.6.728

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*(3), 199–209. https://doi.org/10.1016/j.chiabu.2010.12.004

Russa, M. B., & Rodriguez, C. M. (2010). Physical discipline, escalation, and child abuse potential: Psychometric evidence for the analog parenting task. *Aggressive Behavior*, *36*(4), 251–260. https://doi.org/10.1002/ab.20345

Russell, B. S. (2010). Revisiting the measurement of shaken baby syndrome awareness. *Child Abuse & Neglect*, *34*(9), 671–676. https://doi.org/10.1016/j.chiabu.2010.02.008

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. https://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*(3), 236–240. https://doi.org/10.1016/j.injury.2010.11.042

Shanahan, M. E., Fliss, M. D., & Proescholdbell, S. K. (2018). Child Maltreatment Surveillance Improvement Opportunities: A Wake County, North Carolina Pilot Project. *North Carolina Medical Journal*, *79*(2), 88–93. https://doi.org/10.18043/ncm.79.2.88

Shen, F. (2017). Multitrait-multimethod matrix. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The international encyclopedia of communication research methods* (pp. 1–6). John Wiley

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*, 458678. https://doi.org/10.1155/2014/458678

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*(19), 3343–3366. https://doi.org/10.1177/0886260514563836

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*(4), 249–270. https://doi.org/10.1016/S0145-2134(97)00174-9

Sugaya, L., Hasin, D. S., Olfson, M., Lin, K. H., Grant, B. F., & Blanco, C. (2012). Child physical abuse and adult mental health: A national study. *Journal of Traumatic Stress*, *25*(4), 384–392. https://doi.org/10.1002/jts.21719

Tang, W., Hu, J., Zhang, H., Wu, P., & He, H. (2015). Kappa coefficient: A popular measure of rater agreement. *Shanghai Archives of Psychiatry*, *27*(1), 62–67. https://doi.org/10.11919/j.issn.1002-0829.215010

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., Lai, J.-S., Choi, S. W., Hays, R. D., Reeve, B. B., Reise, S. P., Pilkonis, P. A., & Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly*, *51*(2), 148–180.

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*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

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

Terwee, C. B., Prinsen, C. A. C., Ricci Garotti, M. G., Suman, A., de Vet, H. C. W., & Mokkink, L. B. (2016). The quality of systematic reviews of health-related outcome measurement instruments. *Quality of Life Research*, *25*(4), 767–779. https://doi.org/10.1007/s11136-015-1122-4

Thalheimer, W., & Cook, S. (2002). *How to calculate effect sizes from published research: A simplified methodology*. Work-Learning Research. http://www.work-learning.com

United Nations. (1989). Convention on the rights of the child. https://www.ohchr.org/EN/ProfessionalInterest/Pages/CRC.aspx

United Nations. (2015). *Transforming our world: The 2030 Agenda for sustainable development*. (A/RES/70/1). https://sustainabledevelopment.un.org.post2015/transformingourworld

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*(6), 2055–2064. https://doi.org/10.1542/peds.2005-2204

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, *3*(1), 25. https://doi.org/10.1186/1471-2288-3-25

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M., & QUADAS-2 Group. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

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

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A systematic review evaluating psychometric properties of parent or caregiver report instruments on child maltreatment: Part 1: Content validity. *Trauma, Violence, & Abuse*. Advanced online publication. https://doi.org/10.1177/1524838019898456

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 Department of Social Work, Education and Community Wellbeing, Northumbria University in the United Kingdom.

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.

Supplementary Appendices

Appendix A. *Database Search Strategies.*

Database	Search Terms (Subject heading and Free text words)	Number of records
CINAHL	(((MH "Child Abuse+") OR (MH "Domestic Violence+") OR (MH "Family Conflict") OR (MH "Aggression+") OR (MH "Punishment")) AND ((MH "Parents+") OR (MH "Parenting") OR (MH "Father-Infant Relations")OR (MH "Father-Child Relations") OR (MH "Fathers+") OR (MH "Mother-Child Relations") OR (MH "Mother-Infant Relations") OR (MH "Mothers+") OR (MH "Family+") OR (MH "Caregivers") OR (MH "Child Rearing+")) AND ((MH "Psychometrics") OR (MH "Measurement Issues and Assessments") OR (MH "Validity") OR (MH "Predictive Validity") OR (MH "Reliability and Validity") OR (MH "Internal Validity") OR (MH "Face Validity") OR (MH "External Validity") OR (MH "Discriminant Validity") OR (MH "Criterion-Related Validity") OR (MH "Consensual Validity") OR (MH "Concurrent Validity") OR (MH "Qualitative Validity") OR (MH "Construct Validity") OR (MH "Content Validity") OR (MH "Questionnaire Validation") OR (MH "Validation Studies") OR (MH "Test-Retest Reliability") OR (MH "Sensitivity and Specificity") OR (MH "Reproducibility of Results") OR (MH "Reliability") OR (MH "Intrarater Reliability") OR (MH "Interrater Reliability") OR (MH "Measurement Error") OR (MH "Bias (Research)") OR (MH "Selection Bias") OR (MH "Sampling Bias") OR (MH "Precision") OR (MH "Sample Size Determination") OR (MH "Repeated Measures") OR (Psychometric* or reliability or validit* or reproducibility or bias))) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) Limiters - Published Date:20181001-20191031)	1,173
Embase	((child abuse/ OR child neglect/ OR emotional abuse/ OR physical abuse/ OR battering/ OR domestic violence/ OR physical violence/ OR family conflict/ OR victim/ OR aggression/ OR punishment/) AND (parent/ OR father/ OR father child relation/ OR mother/ OR mother child relation/ OR family/ OR caregiver/ OR child rearing/) AND (psychometry/ or validity/ or reliability/ or measurement error/ or measurement precision/ or measurement repeatability/ or error/ or statistical bias/ or test retest reliability/ or intrarater reliability/ or interrater reliability/ or accuracy/ or criterion validity/ or internal validity/ or face validity/ or external validity/ or discriminant validity/ or concurrent validity/ or qualitative validity/ or construct validity/ or content validity/)) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) limit to yr="2019 -Current")	456
ERIC	((Child abuse/ OR Child neglect/ OR violence/ OR family violence/) AND (parenting styles/ OR parents/ OR child rearing/ OR father attitudes/ OR fathers/ OR mother attitudes/ OR mothers/ OR family attitudes/ OR caregiver attitudes/ OR caregiver child relationship/ OR caregiver role/ OR family environment/) AND (Psychometrics/ OR Validity/ OR Reliability/ OR Error of Measurement/ OR Bias/ OR Interrater Reliability/ OR Accuracy/ OR Predictive Validity/ OR Construct Validity/ OR Content Validity/)) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) limit to yr="Last year")	523

(continued)

Appendix A. *(continued)*

Database	Search Terms (Subject heading and Free text words)	Number of records
PsycINFO	((child abuse/ OR child neglect/ OR violence/ OR domestic violence/ OR physical abuse/ OR family conflict/ OR victimization/ OR aggressive behaviOR/ OR aggressiveness/ OR punishment/) AND (parent child communication/ OR parent child relations/ OR parenting/ OR parenting style/ OR parents/ OR father child communication/ OR father child relations/ OR fathers/ OR mother child communication/ OR mother child relations/ OR mothers/ OR family/ OR caregivers/) AND (Psychometrics/ OR Statistical Validity/ OR Test Validity/ OR Statistical Reliability/ OR Test Reliability/ OR Error of Measurement/ OR Errors/ OR Response Bias/ OR Interrater Reliability/ OR Repeated Measures/)) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) limit to yr="2019 -Current")	285
PubMed	(("Child Abuse"[Mesh] OR "Physical Abuse"[Mesh] OR "Domestic Violence"[Mesh] OR "Violence"[Mesh] OR "Family Conflict"[Mesh] OR "Aggression"[Mesh] OR "Punishment"[Mesh]) AND ("Parents"[Mesh] OR "Parent-Child Relations"[Mesh] OR "Parenting"[Mesh] OR "Fathers"[Mesh] OR "Father-Child Relations"[Mesh] OR "Mothers"[Mesh] OR "Mother-Child Relations"[Mesh] OR "Family"[Mesh] OR "Caregivers"[Mesh] OR "Child Rearing"[Mesh]) AND ("Psychometrics"[Mesh] OR "Reproducibility of Results"[Mesh] OR "Validation Studies as Topic"[Mesh] OR "Validation Studies" [Publication Type] OR "Bias"[Mesh] OR "Observer Variation"[Mesh] OR "Selection Bias"[Mesh] OR "Diagnostic Errors"[Mesh] OR "Dimensional Measurement Accuracy"[Mesh] OR "Predictive Value of Tests"[Mesh] OR "Discriminant Analysis"[Mesh])) OR (((child OR children OR infant* OR toddler* OR neonate* OR baby OR babies OR adolescent* OR teen* OR minor*) AND (victim* OR aggress* OR punish* OR abus* OR maltreat* OR neglect* OR mistreat* or violen* or conflict* or batter* or molest*) AND (rear* OR parent* OR father* OR mother* OR family OR families OR domestic* OR caregiver* OR carer* OR caring OR home OR homes) AND (psychometric* OR reliabilit* OR validit* OR reproducibilit* OR bias)) Filters: Publication date from 2018/10/05 to 2019/10/05)	1,092
Sociological Abstracts	(MAINSUBJECT.EXACT("Child Neglect") OR MAINSUBJECT.EXACT("Child Abuse") OR (MAINSUBJECT.EXACT("Violence") OR MAINSUBJECT.EXACT("Family Violence")) OR MAINSUBJECT.EXACT("Family Conflict") OR MAINSUBJECT.EXACT("Victimization") OR MAINSUBJECT.EXACT("Victims") OR MAINSUBJECT.EXACT("Aggression") OR (MAINSUBJECT.EXACT("Punishment") OR MAINSUBJECT.EXACT("Corporal Punishment")) OR MAINSUBJECT.EXACT("Emotional Abuse")) AND (MAINSUBJECT.EXACT("Parent Child Relations") OR MAINSUBJECT.EXACT("Parental Influence") OR MAINSUBJECT.EXACT("Parents") OR MAINSUBJECT.EXACT("Parental Attitudes") OR MAINSUBJECT.EXACT("Parenthood")) OR MAINSUBJECT.EXACT("Childrearing Practices") OR MAINSUBJECT.EXACT("Fathers") OR MAINSUBJECT.EXACT("Mothers") OR (MAINSUBJECT.EXACT("Family") OR MAINSUBJECT.EXACT("Family Relations") OR MAINSUBJECT.EXACT("Family Conflict") OR MAINSUBJECT.EXACT("Family Violence")) OR MAINSUBJECT.EXACT("Caregivers")) AND (MAINSUBJECT.EXACT("Psychometric Analysis") OR MAINSUBJECT.EXACT("Validity") OR MAINSUBJECT.EXACT("Reliability") OR MAINSUBJECT.EXACT("Error of Measurement") OR MAINSUBJECT.EXACT("Errors") OR MAINSUBJECT.EXACT("Test Bias") OR MAINSUBJECT.EXACT("Statistical Bias") OR MAINSUBJECT.EXACT("Bias") OR MAINSUBJECT.EXACT("Accuracy") OR MAINSUBJECT.EXACT("Agreement") OR MAINSUBJECT.EXACT("Research Design Error") OR MAINSUBJECT.EXACT("Specificity") OR MAINSUBJECT.EXACT("Sampling"))	133

Notes. All searches performed on the 29th of January 2018 with an update on the 5th of October 2019.

Appendix B. *Criteria for Good Psychometric Properties Adapted from Prinsen et al. (2018).*

Psychometric property	Rating[a]	Quality criteria
Structural validity	+	**CTT:** CFA: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08 (e.g., If at least one of CFI and TLI > 0.95) **IRT/Rasch:** CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08 AND residual correlations between the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 AND adequate looking graphs for monotonicity OR item scalability > 0.30 AND IRT x^2 > 0.01; Rasch: 0.5 ≤ infit and outfit mean squares ≤ 1.5 OR -2 < Z-standardised values < 2
	?	Not all information for '+' reported (e.g., **CTT:** If no psychometric data on any of CFI, TLI, RMSEA, or SRMR)
	-	Criteria for '+' not met (e.g., **CTT**: If both CFI and TLI ≤ 0.95)
	NR	No information found on structural validity
Hypotheses testing for construct validity	+	Correlations with instruments measuring similar constructs ≥ 0.50 OR meaningful differences between relevant (sub)groups (e.g., Cohen's d ≥ 0.50) OR at least 75% of the results are in accordance with the hypotheses
	?	Not all information for '+' reported (e.g., If only p-value and lack of information to calculate Cohen's d)
	-	Criteria for '+' not met (e.g., If Correlation r or Cohen's d < 0.50 or less than 75% of the results not in accordance with the hypotheses)
	NR	No information found on hypotheses testing for construct validity
Cross-cultural validity	+	No important differences found between group factors such as age, gender, and language in multiple group factor analysis OR DIF analysis: McFadden's R-Squared < 0.02
	?	Not all information for '+' reported (e.g., If no psychometric data on multiple group factor or DIF analysis)
	-	Criteria for '+' not met (e.g., If McFadden's R-Squared ≥ 0.02)
	NR	No information found on Cross-cultural validity\measurement invariance
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for '+' reported (e.g., If no psychometric data on AUC)
	-	Criteria for '+' not met (e.g., if AUC < 0.70)
	NR	No information found on criterion validity
Measurement error	+	SDC or LoA < MIC
	?	Not all information for '+' reported (e.g., If no psychometric data on MIC)
	-	Criteria for '+' not met (e.g., If LoA ≥ MIC)
	NR	No information found on measurement error
Internal consistency	+	At least low evidence[b] for sufficient structural validity AND Cronbach's alpha(s) ≥ 0.70
	?	Not all information for '+' reported OR Criteria for "At least low evidence[b] for sufficient structural validity not met (e.g., If no psychometric data on Cronbach's alpha or very low evidence for sufficient structural validity regardless of Cronbach alpha)
	-	Criteria for '+' not met (e.g., If low evidence for sufficient structural validity but Cronbach's alpha < 0.70)
	NR	No information found on internal consistency
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	Not all information for '+' reported (e.g., If no psychometric data on ICC)
	-	Criteria for '+' not met (e.g., If ICC < 0.70)
	NR	No information found on reliability

Note. AUC = Area Under the Curve; CFA = Confirmatory Factor Analysis; CFI = Comparative Fit Index; CTT = Classical Test Theory; DIF = Differential Item Functioning; ICC = Intraclass Correlation Coefficient; IRT = Item Response Theory; LoA = Limits of Agreement; MIC = Minimal Important Change; RMSEA: Root Mean Square Error of Approximation; SEM = Standard Error of Measurement; SDC = Smallest Detectable Change; SRMR: Standardised Root Mean Residuals; TLI = Tucker-Lewis Index.

[a] + = Sufficient; - = Insufficient; ? = Indeterminate; ± = Inconsistent; NR = Not Reported.

[b] As defined by grading the evidence according to the GRADE approach (Mokkink, Prinsen, et al., 2018).

Appendix C. *Modified GRADE Approach for Rating the Quality of Evidence on Measurement Properties Adapted from Prinsen et al. (2018).*

Level of evidence quality (sum of scores per factor)	Factor	Score	Criteria
High (0)	*Risk of bias*	0	Multiple studies of at least adequate methodological quality *OR* One study of very good methodological quality
Moderate (-1)		-1	Multiple studies of doubtful methodological quality *OR* Only one study of adequate methodological quality
Low (-2)		-2	Multiple studies of inadequate methodological quality *OR* Only one study of doubtful methodological quality
Very low (< -3)		-3	Only one study of inadequate methodological quality
	Inconsistency	0	All studies show the same results
		-1	Less than 75% of studies show either sufficient or insufficient results
		-2	50% of studies displayed sufficient results against the criteria *AND* Other 50% of studies displayed insufficient results against the criteria
	Imprecision	0	Total sample size > 100
		-1	Total sample size = 50–100
		-2	Total sample size = n < 50
	Indirectness	0	All studies addressing construct or target population of the review
		-1	At least one study not addressing construct or target population of the review, but not all
		-2	All studies not addressing construct or target population of the review

Note. The starting point of evidence quality is 'high' quality of evidence; the level of evidence quality is downgraded by the sum of scores per factors.

Appendix D. *Overview of Child Maltreatment Instrument: Reasons for Exclusion.*

No	Instrument (References)[a]	Abbreviation	Reason for exclusion
1	Adolescent Clinical Sexual Behavior Inventory (William N. Friedrich, Lysne, Sim, & Shamos, 2004)	ACSBI	Not a measure of child maltreatment
2	Adolescent Sexual Behavior Inventory–Self report (Wherry, Berres, Sim, & Friedrich, 2009)	ACSBI-S	Not a measure of child maltreatment
3	Adult Attachment Interviews (Hesse, 2008)	AAIs	Not a parent-report measure
4	Adult–Adolescent Parenting Inventory (Bavolek, 1984)	AAPI	Old version of a revised measure
5	Adverse Childhood Experiences questionnaire (Felitti et al., 1998)	ACEs	Not a parent-report measure
6	Alabama Parenting Questionnaire (Shelton, Frick, & Wootton, 1996)	APQ	Not a measure of child maltreatment
7	Assessing Environments III (Berger, Knutson, Mehm, & Perkins, 1988)	AEIII	Not a parent-report measure
8	Assessment of parental awareness of the shaken baby syndrome[b] (Mann, Rai, Sharif, & Vavasseur, 2015)	N/A	No psychometric data found
9	Body Image Victimization Experiences Scale (Duarte & Pinto-Gouveia, 2017)	BIVES	Not a measure of child maltreatment
10	Brief Child Abuse Potential Inventory (Ondersma, Chaffin, Mullins, & LeBreton, 2005)	BCAP	Not a measure of child maltreatment
11	Brigid Collins Risk Screener (Weberling, Forgays, Crain-Thoreson, & Hyman, 2003)	BCRS	Not a measure of child maltreatment
12	California Family Risk Assessment (W. L. Johnson, 2011)	CFRA	Not a parent-report measure
13	Caregiver–Child Social/Emotional and Relationship Rating Scale (McCall, Groark, & Fish, 2010)	CCSERRS	Not a measure of child maltreatment
14	CHild Abuse InveNtory at Emergency Rooms (Sittig et al., 2016)	CHAINER	Not a parent-report measure
15	Child Abuse Potential Inventory (Milner, 1986)	CAP	Not a measure of child maltreatment
16	Child Abuse Risk Assessment Scale (Chan, 2012)	CARAS	Not developed in English
17	Child and Adolescent Trauma Screen (Sachser et al., 2017)	CATS	Not a measure of child maltreatment
18	Child Behavior CheckList (Achenbach & Rescorla, 2000)	CBCL	Not a measure of child maltreatment
19	Child emotional maltreatment module[b] (A. M. Slep, Heyman, & Snarr, 2011)	N/A	No psychometric data found
20	Child maltreatment assessment (Salum et al., 2016)	N/A	Not developed in English
21	Child maltreatment measure[b] (Tajima, Herrenkohl, Huang, & Whitney, 2004)	N/A	No psychometric data found
22	Child Protective Services Review Document (Fanshel, Finch, & Grundy, 1994)	CPSRD	Not a parent-report measure
23	Child Reflective Functioning scale (Ensink et al., 2015)	CRF	Not a measure of child maltreatment
24	Child Sexual Behavior Inventory (W. N. Friedrich et al., 2001)	CSBI	Not a measure of child maltreatment
25	Child Well-Being Scales (Gaudin, Polansky, & Kilpatrick, 1992)	CWBS	Not a parent-report measure
26	Childhood Experience of Care and Abuse (Brown, Craig, Harris, Handley, & Harvey, 2007)	CECA	Not a parent-report measure
27	Childhood Experience of Care and Abuse Questionnaire (N. Smith, Lam, Bifulco, & Checkley, 2002)	CECA.Q	Not a parent-report measure
28	Childhood Experiences of Violence Questionnaire (Walsh, MacMillan, Trocme, Jamieson, & Boyle, 2008)	CEVQ	Not a parent-report measure
29	Childhood Trauma Interview (Fink, Bernstein, Handelsman, Foote, & Lovejoy, 1995)	CTI	Not a parent-report measure
30	Childhood Trauma Questionnaire (Bernstein, Ahluvalia, Pogge, & Handelsman, 1997)	CTQ	Not a parent-report measure
31	Childhood Trauma Questionnaire Short Form (Forde, Baron, Scher, & Stein, 2012)	CTQ-SF	Not a parent-report measure
32	Child–Parent Relationship Scale (Driscoll & Pianta, 2011)	CPRS	Not a measure of child maltreatment
33	Child–Parent Relationship Scale–Short Form (Pianta, 1992)	CPRS-SF	Not a measure of child maltreatment
34	Children Intimate Relationships, and Conflictual Life Events interview (Marshall, Feinberg, Jones, & Chote, 2017)	CIRCLE	Not a parent-report measure
35	Children's Impact of Traumatic Events Scale–Revised (Chaffin & Shultz, 2001)	CITES-R	Not a measure of child maltreatment
36	Christchurch trauma assessment (Nelson, Lynskey, Heath, & Martin, 2010)	N/A	Not a parent-report measure
37	Cleveland Child Abuse Potential Scale (Ezzo & Young, 2012)	C-CAPS	Not a parent-report measure
38	Comprehensive Childhood Maltreatment Inventory (Riddle & Aponte, 1999)	CCMI	Not a parent-report measure
39	Conflict Tactic Scale 2 (Straus et al., 2003)	CTS 2	Not a measure of child maltreatment

(Continued)

Appendix D. *(continued)*

No	Instrument (References)[a]	Abbreviation	Reason for exclusion
40	Conflict Tactics Scales (Straus et al., 2003)	CTS	Not a measure of child maltreatment
41	Defense Style Questionnaire (Bond & Wesley, 1996)	DSQ	Not a parent-report measure
42	Disciplinary methods interview[b] (Thompson, 2017)	N/A	Not a measure of child maltreatment
43	Discipline survey (Socolar, Savage, Devellis, & Evans, 2004)	N/A	Not a measure of child maltreatment
44	Dunedin Family Services Indicator (Muir et al., 1989)	DFSI	Not a parent-report measure
45	Dyadic Parent–child Interaction Coding System-II (Eyberg, Bessmer, Newcomb, Edwards, & Robinson, 1994)	DPICS-II	Not a parent-report measure
46	Egna Minnen Beträffande Uppfostran (My Memories of Upbringing) (Castro, de Pablo, Gomez, Arrindell, & Toro, 1997)	EMBU	Not developed in English
47	Egna Minnen Betrffånde Uppfostran for Children (Castro et al., 1997; Markus, Lindhout, Boer, Hoogendijk, & Arrindell, 2003)	EMBU-C	Not a parent-report measure
48	Emotional and Physical Abuse Questionnaire (Kemper, Carlin, & Buntain-Ricklefs, 1994)	EPAB	Not a parent-report measure
49	Environmental harshness, health, and life history strategy Indicators[b] (Chua, Lukaszewski, Grant, & Sng, 2017)	N/A	Not a measure of child maltreatment
50	Exposure To community Violence (Richters & Martinez, 1993)	ETV	Not a measure of child maltreatment
51	Exposure to violence questionnaire[b] (Kuo, Mohler, Raudenbush, & Earls, 2000)	N/A	Not a measure of child maltreatment
52	Familial Experiences Questionnaire (Wheelock, Lohr, & Silk, 1997)	FEQ	Not a parent-report measure
53	Family Affective Attitude Rating Scale (Waller, Gardner, Dishion, Shaw, & Wilson, 2012)	FAARS	Not a measure of child maltreatment
54	Family Aggression Screening Tool (Cecil, McCrory, Viding, Holden, & Barker, 2016)	FAST	Not a parent-report measure
55	Family Background Questionnaire–Brief (Melchert & Kalemeera, 2009)	FBQ-B	Not a parent-report measure
56	Family Behaviors Screen (Simmons, Craun, Farrar, & Ray, 2017)	FBS	Not a measure of child maltreatment
57	Family Betrayal Questionnaire (Delker, Smith, Rosenthal, Bernstein, & Freyd, 2017)	FBQ	Not a measure of child maltreatment
58	Family Law Detection Of Overall Risk Screen (McIntosh, Wells, & Lee, 2016)	FL-DOORS	Not a measure of child maltreatment
59	Family maltreatment diagnostic criteria (Heyman & Smith Slep, 2009)	N/A	Not a parent-report measure
60	Family Risk of Abuse And Neglect (Lennings, Brummert Lennings, Bussey, & Taylor, 2014)	FRAAN	Not a measure of child maltreatment
61	Family Therapy Alliance Scale (L. N. Johnson, Ketring, & Anderson, 2013)	FTAS	Not a measure of child maltreatment
62	Family Unpredictability Scale (Ross & Hill, 2000)	FUS	Not a measure of child maltreatment
63	Go/No–go Association Task Physical Discipline (Sturge-Apple, Rogge, Peltz, Suor, & Skibo, 2015)	GNAT-Physical Discipline	Not a measure of child maltreatment
64	Home Observation Measure of the Environment (Caldwell & Bradley, 2003)	HOME	Not a parent-report measure
65	Home safety screening (Scribano, Stevens, Marshall, Gleason, & Kelleher, 2011)	N/A	Not a measure of child maltreatment
66	Identification of Parents At Risk for child Abuse and Neglect (van der Put et al., 2017)	IPARAN	Not developed in English
67	Index of Child Care Environment (Anme et al., 2013)	ICCE	Not developed in English
68	Invalidating Childhood Environments Scale (Mountford, Corstorphine, Tomlinson, & Waller, 2007)	ICES	Not a measure of child maltreatment
69	Inventory on beliefs and attitudes towards domestic violence (Hutchinson & Doran, 2017)	N/A	Not a measure of child maltreatment
70	ISPCAN Child Abuse Screening Tool Children's version (Zolotor et al., 2009)	ICAST-C	Not a parent-report measure
71	ISPCAN Child Abuse Screening Tool Parents' version (Runyan et al., 2009)	ICAST-P	Developed in multiple languages
72	ISPCAN Child Abuse Screening Tools Retrospective version (Dunne et al., 2009)	ICAST-R	Not a parent-report measure
73	Japanese version of Conflict Tactics Scale[b] (Baba et al., 2017)	CTS1: Japanese version	Developed in English but translated and validated in other languages
74	Juvenile Victimization Questionnaire (Finkelhor, Hamby, Ormrod, & Turner, 2005)	JVQ	Not a parent-report measure

(Continued)

Appendix D. *(continued)*

No	Instrument (References)[a]	Abbreviation	Reason for exclusion
75	Maternal Characteristics Scale (Polansky, Gaudin, & Kilpatrick, 1992)	MCS	Not a measure of child maltreatment
76	Maternal discipline and appropriateness[b] (Padilla-Walker, 2008)	N/A	Not a parent-report measure
77	Maternal Responsiveness Questionnaire (Leerkes & Qu, 2017)	MRQ	Not a measure of child maltreatment
78	Maternal Self-report Support Questionnaire (D. W. Smith et al., 2010)	MSSQ	Not a measure of child maltreatment
79	Maternal Support Questionnaire–Child Report (D. W. Smith et al., 2017)	MSQ-CR	Not a measure of child maltreatment
80	Meaning of the Child interview (Grey & Farnfield, 2017)	MotC	Not a measure of child maltreatment
81	Measure Of Parenting Style (Parker et al., 1997)	MOPS	Not a parent-report measure
82	MeaSure trauma associated with Child Sexual Abuse (Choudhary, Satapathy, & Sagar, 2018)	MSCSA	Not a measure of child maltreatment
83	Measures of community–relevant outcomes for violence prevention programs[b] (Hausman et al., 2013)	N/A	Not a measure of child maltreatment
84	Medical history questionnaire[b] (Famularo, Fenton, & Kinscherff, 1992)	N/A	Not a measure of child maltreatment
85	Minnesota Multiphasic Personality Inventory-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kreammer, 1989)	MMPI-2	Not a measure of child maltreatment
86	Multidimensional Assessment of Parenting Scale (Parent & Forehand, 2017)	MAPS	Not a measure of child maltreatment
87	Multidimensional inventory for assessment of parental functioning (Reis, Orme, Barbera-Stein, & Herz, 1987)	N/A	Not a measure of child maltreatment
88	Multidimensional Neglectful Behavior Scale: Adolescent and adult recall version (Dubowitz et al., 2011)	MNBS-A	Not a parent-report measure
89	Multidimensional Neglectful Behavior Scale–Child Report (Beyazit & Ayhan, 2018)	MNBS-CR	Not a parent-report measure
90	National council on crime and delinquency indicators (Wood, 1997)	N/A	Not a parent-report measure
91	Needs-based Assessment of Parental (guardian) Support (Bolen, Lamb, & Gradante, 2002)	NAPS	Not a measure of child maltreatment
92	Neglect scale (Harrington, Zuravin, DePanfilis, Ting, & Dubowitz, 2002)	N/A	Not a parent-report measure
93	Parent cognition scale[b] (Snarr, Slep, & Grande, 2009)	N/A	Not a measure of child maltreatment
94	Parent discipline style[b] (Mezzich et al., 2007)	N/A	Not a measure of child maltreatment
95	Parent Perception Inventory (Glaser, Horne, & Myers, 1995)	PPI	Not a measure of child maltreatment
96	Parent Perception Inventory–Child version (Bruce et al., 2006)	PPIC	Not a measure of child maltreatment
97	Parent Problem Checklist (Stallman, Morawska, & Sanders, 2009)	PPC	Not a measure of child maltreatment
98	Parent Qualities Measure (Crick, 2006; Stallman et al., 2009)	PQM	Not a measure of child maltreatment
99	Parent Threat Inventory (Crick, 2006; Scher, Stein, Ingram, Malcarne, & McQuaid, 2002)	PTI	Not a parent-report measure
100	Parental Acceptance–Rejection Questionnaire (Rohner & Khaleque, 2005)	PARQ	Not a parent-report measure
101	Parental Anger Inventory (Scher et al., 2002; Sedlar & Hansen, 2001)	PAI	Not a measure of child maltreatment
102	Parental Authority Questionnaire (Buri, 1991)	PAQ	Not a measure of child maltreatment
103	Parental Emotion Regulation Inventory (Lorber, Del Vecchio, Feder, & Smith Slep, 2017; Sedlar & Hansen, 2001)	PERI	Not a measure of child maltreatment
104	Parental Empathy Measure (Kilpatrick, 2005; Lorber et al., 2017)	PEM	Not a measure of child maltreatment
105	Parent–Child Activities interview (Kilpatrick, 2005; Lefever et al., 2008)	PCA	Not a parent-report measure
106	Parent–Infant Relationship Global Assessment Scale (Lefever et al., 2008; THREE, 2005)	PIR-GAS	Not a measure of child maltreatment
107	Parenting Anxious Kids Ratings Scale–Parent Report (Flessner, Murphy, Brennan, & D'Auria, 2017; THREE, 2005)	PAKRS-PR	Not a measure of child maltreatment
108	Parenting behavior rating scales (Flessner et al., 2017; G. A. King, Rogers, Walters, & Oldershaw, 1994)	N/A	Not a parent-report measure
109	Parenting daily diary (G. A. King et al., 1994; Peterson, Tremblay, Ewigman, & Popkey, 2002)	N/A	Not a parent-report measure
110	Parenting Practices Questionnaire–Corporal Punishment (Avinun, Davidov, Mankuta, Knafo-Noam, & Knafo-Noam, 2018)	PPQ-CP	Not a measure of child maltreatment
111	Parenting Scale (Peterson et al., 2002; Salari, Terreros, & Sarkadi, 2012)	PS	Not a measure of child maltreatment

(Continued)

Appendix D. *(continued)*

No	Instrument (References)[a]	Abbreviation	Reason for exclusion
112	Parenting Support Needs Assessment (Murry & Lewin, 2014; Salari et al., 2012)	PSNA	Not a measure of child maltreatment
113	Plotkin Child Vignettes (Plotkin, 1983)	PCV	Not a measure of child maltreatment
114	Post–divorce Parental Conflict Scale (Morris & West, 2000; Murry & Lewin, 2014)	PPCS	Not a measure of child maltreatment
115	PREschool Symptom Self-report (Martini, Strayhorn, & Puig-Antich, 1990)	PRESS	Not a measure of child maltreatment
116	Production of Discipline Alternatives (Rodriguez, Wittig, & Christl, 2019)	PDA	Not a parent-report measure
117	Protective Factors Survey (Counts, Buffington, Chang-Rios, Rasmussen, & Preacher, 2010; Martini et al., 1990)	PFS	Not a measure of child maltreatment
118	Psychological Maltreatment Rating Scales (Brassard, Hart, & Hardy, 1993; Counts et al., 2010)	PMRS	Not a parent-report measure
119	Psychological neglect (Brassard et al., 1993; Christ, Kwak, & Lu, 2017)	N/A	Not a parent-report measure
120	Psychologically Violent Parental Practices Inventory (Christ et al., 2017; Gagne, Pouliot-Lapointe, & St-Louis, 2007)	PVPPI	Not developed in English
121	Questionnaire for evaluating maltreatment and neglect (Calheiros, Patrício, Graça, & Magalhães, 2018)	N/A	Not developed in English
122	Reflective Parenting Assessment (Ensink, Leroux, Normandin, Biberdzic, & Fonagy, 2017; Gagne et al., 2007)	RPA	Not a measure of child maltreatment
123	Responsiveness index (Ensink et al., 2017; Yates, Hull, & Huebner, 1983)	N/A	Not a parent-report measure
124	Revised Child Anxiety and Depression Scale Parent version (Ebesutani, Tottenham, & Chorpita, 2015; Yates et al., 1983)	RCADS-P	Not a measure of child maltreatment
125	Risk scale[b] (Ebesutani et al., 2015; Grietens, Geeraert, & Hellinckx, 2004)	N/A	Not a parent-report measure
126	Rorschach Inkblot Method (Choca, 2013; Grietens et al., 2004)	RIM	Not a measure of child maltreatment
127	Scale of Negative Family Interactions (Choca, 2013; Simonelli, Mullis, & Rohde, 2005)	SNFI	Not a parent-report measure
128	Screen for Adolescent Violence Exposure for children version (Flowers, Lanclos, & Kelley, 2002; Simonelli et al., 2005)	KID-SAVE	Not a parent-report measure
129	Sexual Abuse Indicators (Flowers et al., 2002; Terrell et al., 2008)	SAI	Not a parent-report measure
130	Sexual behavior problems questionnaire[b] (Hall, Mathews, & Pearce, 1998; Terrell et al., 2008)	N/A	Not a parent-report measure
131	Sexual Events Questionnaire (Finkelhor, 1979; Hall et al., 1998)	SEQ	Not a parent-report measure
132	Sexual Experiences Survey (Finkelhor, 1979; Koss & Gidycz, 1985)	SES	Not a parent-report measure
133	Shaken Baby Syndrome awareness assessment (Koss & Gidycz, 1985; Russell & Britner, 2006)	SBS	Old version of a revised measure
134	Sixteen Personality Factor questionnaire (Francis, Hughes, & Hitz, 1992; Russell & Britner, 2006)	16-PF	Not a measure of child maltreatment
135	Social Factors and Children Violence Questionnaire (Francis et al., 1992; Oni & Adetoro, 2014)	SPCVQ	No psychometric data found
136	Standardized Observation Codes III (Cerezo, Keesler, Dunn, & Wahler, 1986; Oni & Adetoro, 2014)	SOC III	Not a measure of child maltreatment
137	Structured Problem Analysis of Raising Kids (Cerezo et al., 1986; Staal, van den Brink, Hermanns, Schrijvers, & van Stel, 2011)	SPARK	Not a measure of child maltreatment
138	Supervisory neglect (Coohey, 2003; Staal et al., 2011)	N/A	Not a parent-report measure
139	Symptoms Of Trauma Scale (Coohey, 2003; Ford et al., 2017)	SOTS	Not a measure of child maltreatment
140	Trauma Experiences Checklist (Cristofaro et al., 2013; Ford et al., 2017)	TEC	Not a measure of child maltreatment
141	Trauma History Questionnaire (Cristofaro et al., 2013; Hooper, Stockton, Krupnick, & Green, 2011)	THQ	Not a parent-report measure
142	Trauma Symptom Checklist for Children (Briere et al., 2001; Hooper et al., 2011)	TSCC	Not a measure of child maltreatment
143	Trauma Symptom Checklist for Young Children (Briere et al., 2001)	TSCYC	Not a measure of child maltreatment
144	U.S. Air Force Family Advocacy Program Severity Index (Briere et al., 2001; A. M. Slep & Heyman, 2004)	USAF-FAP Severity Index	Not a parent-report measure

(Continued)

Appendix D. *(continued)*

No	Instrument (References)[a]	Abbreviation	Reason for exclusion
145	Violent Experiences Questionnaire–Revised (A. R. King & Russell, 2017; A. M. Slep & Heyman, 2004)	VEQ-R	Not a parent-report measure
146	Weekly Problems Scales (A. R. King & Russell, 2017; Sawyer, Tsao, Hansen, & Flood, 2006)	WPS	Not a measure of child maltreatment
147	When Bad Things Happen scale (Fletcher, 1995; Sawyer et al., 2006)	WBTH	Not a measure of child maltreatment
148	Young Parenting Inventory (Young, Klosko, & Weishaar, 2003)	YPI	Not a parent-report measure
149	Young Parenting Inventory–Revised (Louis, Wood, & Lockwood, 2018)	YPI-R2	Not a parent-report measure
150	Young Schema Questionnaire–Short form 3 (Young, 2005)	YSQ-S3	Not a parent-report measure

Notes. N/A = Not Applicable (No Abbreviation).

[a] References of the excluded instruments in this review are available from the first author upon request.

[b] Unofficial title retrieved from publication content as an instrument published without a title or abbreviation.

Appendix E. *Descriptions of Included Studies on Psychometric Properties of Instruments for the Assessment of Child Maltreatment.*

Instrument (Abbreviation)	Reference	Purpose of study	Assessed Psychometric properties	Study population	Age (Range [R] and/or Mean [MN] and/or Standard Deviation [SD])
Adult Adolescent Parenting Inventory-2 (AAPI-2)	Bavolek and Keene (1999)	To develop and validate the AAPI-2	Structural validity Internal consistency Hypotheses testing for construct validity	N = 1427 (Stage: Construct development): (I) Adolescents and adult parents N = 989 (Stage: Validation of the AAPI-2): (II) Non-Abusive parents (F = 677; M = 225); (III) Abusive parents (F = 58; M = 29)	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR; (III) R = NR, MN = NR, SD = NR
	Conners et al. (2006)	To examine the psychometric properties of the AAPI-2	Structural validity Internal consistency Hypotheses testing for construct validity	N = 309: Low-income parents of preschool age children (F = NR; M = NR)	R = 21–30y, MN = NR, SD = NR
	Lawson et al. (2017)	To examine the construct and predictive validity of the AAPI-2	Structural validity Internal consistency Hypotheses testing for construct validity	N = 2,610: Participating parents in child maltreatment prevention programs (F = 2,583; M = 27): (I) n = 1,271: Parents completing the AAPI-2 only before the programs (F = 1,258; M = 13); (II) n = 1,339: Parents completing the AAPI-2 both before and after the programs (F = 1,325; M = 14)	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR
	Rodriguez et al. (2011)	To develop and validate the P-CAAM (correlation with AAPI-2, CAP, and APT)	Internal consistency Hypotheses testing for construct validity	N = 147 (Stage: Pilot Testing): (I) Undergraduate students (F = 105; M = 42) N = 70 (Stage: Validation of P-CAAM by comparing with AAPI-2): (II) Mothers of children younger than 12 years old (F = 70; M = 0)	(I) R = NR, MN = 18.91y, SD = 2.35y; (II) R = NR, MN = 36.71y, SD = 6.59y
	Russa and Rodriguez (2010)	To support the validity of the APT as a questionnaire to assess risk for harsh, physically aggressive parenting (correlation with AAPI-2 and CAP)	Hypotheses testing for construct validity	N = 66 (Stage: Correlation study between APT and AAPI-2): (I) Pre-parent undergraduate students (F = 55; M = 11) N = 181 (Stage: Correlation study between APT, ATS and AAPI-2): (II) Pre-parent undergraduate students (F = 134; M = 47) N = 324 (Stage: Correlation study between APT, ATS and CAP): (III) Pre-parent undergraduate students (F = 220; M = 104)	(I) R = NR, MN = 18.76y, SD = 1.56y; (II) R = NR, MN = 18.91y, SD = 2.40y; (III) R = NR, MN = 19.13y, SD = 2.45y
Analog Parenting Task (APT)	Rodriguez et al. (2011)	To develop and validate the P-CAAM (correlation with AAPI-2, CAP, and APT)	Internal consistency Hypotheses testing for construct validity	N = 147 (Stage: Pilot Testing): (I) Undergraduate students (F = 105; M = 42) N = 70 (Stage: Validation of P-CAAM by comparing with AAPI-2): (II) Mothers of children younger than 12 years old (F = 70; M = 0)	(I) R = NR, MN = 18.91y, SD = 2.35y; (II): R = NR, MN = 36.71y, SD = 6.59y
	Russa and Rodriguez (2010)	To support the validity of the APT as an questionnaire to assess risk for harsh, physically aggressive parenting (correlation with AAPI-2, ATS, and CAP)	Internal consistency Hypotheses testing for construct validity	N = 66 (Stage: Correlation study between APT and AAPI-2): (I) Pre-parent undergraduate students (F = 55; M = 11) N = 181 (Stage: Correlation study between APT, ATS and AAPI-2): (II) Pre-parent undergraduate students (F = 134; M = 47) N = 324 (Stage: Correlation study between APT, ATS and CAP): (III) Pre-parent undergraduate students (F = 220; M = 104)	(I) R = NR, MN = 18.76y, SD = 1.56y; (II) R = NR, MN = 18.91y, SD = 2.40y; (III) R = NR, MN = 19.13y, SD = 2.45y

(Continued)

Appendix E. *(continued)*

Instrument (Abbreviation)	Reference	Purpose of study	Assessed Psychometric properties	Study population	Age (Range [R] and/or Mean [MN] and/or Standard Deviation [SD])
Child Neglect Questionnaire (CNQ)	Kirisci et al. (2001)	To develop and evaluate psychometric properties of the CNQ	Structural validity Internal consistency Hypotheses testing for construct validity	N = 172: (I) n = 76: Parents of children having fathers with Substance Use Disorder (SUD); (II) n = 96: Parents of children having fathers without SUD	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR
Child Neglect Scales–Maternal Monitoring and Supervision Scale (CNS-MMS)	Stewart et al. (2015)	To develop and evaluate validity and reliability of the Child Neglect Scales (CNS)	Structural validity Internal consistency Hypotheses testing for construct validity	N = 344: (I) n = 122: Mothers of boys having fathers with Substance Use Disorder (SUD) (F = 122; M = 0); (II) n = 222: Mothers of boys having fathers without SUD (F = 222; M = 0)	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR
Child Trauma Screen–Exposure Score (CTS-ES)	Lang and Connell (2017)	To develop and validate the Child Trauma Screen (CTS)	Hypotheses testing for construct validity	N = 923 (Stage: CTS Development): (I) Parents of children receiving care at outpatient behavioral health clinics N = 69 (Stage: CTS Validation): (II) Parents of children receiving care at outpatient behavioral health clinics	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR
Conflict Tactics Scales: Parent–Child version (CTSPC)	Compier-de Block et al. (2017)	To examine to what extent parents and children agree on the occurrence of various types of parent-to-child maltreatment by comparing parent and child report CTSPC	Internal consistency Reliability Hypotheses testing for construct validity	N = 83: Parents reported on perpetrated maltreatment (F = 48; M = 35)	R = 33–88y, MN = 57.4y, SD = 11.5y
	Cotter et al. (2018)	To examine the factor structure of the CTSPC as well as its convergent validity with the DPICS	Structural validity Internal consistency Hypotheses testing for construct validity	N = 110: Parents with a substantiated physical abuse incident who were court-mandated to receive treatment (F = 72; M = 38)	R = NR, MN = 32.24y, SD = 8.68y
	Grasso et al. (2016)	To examine the overlap between specific forms of psychological and physical Intimate partner violence by using the CTS-2 and mothers' harsh parenting behaviors by using the CTSPC, and their relationship to child disruptive behavior by using the MAP-DB	Internal consistency	N = 162: (I) n = 81: Mothers of children ages 4 to 6 years (F = 81; M = 0); (II) n = 81: Children ages 4 to 6 years (F = 31; M = 50)	(I) R = NR, MN = 31.1y, SD = 5.4y; (II) R = 4–6y, MN = 4.74y, SD = 0.91y
	Kobulsky et al. (2017)	To investigate the concordance of parent and child reports of current physical abuse by using the CTSPC, and the relation between concordance and parent and child reports of current child behavioral problems by using the CBCL and the YSR.	Reliability	N = 1,376: (I) n = 638: Parents reported on child physical abuse (F = 572; M = 66); (II) n = 638: Children reported on child physical abuse (F = 369; M = 269)	(I) R = 22–87y, MN = 39.5y, SD = 8.4y; (II) R = 11–17y, MN = 13.6y, SD = 1.9y

(Continued)

Appendix E. *(continued)*

Instrument (Abbreviation)	Reference	Purpose of study	Assessed Psychometric properties	Study population	Age (Range [R] and/or Mean [MN] and/or Standard Deviation [SD])
Conflict Tactics Scales: Parent–Child version (CTSPC)	Lorber and Slep (2017)	To prove the reliability of CTSPC by using Item Response Theory (IRT) analyses	Structural validity Internal consistency	N = 453: parents with 3- to 7-year old children (F = 235; M = 218)	R = NR, MN = NR, SD = NR
	O'Dor et al. (2017)	To exam the psychometric properties of the FSI-R (correlation with CTSPC)	Internal consistency Hypotheses testing for construct validity	N = 772: (I) n = 386: Mothers of 3–6 year-old children with disruptive behavior or experience of intimate partner violence (IPV); (II) n = 386: 3–6 year-old children with disruptive behavior or parents exposed intimate partner violence (IPV)	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = 56.72m, SD = 10.27m
	Rodriguez (2010)	To explore relationships between parent–child aggression by using the CTSPC and parenting styles by using the PS associated with child maltreatment potential by using the CAP	Hypotheses testing for construct validity	N = 772: (I) n = 327: Parents of children younger than 12 (F = 275; M = 52); (II) n = 115: parents of children between ages 7 and 12 (F = 86; M = 29); (III) n = 74: Mothers of 7- to 12-year-old children with diagnosed externalising behavior problems (F = 74; M = 0)	(I) R = NR, MN = 30.48y, SD = 6.22y; (II) R = NR, MN = 37.62y, SD = 7.91y; (III) R = NR, MN = 40.65y, SD = 10.53y
	Straus et al. (1998)	To develop and test the reliability and validity of CTSPC	Internal consistency Hypotheses testing for construct validity	N = 1,000: Parents of children under 18 years old participated in an U.S. national survey (F = 660; M = 340)	R = NR, MN = 36.8y, SD = NR
Family Maltreatment–Child Abuse criteria (FM-CA)	Heyman, et al. (2019).	To develop and validate the FM-CA	Hypotheses testing for construct validity	N = 126: U.S. Air Force service members and their spouses (F = 41; M = 85)	R = NR, MN = NR, SD = NR
ISPCAN Child Abuse Screening Tool for use in Trials (ICAST-Trial)	Meinck et al. (2018)	To develop and validate the ICAST-Trial	Structural validity Internal consistency Hypotheses testing for construct validity	N = 115 (Stage: Pilot study) (I) Parents of adolescents participated in a parenting program to prevent child abuse (F = 112; M = 3) N = 552 (Stage: Validation of ICAST-Trial) (II) Parents of adolescents participated in a parenting program to prevent child abuse (F = 523; M = 29)	(I) R = NR, MN = 48y, SD = 13.6y; (II) R = NR, MN = 49.4y, SD = 14.69y
Intensity of Parental Punishment Scale (IPPS)	Gordon et al. (1979)	To develop and validate the IPPS	Structural validity Internal consistency Cross-cultural validity Reliability Hypotheses testing for construct validity	N = 417: (I) n = 301: Parents of 5- to 10-year-old children; (II) n = 50: Upper-middle-class parents of 7- to 12-year old children; (III) n = 26: Mothers of 6- to 9-year-old children; (IV) n = 40: Mothers of 6- to 14-year-old children	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR; (III) R = NR, MN = NR, SD = NR; (IV) R = NR, MN = NR, SD = NR
Mother–Child Neglect Scale (MCNS)	Lounds et al. (2004)	To evaluate reliability and validity of the MCNS and MCNS-SF	Internal consistency Reliability Hypotheses testing for construct validity	N = 100: Adolescent mothers of children ages 9 to 15 years	R = 14.2–19.2y, MN = 17y, SD = 1.16y

(Continued)

Appendix E. *(continued)*

Instrument (Abbreviation)	Reference	Purpose of study	Assessed Psychometric properties	Study population	Age (Range [R] and/or Mean [MN] and/or Standard Deviation [SD])
Mother–Child Neglect Scale–Short Form (MCNS-SF)	Lounds et al. (2004)	To evaluate reliability and validity of the MCNS and MCNS-SF	Internal consistency Criterion validity Hypotheses testing for construct validity	N = 100: Adolescent mothers of children ages 9 to 15 years	R = 14.2–19.2y, MN = 17y, SD = 1.16y
Parent–Child Aggression Acceptability Movie task (P-CAAM)	Rodriguez et al. (2011)	To develop and validate the P-CAAM (correlation with AAPI-2, CAP, and APT)	Internal consistency Hypotheses testing for construct validity	N = 147 (Stage: Pilot Testing): (I) Undergraduate students (F = 105; M = 42) N = 70 (Stage: Validation of P-CAAM by comparing with AAPI-2): (II) Mothers of children younger than 12 years old (F = 70; M = 0)	(I) R = NR, MN = 18.91y, SD = 2.35y; (II): R = NR, MN = 36.71y, SD = 6.59y
Parent Opinion Questionnaire (POQ)	Azar and Rohrbeck (1986)	To assess validation of the POQ by comparing the unrealistic expectations of child abusing mothers with mothers whose partners perpetrated the abuse	Reliability Hypotheses testing for construct validity	N = 30: (I) n = 16 Mothers abusing their children; (II) n = 14: Non-abusing mothers with partners abusing their children	(I) R = NR, MN = NR, SD = NR; (II) R = NR, MN = NR, SD = NR
	Haskett et al. (2006)	To exam psychometric properties of the POQ and CV	Structural validity Internal consistency Hypotheses testing for construct validity	N = 155: (I) n = 77: Abusive parents documented history of child physical abuse with 4- to 10-year-old children (F = 64; M = 13); (II) n = 78: Non-abusive parents with 4- to 10-year-old children (F = 64; M = 14)	(I) R = NR, MN = 34.3y, SD = 7.2y; (II) R = NR, MN = 34.7y, SD = 9.3y
	Mammen et al. (2003)	To exam convergence among cognitions by using the POQ and satisfaction with the child by using the CRI in child abusive parents, and their relationships to parental aggression by using the CTS	Hypotheses testing for construct validity	N = 52: Parent participants in a treatment study because of physical abuse towards their children ages 6 to 13 years (F = 44; M = 8)	R = NR, MN = 31.9y, SD = 5.88y
Parental Response to Child Misbehavior questionnaire (PRCM)	Vittrup et al. (2006)	To exam the emergence of discipline techniques by mothers of young children by using the PRCM and assess the predictive validity of spanking attitudes with subsequent reports of spanking by using the ATS	Hypotheses testing for construct validity	N = 132: Mothers of 12- to 48-month-old children (F = 132; M = 0)	R = 20–44y, MN = 31.4y, SD = 4.5y
Shaken Baby Syndrome awareness assessment–Short Version (SBS-SV)	Russell (2010)	To develop and assess the psychometric properties of the SBS-SV	Internal consistency	N = 370: Public college students (F = 270; M = 100)	R = NR, MN = 21y, SD = NR

Note. AEIII = Assessing Environments III; ATS = Attitude Towards Spanking; CAP = Child Abuse Potential inventory; CBCL = Child Behavior CheckList; CRI = Child Rearing Inventory; CTS = Conflict Tactics Scale; CTS-2 = Conflict Tactics Scale-2; CV = Child Vignettes; DPICS = Dyadic Parent–child Interaction Coding System; FSI-R = Family Socialization Interview–Revised; ISPCAN = International Society for the Prevention of Child Abuse and Neglect; MAP-DB = Multidimensional Assessment of Preschool Disruptive Behavior; PS = Parenting Scale; YSR = Youth Self-Report.

Appendix F. *Study Results and Ratings on Psychometric Properties.*

Psychometric Property	Instrument	Reference	Risk of bias	Sample size	Results (rating)	Overall rating	Quality of evidence (reasons)
Structural validity	**AAPI-2**	Bavolek and Keene (1999)	Very good	1,427	*EFA*: 5-factor structure **(?)**	±	**Moderate** (partly inconsistent results)
		Conners et al. (2006)	Very good	309	*CFA*: 5-factor structure: 1st factor CFI = 0.86; 2nd factor = 0.68; 3rd factor = 0.82; 4th factor = 0.75; 5th factor = 1.0 **(-)**		
		Lawson et al. (2017)	Adequate	1,271	*CFA*: 5-factor structure: CFI = 0.84; TLI = 0.84; RMSEA = 0.14 **(-)**		
			Adequate	1,339	*CFA*: 2-factor structure: CFI = 0.901; TLI = 0.968; RMSEA = 0.058 **(+)**		
	CNQ	Stewart et al. (2015)	Adequate	172	*CFA*: 1-factor structure: Total scale and each subscale RMSEA < 0.001 **(+)**	+	**Moderate** (only one adequate study)
	CNS-MMS	Kirisci et al. (2001)	Very good	344	*CFA*: 1-factor structure: RMSEA = 0.028 **(+)**	+	**High** (no concern)
	CTSPC	Cotter et al. (2018)	Very good	110	*EFA:* 4-factor structure **(?)**	?	**Not evaluated** (lack of evidence)
		Lorber and Slep (2017)	Very good	453	*CFA:* Physical assault subscale CFI > 0.95; Total scale and other subscales: NR **(?)**		
			Very good	453	IRT: Physical assault subscale CFI > 0.95; Evidence of Local Independence: NR; Adequate looking graphs for Monotonicity: NR; $\chi^2 > 0.01$; Total scale and other subscales: NR **(?)**		
	ICAST-Trial	Meinck et al. (2018)	Very good	552	CFA: 4-factor structure: CFI 0.975, TLI 0.965, RMSEA 0.025, SRMR 0.036 **(+)**	+	**High** (no concern)
	IPPS	Gordon et al. (1979)	Adequate	217	*EFA*: 5-factor structure **(?)**	?	**Not evaluated** (lack of evidence)
	POQ	Haskett et al. (2006)	Doubtful	128	*EFA*: 6-factor structure **(?)**	?	**Not evaluated** (lack of evidence)
Internal consistency	**AAPI-2**	Bavolek and Keene (1999)	Very good	1,427	Conflicting structural validity *AND* Cronbach's Alpha = 0.82 (1st factor); 0.88 (2nd factor); 0.92 (3rd factor); 0.82 (4th factor); 0.80 (5th factor) **(?)**	?	**Not evaluated** (lack of evidence)
		Conners et al. (2006)	Very good	309	Conflicting structural validity *AND* Cronbach's alpha = 0.85 (Total scale); 0.79 (1st factor); 0.64 (2nd factor); 0.79 (3rd factor); 0.59 (4th factor); 0.50 (5th factor) **(?)**		
		Lawson et al. (2017)	Very good	1,271	Conflicting structural validity *AND* Cronbach's Alpha = 0 .89 (Total scale); 0.70 (1st factor); 0.69 (2nd factor); 0.70 (3rd factor); 0.56 (4th factor); 0.48 (5th factor) **(?)**		
		Rodriguez et al. (2011)	Adequate	147	Conflicting structural validity *AND* Cronbach's Alpha = 0.87 **(?)**		
			Very good	70	Conflicting structural validity *AND* Cronbach's Alpha = 0.83 **(?)**		
	APT	Rodriguez et al. (2011)	Very good	147	No evidence on the structural validity AND Cronbach's Alpha = 0.92 (Physical Discipline); 0.72 (Escalation) **(?)**	?	**Not evaluated** (lack of evidence)
		Russa and Rodriguez (2010)	Very good	66	No evidence on the structural validity AND Cronbach's Alpha = 0.93 (Physical Discipline); 0.80 (Escalation) **(?)**		
			Very good	181	No evidence on the structural validity AND Cronbach's Alpha = 0.91 (Physical Discipline); 0.77 (Escalation) **(?)**		
			Very good	324	No evidence on the structural validity AND Cronbach's Alpha = 0.92 (Physical Discipline); 0.72 (Escalation) **(?)**		

(Continued)

Appendix F. *(Continued).*

Psychometric Property	Instrument	Reference	Risk of bias	Sample size	Results (rating)	Overall rating	Quality of evidence (reasons)
Internal consistency	**CNQ**	Stewart et al. (2015)	Doubtful	172	Moderate evidence for sufficient structural validity *AND* Cronbach's Alpha = 0.86 (mother report); 0.92 (father report) **(+)**	**+**	**Low** (only one doubtful study)
	CNS-MMS	Kirisci et al. (2001)	Very good	344	High evidence for sufficient structural validity *AND* Cronbach's Alpha = 0.72 **(+)**	**+**	**High** (no concern)
	CTSPC	Compier-de Block et al. (2017)	Very good	35	Indeterminate structural validity *AND* Cronbach's Alpha (Father report) = 0.74 (Total scale); 0.64 (Psychological Aggression); 0.71 (Physical Assault); 0.78 (Emotional Neglect) **(?)**	**?**	**Not evaluated** (lack of evidence)
			Very good	48	Indeterminate structural validity *AND* Cronbach's Alpha (Mother report) = 0.82 (Total scale); 0.75 (Psychological Aggression); 0.68 (Physical Assault); 0.79 (Emotional Neglect) **(?)**		
		Cotter et al. (2018)	Adequae	110	Indeterminate structural validity *AND* Cronbach Alpha = 0.72 (Corporal Punishment/Minor Physical Assault); 0.75 (Psychological Aggression scale); 0.72 (Nonviolent Discipline scale); 0.46 (Severe Physical Assault) **(?)**		
		Grasso et al. (2016)	Very good	81	Indeterminate structural validity *AND* Cronbach's Alpha = 0.61 (Psychological Aggression); 0.59 (Mild/Moderate Child Physical Assault); 0.54 (Severe Child Physical Assault) **(?)**		
		Lorber and Slep (2017)	Adequate	453	Indeterminate structural validity *AND* Cronbach's Alpha ≤ 0.59 **(?)**		
		O'Dor et al. (2017)	Very good	386	Indeterminate structural validity *AND* Cronbach's Alpha = 0.65 (Physical Aggression); 0.68 (Psychological Aggression) **(?)**		
		Straus et al. (1998)	Adequate	1,000	Indeterminate structural validity *AND* Cronbach's Alpha = 0.55 (Physical Assault); 0.60 (Psychological Aggression); 0.70 (Nonviolent Discipline); 0.22 (Neglect) **(?)**		
	ICAST-Trial	Meinck et al. (2018)	Very good	552	High-quality evidence for sufficient structural validity *AND* Cronbach's Alpha = 0.84 (1st factor); 0.62 (2nd factor); 0.62 (3rd factor); 0.58 (4th factor) **(-)**	**-**	**High** (no concern)
	IPPS	Gordon et al. (1979)	Very good	217	Indeterminate structural validity *AND* Spilt-half reliability coefficient (r = 0.75, p < 0.01) **(?)**	**?**	**Not evaluated** (lack of evidence)
	MCNS	Lounds et al. (2004)	Very good	100	No evidence on the structural validity *AND* Coefficient Alpha = 0.94 (Total Scale); 0.80 (Emotional Needs); 0.86 (Cognitive Needs); 0.78 (Supervision); 0.90 (Physical Neglect) **(?)**	**?**	**Not evaluated** (lack of evidence)
			Very good	100	No evidence on the structural validity *AND* Coefficient Alpha = 0.95 (Total Scale); 0.85 (Emotional Needs); 0.86 (Cognitive Needs); 0.86 (Supervision); 0.91 (Physical Neglect) **(?)**		
	MCNS-SF	Lounds et al. (2004)	Very good	100	No evidence on the structural validity *AND* Coefficient Alpha = 0.90 **(?)**	**?**	**Not evaluated** (lack of evidence)
	P-CAAM	Rodriguez et al. (2011)	Adequate	147	No evidence on the structural validity *AND* Alpha = 0.77 (university student sample) **(?)**	**?**	**Not evaluated** (lack of evidence)
			Very good	70	No evidence on the structural validity *AND* Alpha = 0.74 (parent sample) **(?)**		
	POQ	Haskett et al. (2006)	Very good	128	Indeterminate structural validity *AND* KR-20 = 0.82 (Total scale); 0.54 (Self Care); 0.45 (Family Responsibility); 0.56 (Help/Affection to Parents); 0.31 (Leave Child Alone); 0.53 (Proper Behavior Feelings); 0.47 (Punishment) **(?)**	**?**	**Not evaluated** (lack of evidence)
	SBS-SV	Russell (2010)	Very good	370	No evidence on the structural validity *AND* Cronbach's Alpha = 0.76 (Soothing Technique); 0.79 (Discipline Techniques); 0.70 (Potential for Injury) **(?)**	**?**	**Not evaluated** (lack of evidence)

(Continued)

Appendix F. *(Continued).*

Psychometric Property	Instrument	Reference	Risk of bias	Sample size	Results (rating)	Overall rating	Quality of evidence (reasons)
Cross-cultural validity	**IPPS**	Gordon et al. (1979)	Inadequate	217	Multiple group factor analysis or DIF: NR **(?)**	**?**	**Not evaluated** (lack of evidence)
Reliability	**CTSPC**	Compier-de Block et al. (2017)	Very good	35	Interrater reliability: ICC = 0.29 **(-)**	**-**	**Moderate** (some indirect evidence from different population other than target population)
			Very good	48	Interrater reliability: ICC = 0.18 **(-)**		
		Kobulsky et al. (2017)	Very good	638	Interrater reliability: Kappa = 0.144 **(?)**		
	IPPS	Gordon et al. (1979)	Doubtful	19	Test-retest reliability: r = 0.85; statistical method not reported **(?)**	**?**	**Not evaluated** (lack of evidence)
			Doubtful	50	Test-retest reliability: r = 0.56; statistical method not reported **(?)**		
	MCNS	Lounds et al. (2004)	Adequate	100	Test-retest reliability: Spearman's rho = 0.60 **(?)**	**?**	**Not evaluated** (lack of evidence)
	POQ	Azar and Rohrbeck (1986)	Doubtful	16	Test-retest reliability: r = 0.85; statistical method not reported **(?)**	**?**	**Not evaluated** (lack of evidence)
Criterion validity	**MCNS-SF**	Lounds et al. (2004)	Very good	100	Correlation with MCNS (long version): r = 0.96 **(+)**	**+**	**High** (no concern)
Hypotheses testing for construct validity	**AAPI-2**	Bavolek and Keene (1999)	Adequate	989	Difference between abusive and non-abusive parents: Cohen’s d = 0.57–3.96 **(+)**	**-**	**Moderate** (some indirect evidence from different population other than target population)
			Adequate	989	Difference between fathers and mothers: Cohen’s d = 0.28–1.40 **(-)**		
		Conners et al. (2006)	Adequate	309	Correlation with PDMI: r = -0.36 **(-)**		
			Very good	309	Correlation with HOME r = 0.19 **(-)**		
			Very good	309	Correlation with Parenting Style: r = -0.45 **(-)**		
			Very good	309	Correlation with PKBS: r = -0.23 **(-)**		
		Lawson et al. (2017)	Adequate	1,339	Difference between parents of children with a substantiated child maltreatment report (SCAR) and without SCAR (2-factor AAPI-2): Cohen’s d = 0.04 **(-)**		
			Adequate	1,339	Difference between parents of children with a substantiated child maltreatment report (SCAR) and without SCAR (5-factor AAPI-2): Cohen’s d = 0.03–0.11 **(-)**		
		Rodriguez et al. (2011)	Very good	147	Correlation with P-CAAM: r = -0.33 **(-)**		
			Very good	70	Correlation with P-CAAM: r = -0.51 **(+)**		
		Russa and Rodriguez (2010)	Very good	66	Correlation with APT: r = 0.353 **(-)**		
			Very good	181	Correlation with APT: r = 0.497 **(-)**		
	APT	Rodriguez et al. (2011)	Very good	147	Correlation with P-CAAM: r = 0.26–0.29 **(-)**	**±**	**Very Low** (partly inconsistent results, all indirect evidence from different population other than target population)
		Russa and Rodriguez (2010)	Very good	66	Correlation with AAPI-2: r = 0.339–0.353 **(-)**		
			Very good	181	Correlation with AAPI-2: r = 0.463–0.497 **(-)**		
			Very good	324	Correlation with CAP: r = 0.158–0.279 **(-)**		
			Adequate	324	Correlation with ATS: r = 0.521–0.565 **(+)**		
			Adequate	181	Correlation with ATS: r = 0.577–0.612 **(+)**		

(Continued)

Appendix F. *(Continued).*

Psychometric Property	Instrument	Reference	Risk of bias	Sample size	Results (rating)	Overall rating	Quality of evidence (reasons)
Hypotheses testing for construct validity	**CNQ**	Stewart et al. (2015)	Very good	172	Correlation with CRPB: r = 0.01–0.19 **(-)**	-	**High** (no concern)
			Very good	172	Correlation with FAM: r = 0.00–0.13 **(-)**		
			Very good	172	Correlation with ACQ: r = 0.02–0.11 **(-)**		
			Very good	172	Correlation with CRC: r = -0.13–0.27 **(-)**		
			Very good	172	Difference between families of fathers with and without substance use disorder: Lack of information to calculate Cohen’s d **(?)**		
	CNS-MMS	Kirisci et al. (2001)	Very good	344	Correlation with child-report CNS: r = -0.10 (boys of fathers with substance use disorder (SUD); r = -0.18 (without SUD) **(-)**	-	**Moderate** (some indirect evidence from different population other than target population)
	CTS-ES	Lang and Connell (2017)	Very good	69	Correlation with CPSS: r = 0.49 **(-)**	±	**Low** (totally inconsistent results)
			Very good	69	Correlation with CPSS: r = 0.71 **(+)**		
	CTSPC	Compier-de Block et al. (2017)	Adequate	83	Difference between parents with younger children and parents with older children: Cohen’s d = 0.54 **(+)**	-	**High** (no concern)
		Cotter et al. (2018)	Very good	110	Correlation with DPICS = -0.21–0.26 **(-)**		
		O'Dor et al. (2017)	Very good	386	Correlation with FSI-R: r = 0.31–0.63 **(-)**		
		Rodriguez (2010)	Very good	327	Correlation with CAPI: r = -0.01–0.39 **(-)**		
			Very good	327	Correlation with PS: r = -0.08–0.56 **(-)**		
			Very good	115	Correlation with CAPI: r = 0.08–0.33 **(-)**		
			Very good	115	Correlation with PS: r = -0.03–0.56 **(-)**		
			Very good	74	Correlation with CAPI: r = -0.14–0.33 **(-)**		
			Very good	74	Correlation with PS: r = -0.27–0.48 **(-)**		
		Straus et al. (1998)	Adequate	1,000	Difference between younger and older parents: Cohen’s d = -0.70–-0.24 **(-)**		
			Adequate	1,000	Difference between parents with younger and older children: Cohen’s d = -0.72—0.12 **(-)**		
			Adequate	182	Difference between European American and African Hispanic American parents: Cohen’s d = 0.68 **(+)**		
			Adequate	1,000	Difference between mothers and fathers: Cohen’s d = 0.1 **(-)**		
	FM-CA	Heyman, et al. (2019).	Doubtful	126	Correlation with CTSPC: Guilford G (inter-rater agreement coefficient) = -0.06–0.94 **(?)**	?	**Not evaluated** (lack of evidence)
	ICAST-Trial	Meinck et al. (2018)	Very good	552	Correlation with Corporal Punishment items of APQ: r = 0.457 **(-)**	-	**High** (no concern)

(Continued)

Appendix F. *(Continued).*

Psychometric Property	Instrument	Reference	Risk of bias	Sample size	Results (rating)	Overall rating	Quality of evidence (reasons)
Hypotheses testing for construct validity	**IPPS**	Gordon et al. (1979)	Adequate	42	Correlation with Intensity of Anger: r = 0.84 **(+)**	±	**Low** (partly inconsistent results, multiple doubtful studies)
			Adequate	26	Correlation with Parent–Child Interaction Code: r = -0.57–0.44 **(-)**		
			Adequate	50	Correlation with Parent's Priorities of Child Behaviors: r = 0.49 **(-)**		
			Adequate	40	Difference between mothers with less and more warmth to children: Lack of information to calculate Cohen's d **(?)**		
			Adequate	64	Difference between parents of children with and without behavior problems: Lack of information to calculate Cohen's d **(?)**		
			Doubtful	192	Correlation with Frustration Tolerance): r = 0.18 **(-)**		
			Doubtful	49	Correlation with children's personality questionnaire: -0.28–0.32 **(-)**		
			Doubtful	43	Correlation with Child Behavior Rating: r = -0.57–0.58 **(+)**		
			Doubtful	205	Difference between parents with older and younger children: Cohen's d = 0.22–0.72 **(-)**		
			Doubtful	40	Difference between more and less out of contact mothers: Lack of information to calculate Cohen's d **(?)**		
			Doubtful	40	Difference between mothers giving more and less critical evaluations: Lack of information to calculate Cohen's d **(?)**		
			Doubtful	217	Difference between parents with lower and higher socioeconomic status: Cohen's d = 0.49 **(-)**		
			Doubtful	217	Difference between parents with less and more education: Cohen's d = 0.68 **(+)**		
	MCNS	Lounds et al. (2004)	Very good	100	Correlation with MIS: r = -0.31 **(-)**	-	**High** (no concern)
			Very good	100	Correlation with CAP: r = 0.16 **(-)**		
			Very good	100	Correlation with NS: r = 0.32–0.36 **(-)**		
	MCNS-SF	Lounds et al. (2004)	Very good	100	Correlation with CAP: r = 0.19 **(-)**	-	**High** (no concern)
			Very good	100	Correlation with NS: r = 0.28 **(-)**		
			Very good	100	Correlation with MIS: r = -0.26 **(-)**		
	P-CAAM	Rodriguez et al. (2011)	Very good	147	Correlation with ATS: r = 0.33–0.43 **(-)**	±	**Low** (partly inconsistent results, some indirect evidence from different population other than target population)
			Very good	147	Correlation with AAPI-2: r = -0.33—0.27 **(-)**		
			Very good	147	Correlation with APT: r = -0.21–0.30 **(-)**		
			Very good	70	Correlation with ATS: r = 0.26–0.30 **(-)**		
			Very good	70	Correlation with AAPI-2: r = -0.51—0.46 **(-)**		
			Very good	70	Correlation with CAPI: r = 0.27–0.30 **(-)**		
			Very good	70	Correlation with Parenting Scale: r = -0.28–0.34 **(-)**		
			Very good	34	Difference between parents with higher and lower scores of CAPI: Cohen's d = 0.70–0.76 **(+)**		
			Very good	34	Difference between parents with higher and lower scores of AAPI-2: Cohen's d = 1.32–1.44 **(+)**		
			Very good	34	Difference between parents with higher and lower scores of Overreactivity: Cohen's d = 0.57–0.89 **(+)**		
			Adequate	74	Difference between parents with higher and lower scores of AAPI-2: Cohen's d = 0.92–1.11 **(+)**		

(Continued)

Appendix F. *(Continued).*

Psychometric Property	Instrument	Reference	Risk of bias	Sample size	Results (rating)	Overall rating	Quality of evidence (reasons)
Hypotheses testing for construct validity	**POQ**	Azar and Rohrbeck (1986)	Very good	30	Difference between mothers who perpetrated child maltreatment and whose partners perpetrated child maltreatment: Cohen's d = 2.01 **(+)**	-	**High** (no concern)
		Haskett et al. (2006)	Very good	128	Correlation with CV: r = 0.33–0.43 **(-)**		
			Very good	128	Correlation with PSI: r = 0.31 **(-)**		
			Very good	128	Correlation with Parent–Child Interactions: r = -0.11–0.08 **(-)**		
			Very good	128	Correlation with K-BIT: r = -0.35 **(-)**		
			Very good	128	Difference between abusive and non-abusive parents: Cohen's d = 0.24 **(-)**		
			Adequate	128	Correlation with ECBI: r = 0.06–0.09 **(-)**		
			Adequate	128	Correlation with SCL-90-R: r = 0.30 **(-)**		
			Adequate	128	Correlation with CTS: r = 0.28 **(-)**		
		Mammen et al. (2003)	Very good	48	Correlation with CRI: r = 0.02 **(-)**		
			Very good	43	Correlation with CRI: r = 0.05 **(-)**		
			Very good	39	Correlation with PAT: r = 0.04 **(-)**		
			Very good	42	Correlation with PAT: r = -0.30 **(-)**		
			Very good	39	Correlation with PPQ: r = 0.08 **(-)**		
			Very good	40	Correlation with PPQ: r = 0.09 **(-)**		
			Very good	46	Correlation with CTSPC: r = -0.23–-0.02 **(-)**		
			Very good	42	Correlation with CTSPC: r = 0.10 (Minor Violence score); r = 0.14 **(-)**		
			Doubtful	49	Difference between abusive and non-abusive parents: Cohen's d = 0.08 **(-)**		
			Doubtful	43	Difference between abusive and non-abusive parents: Cohen's d = 0.02 **(-)**		
	PRCM	Vittrup et al. (2006)	Very good	244	Difference between mothers of 12-month-old and 48-month-old babies: Cohen's d = 1.79 **(+)**	+	**High** (no concern)

AAPI-2: Adult Adolescent Parenting Inventory-2; ACQ: Areas of Change Questionnaire; APQ: Alabama Parenting Questionnaire; APT: Analog Parenting Task; APT: Attitudes Toward Spanking; CAP: Child Abuse Potential inventory; CNQ: Child Neglect Questionnaire; CNS-MMS: Child Neglect Scales–Maternal Monitoring and Supervision scale; CPSS: Child Posttraumatic Stress Scale; CRC: Child's Relationship with Caretaker; CRI: Child Rearing Inventory; CRPB: Child Report on Parental Behavior; CTS: Conflict Tactics Scale; CTS-ES: Child Trauma Screen–Exposure Score; CTSPC: Conflict Tactics Scales: Parent–Child version; CV: Child Vignettes; DPICS: Dyadic Parent–child Interaction Coding System; ECBI: Eyberg Child Behavior Inventory; FAM :Family Assessment Measure; FM-CA: Family Maltreatment–Child Abuse criteria; FSI-R: Family Socialization Interview–Revised; HOME: Home Observations for the Measurement of the Environment; 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; K-BIT: Kaufman Brief Intelligence Test; MCNS: Mother–Child Neglect Scale; MCNS-SF: Mother–Child Neglect Scale–Short Form; MIS: Maternal Interaction Scale; NS: Neglect Scale; PAT: Parent Attribution Test; P-CAAM: Parent–Child Aggression Acceptability Movie task; PDMI: Parental Discipline Methods Interview; PKBS: Preschool and Kindergarten Behavior Scales; POQ : Parent Opinion Questionnaire; PPQ: Parent Practices Questionnaire; PRCM : Parental Response to Child Misbehavior questionnaire; PS: Parenting Scale; PSI-SF: Parenting Stress Index–Short Form; SBS-SV: Shaken Baby Syndrome awareness assessment–Short Version; SCL-90-R: Symptom CheckList 90–Revised; CFA: Confirmatory Factor Analysis; CFI: Comparative Fit Index; CTT: Classical Test Theory; EFA: Exploratory Factor Analysis; IRT: Item Response Theory; NR: Not Reported; RMSEA: Root Mean Square Error of Approximation; TLI: Tucker–Lewis Index; DIF: Differential Item Functioning; ICC: Intraclass Correlation Coefficient; + = Sufficient rating; ? = Indeterminate rating; - = Insufficient rating; ± = Inconsistent rating; High = High level of confidence in overall ratings; Moderate = Moderate level of confidence in overall ratings; Low = Low level of confidence in overall ratings; Very Low = Very low level of confidence in overall ratings; For the hypothesis testing on difference between subgroups, Cohen's d was calculated using the formulas presented by Friedman (1968), and Thalheimer and Cook (2002).

Article 3

Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Responsiveness of Parent- or Caregiver-Reported Child Maltreatment Instruments to Parenting Interventions. *Trauma, Violence, & Abuse.* Manuscript submitted for publication.

Trauma, Violence, & Abuse

A Systematic Review on Evaluating Responsiveness of Parent- or Caregiver-Reported Child Maltreatment Measures for Interventions

Journal:	*Trauma, Violence, & Abuse*
Manuscript ID	Draft
Manuscript Type:	Review Manuscripts
Keywords:	Child Abuse, Prevention of Child Abuse < Child Abuse, Treatment/Intervention < Child Abuse, Assessment < Domestic Violence < Cultural Contexts, Neglect < Child Abuse

A Systematic Review on Evaluating Responsiveness of Parent- or Caregiver-Reported Child Maltreatment Measures for Interventions

Abstract

Aims: Child maltreatment (CM) is a global public health and social problem, resulting in serious long-term health and socioeconomic consequences. As parents are the most common perpetrators of CM, parenting interventions is an appropriate strategy to prevent CM. However, research on parenting interventions on CM has been hampered by lack of consensus on what measures are most responsive to detect a reduction in parental maltreating behaviours after parenting intervention. This systematic review aimed to evaluate the responsiveness of all current parent- or caregiver-reported CM measures.

Methods: A systematic search was conducted in CINAHL, Embase, ERIC, PsycINFO, PubMed, and Sociological Abstracts. The quality of studies and responsiveness of the measures were evaluated using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines for systematic reviews of patient-reported outcome measures. Only measures developed and published in English were included.

Results: Sixty-nine articles reported on responsiveness of fifteen identified measures. The study quality was overall adequate. The responsiveness of the measures was overall insufficient or not reported; high-quality evidence on responsiveness was limited.

Conclusions: Only the Physical Abuse subscale of the International Society for the Prevention of Child Abuse and Neglect (ISPCAN) Child Abuse Screening Tool for use in trials can be recommended as most responsive for use in parenting interventions, with high quality evidence supporting sufficient responsiveness. All other overall scales or subscales of the fifteen included measures were identified as promising based on current data on responsiveness. Additional psychometric evidence is required before they can be recommended.

Keywords: assessment; child abuse; COSMIN; measure; parent report; measurement properties; responsiveness

Introduction

Child maltreatment (CM) refers to the abuse and neglect experienced by a child under the age of 18 years, resulting in actual or potential harm to the child (World Health Organization [WHO], 2016). This conceptual definition can be categorised into four subtypes of CM (Slep et al., 2015; WHO, 2006): (1) physical abuse (non-accidental acts of physical force causing actual or potential physical harm), (2) emotional abuse (non-accidental verbal or symbolic acts causing significant psychological harm), (3) sexual abuse (sexual acts using a child for sexual gratification), and (4) neglect (failure in providing a child with needed age-appropriate care in health, education, emotional development, nutrition, shelter, and safe living conditions).

CM is a pervasive public health problem and societal burden. Worldwide, more than 1 billion children (aged 2 to 17 years) are annually exposed to at least one type of CM (Hillis et al., 2016). Early exposure to multiple types and repeated episodes of CM can cause childhood adverse outcomes such as physical injuries, mental health problems and death (Coley et al., 2014; Gilbert et al., 2009; Louwers et al., 2011; MacKenzie et al., 2015). Childhood physical and mental health problems due to exposure to CM can also persist into adulthood and cause adverse outcomes such as chronic diseases, depression, substance use, and suicidal behaviour (Currie & Widom, 2010; Hughes et al., 2017). Furthermore, CM is associated with high economic burden. For example, the lifetime estimated financial cost for each victim of CM is approximately USD 210,012 which is higher than other costly health conditions such as stroke (USD 159,846) or type 2 diabetes (USD 181,000; Fang et al., 2012). Given the great health and societal impact of CM, the importance of preventing CM cannot be overstated.

One of the main strategies to prevent CM is interventions aimed at improving parenting skills (Hinds & Giardino, 2017; WHO, 2016). Parents make up the majority of CM perpetrators (Devries et al., 2018; Sedlak et al., 2010). For example, every year more than 80 percent of CM perpetrators in the US are parents (Institute of Medicine & National Research Council, 2014). Poor parenting skills are a significant risk factor for CM (Knerr et al., 2013). For this reason, a number of parenting skill interventions have been developed targeting parents with the aim to reduce CM (Gubbels et al., 2019).

Research on parenting interventions to reduce CM is hampered by the lack of consensus on which CM measures is most responsive to detecting treatment effects following interventions for reducing CM by parents (Fluke et al., 2020). Many CM efficacy studies used indirect measures (e.g., measures evaluating parental depression and parental stress) that do not capture actual reductions in CM (Mikton & Butchart, 2009), and parent survey measures (e.g., measures estimating prevalence of CM) that may be less sensitive to measure actual reductions in parental maltreating behaviours in intervention studies (Cluver et al., 2016). Furthermore, some studies used CM observational measures (i.e., outsiders' observation parenting behaviours) that cannot capture extreme cases of parental maltreating behaviours, such as using harsh physical discipline (Presser & Stinson, 1998) and leaving a child at home without supervision (Singer et al., 1995). Furthermore, they are considerably more complex, costly, and time-consuming to administer compared with parent report measures (Morsbach & Prinz, 2006). However, the accuracy of parents reporting on their own perpetration of CM is also controversial as parents tend to respond in socially desirable ways (i.e., social desirability bias; Milner & Crouch, 1997) and struggle remembering past events (i.e., recall bias, Greenhoot, 2013). Therefore, identifying high-quality parent- or caregiver-reported measures that are sensitive enough to measure change over time in response to a parenting intervention, is essential to detect intervention effects accurately.

The quality of a measure is largely determined by its psychometric properties (Karanicolas et al., 2009) and consists of the following three overarching constructs: validity (the extent to which a measure assesses the construct it is intended to assess), reliability (the extent to which scores for patients who have not changed are the same for repeated assessments), and responsiveness(the ability to detect change over time in the construct measured; Prinsen et al., 2018). The best way for selecting the most valid, reliable, and responsive measures is to systematically review the psychometric properties of existing measures (Scholtes et al., 2011). Recently, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) group has updated comprehensive guidelines for conducting systematic reviews on psychometric properties of health measures (Prinsen et al., 2018; Terwee et al., 2018). The COSMIN guidelines provide the following useful tools: a taxonomy on terms and definitions of each psychometric property (Mokkink et al., 2010b); a checklist for assessing the methodological quality of psychometric studies (Mokkink, de Vet, et al., 2018); quality criteria for evaluating single-study results on a psychometric property (Prinsen et al., 2018; Terwee et al., 2018); and a rating system summarising all study results on each psychometric property and grading quality of all evidence used for assessing both the methodological and the psychometric quality (Prinsen et al., 2018; Terwee et al., 2018).

For evaluating responsiveness, the COSMIN guidelines suggest testing the following two approaches: criterion and construct (Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018). The criterion approach assesses the relationship of change scores between the measures and a gold standard(i.e., a single error-free reference measure; Naaktgeboren et al., 2013) for detecting the effect of intervention for preventing CM (i.e., comparison to a gold standard; Mokkink, Prinsen, et al., 2018). If there is no gold standard assessment available, as is the case of measuring the construct CM (Bailhache et al., 2013), the COSMIN guidelines (Mokkink, Prinsen, et al., 2018) recommend using the construct approach instead. The construct approach assesses the following three aspects: (1) the relationship between the change scores on the reviewed measures and other measures used to assess the same construct (i.e., comparison with other outcome measures); (2) the mean difference in change scores for measures between different subgroups (i.e., comparison between subgroups); and (3) the mean difference in change scores for measures before and after intervention (i.e., comparison before and after intervention).

Only one systematic review to date has evaluated responsiveness of CM measures (Saini et al., 2019), which identified child or clinician report CM measures and evaluated the measures' responsiveness. However, the authors did not include parent- or caregiver-reported measures. Furthermore, the authors did not use the recently revised COSMIN guidelines (Prinsen et al., 2018; Terwee et al., 2018), but old versions of the COSMIN checklist (Mokkink et al., 2010a) and quality criteria (Terwee et al., 2007) to assess the methodological quality of included studies and the responsiveness of measures. These older versions of the checklist and quality criteria have neither a standardised method for summarising evidence on each psychometric property including responsiveness, nor for grading quality of evidence when deciding whether to recommend a measure for research and clinical use (Prinsen et al., 2018; Terwee et al., 2018). To overcome these limitations of older versions, the COSMIN guidelines have been thoroughly revised in recent years (Prinsen et al., 2018; Terwee et al., 2018).

Authors et al. (2020a; 2020b [reference blinded for review]) published two psychometric reviews on parent- or caregiver-reported measures on CM using the latest versions of the COSMIN guidelines (Prinsen et al., 2018; Terwee et al., 2018). Firstly, Authors et al. (2020a [reference blinded for review]) assessed measures' content validity for being the most important psychometric property when selecting a measure (Prinsen et al., 2018; Prinsen et al., 2016); if the content (e.g., items) of measures inadequately represents the construct(s) to be assessed, the evaluation of other psychometric properties is of limited value. This review by Authors et al. (2020a [reference blinded for review]) identified 15 parent- or caregiver-reported measures developed and published in English, assessed parents' or caregivers' attitude toward CM or perpetration of CM, and assessed one or more of the four categories of CM (i.e., physical abuse, emotional abuse, sexual abuse, neglect; Slep et al., 2015; WHO, 2006; WHO, 1999). No high-quality evidence supporting insufficient content validity was found for any of the 15 included measures, thus rendering them suitable for further psychometric evaluation. In a subsequent psychometric review, Authors et al. (2020b [reference blinded for review]) reported on the other psychometric properties (reliabilities and validities other than content validity) of the 15 included measures (Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018). However, responsiveness was outside the scope of this review by Authors et al. (2020b [reference blinded for review]), given that the search strategy needed to be adjusted to identify studies appropriate to determine responsiveness. No systematic review on the responsiveness of parent- or caregiver-reported measures on CM has been published to date.

Study Aim

The aim of this systematic review was to evaluate responsiveness of all current parent- or caregiver-reported CM measures limited to one aspect of the construct approach for responsiveness (i.e., the comparison before and after interventions using the COSMIN guidelines; Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018). Due to the size, scope, and complexity of reporting, the remaining aspects of the construct approach for responsiveness (i.e., comparison with other outcome measures and comparison between subgroups) were beyond the scope of the present review.

Method

This systematic review followed the guidelines of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement (Moher et al., 2009) and the COSMIN guidelines (Prinsen et al., 2018). This review followed the following three consecutive steps (see Figure 1):

- Step 1: *Systematic literature search* formulating eligibility criteria (Step 1.1), searching the literature, and selecting studies (Step 1.2);
- Step 2: *Evaluation of the methodological quality of studies* on responsiveness of measures using the COSMIN Risk of Bias checklist; and
- Step 3: *Evaluation of responsiveness of measures* by rating the result of single studies against the criteria for responsiveness (Step 3.1), rating the pooled results of all studies per measure (Step 3.2), and grading the quality of evidence on responsiveness (Step 3.3).

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

Insert Figure 1 about here

Step 1. Systematic Literature Search

The systematic literature search was performed formulating eligibility criteria (Step 1.1) and searching literature and selecting studies (Step 1.2) in accordance with the PRISMA statement (Moher et al., 2009).

Eligibility criteria (Step 1.1)

To be selected for this current review, articles had to meet the following three eligibility criteria: (1) journal articles were published in English; (2) articles involved parents or caregivers to assess their attitudes toward CM or change maltreating behaviours toward their children; (3) articles reported on responsiveness data (i.e., change scores of a measure before and after an intervention) for one or more of the fifteen parent- or caregiver-reported CM measures (see Table 1) as identified in the companion systematic reviews by Authors et al. (2020a; 2020b [reference blinded for review]).

Insert Table 1 about here

Literature search and study selection (Step 1.2)

To identify eligible articles that reported on responsiveness of the selected 15 measures, systematic literature searches were performed in six electronic databases: CINAHL, Embase, ERIC, PsycINFO, PubMed, and Sociological Abstracts. All database searches were conducted in January 2020 with an updated search conducted in March 2021. Free text terms were used to search databases and to retrieve all publication prior to March 2021 (see Online Supplemental Table S1).

Titles and abstracts retrieved from database searches were screened to identify eligible journal articles on responsiveness of the 15 measures by two reviewers independently; one reviewer screened all abstracts, while the other reviewer screened a random selection of fifty percent of all abstracts. All full texts of eligible abstracts were retrieved and assessed by both reviewers independently. Any disagreements between both reviewers were resolved via a consensus decision including a third reviewer. Inter-rater agreement was determined using Cohen's weighted κ (Cohen & Humphreys, 1968) and interpreted as: very good (κ = 0.81–1.00), good (κ = 0.61–0.80), moderate (κ = 0.41–0.60), fair (κ = 0.21–0.40), and poor (κ = 0.00–0.20) agreement (Altman, 1991). Reference lists of all included full-text articles were searched manually to identify additional eligible journal articles. Hand searching of reference lists was performed by one reviewer and identified journal articles were checked by the second reviewer.

After identifying eligible articles, a distinction was made between 'an article' and 'an analysis at scale level. An article may assess responsiveness of: a) one overall scale or b) one overall scale and several unidimensional subscales (i.e., subscale(s) consisting of multiple items that assess a single underlying construct) or c) several unidimensional subscales. Conversely, an analysis at scale level assess only one overall scale or one unidimensional subscale, thus making it the lowest unit of analysis to determine responsiveness (Mokkink, Prinsen, et al., 2018). This is an important distinction as authors report on the effectiveness of interventions using both overall scales and subscales; hence the need to assess responsiveness of both all overall scales as well as unidimensional subscales. The unidimensionality of a subscale was confirmed if data could be identified in the literature supporting the internal structure of the subscale (i.e., conducted factor analysis and internal consistency using Cronbach's alpha for each subscale) (i.e., conducted factor analysis and internal consistency using Cronbach's alpha for each subscale; Mokkink, de Vet, et al., 2018). The confirmed subscale can be used as an independent measure besides an overall scale (Mokkink, Prinsen, et al., 2018). Included articles reporting data on responsiveness of overall scales or confirmed subscales were divided into separate 'analyses at scale level' (i.e., each assessment of responsiveness per scale or unidimensional subscale) for evaluation of methodological quality of studies (Step 2).

Step 2. Evaluation of Methodological Quality of Studies

The methodological quality of the included studies on the responsiveness of the selected 15 measures was assessed using the COSMIN Risk of Bias checklist (Mokkink, de Vet, et al., 2018). The checklist contains three items for responsiveness on comparison *before and after intervention* (see Online Supplemental Table S2), which rate the quality of study design and the robustness of statistical methods used in studies on a measure's responsiveness to change following intervention (Mokkink, de Vet, et al., 2018). Each checklist item was scored on a four-point rating scale: *inadequate* = 1, *doubtful* =2, *adequate* = 3; and *very good* = 4 (Mokkink, de Vet, et al., 2018). A total rating for responsiveness was determined by the ratio of *'the obtained total score minus the minimum possible score'* to *'the maximum possible score minus the minimum possible score'* (Cordier et al., 2015). This ratio score method was preferred over the worst score counts method as suggested by the COSMIN guidelines (i.e., determining total ratings based on the lowest rating of any of the checklist items; Mokkink, Prinsen, et al., 2018). The worst score counts method is likely to prohibit detecting subtle differences in methodological quality between studies (Speyer et al., 2014). Accordingly, the total score of methodological quality ratings on responsiveness was reported as a percentage rating and can be interpreted as follows: 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 independent reviewers rated the methodological quality. Any disagreements were resolved by consensus. The interrater agreement between both reviewers was determined by weighted κ (Cohen & Humphreys, 1968).

After assessing methodological quality of the included studies on responsiveness, the following data from the included studies and measures were extracted using a data extraction template that is part of the COSMIN manual (Mokkink, Prinsen, et al., 2018): (1) study characteristics; (2) measure characteristics; and (3) study results on responsiveness. (i.e., conducted factor analysis and internal consistency using Cronbach's alpha for each subscale; Mokkink, de Vet, et al., 2018) The extraction was done by one reviewer and a second reviewer cross-checked the accuracy and completeness of the extracted data. All extracted data were used for evaluation of responsiveness of measures (Step 3).

Step 3. Evaluation of responsiveness of measures

The responsiveness of measures was assessed in three sequential steps: Step 3.1 rating the results of single studies, Step 3.2 rating the pooled results of all studies per measure, and Step 3.3 grading the quality of evidence on responsiveness. All ratings were scored by two independent reviewers separately, after which consensus ratings were determined based on reviewers group discussion.

Rating the results of single studies (Step 3.1)

Rating the results of single studies using quality criteria for responsiveness was limited to the comparison of *before and after intervention*. The results of responsiveness to change in scores following an intervention for each individual study were rated as *sufficient*

(+ = meeting the quality criteria), *insufficient* (- = below the quality criteria), or *indeterminate* (? = lack of robust evidence of meeting the quality criteria) against predefined criteria for good responsiveness (Mokkink, Prinsen, et al., 2018; see Online Supplemental Table S3). For a sufficient (+) rating on single study results, robust data on change scores before and after intervention on the selected measures should be available to allow calculation of the standardised mean difference (SMD) and confirm at least medium effect size (i.e., Hedges' g ≥ 0.50; Cohen, 1988); insufficient (-) ratings showed calculated SMDs below medium effect size (i.e., Hedges' g < 0.50; Cohen, 1988). Single study results that did not provide robust data to allow SMD calculations (Hedges' g; Hedges & Olkin, 2014) were rated as indeterminate (?).

Rating the pooled results of all studies per measure (Step 3.2)

All results on responsiveness from available studies per measure were quantitatively pooled into overall ratings of the responsiveness per measure (Prinsen et al., 2018). An overall sufficient (+), insufficient (-), or indeterminate (?) rating for responsiveness was given using the same quality criteria for good responsiveness (Mokkink, Prinsen, et al., 2018) (see Online Supplemental Table S3). For an overall sufficient (+) rating on responsiveness per measure, the pooled SMD must be at least medium effect size (i.e., Hedges' g ≥ 0.50; Cohen, 1988). For an overall insufficient (-) rating, the pooled SMD falls below medium effect size (i.e., Hedges' g < 0.50; Cohen, 1988). For an overall indeterminate (?) rating, all results represent insufficiently robust data, thus not supporting the calculation of the pooled SMD (Hedges' g; Hedges & Olkin, 2014). Hedges' *g* for both single study results (Step 3.1) and all study results per measure (Step 3.2) was calculated as proposed by Borenstein et al. (2009) and using the Comprehensive Meta-Analysis (CMA) software version 3.0 (Borenstein et al., 2013). In cases where at least moderate heterogeneity (i.e., Higgins' $I^2 \geq 50\%$; Higgins et al., 2003) in effect sizes across studies were calculated (Higgins et al., 2003), a random effect model (Borenstein et al., 2009) was used to calculate pooled effect size. In cases where low heterogeneity (i.e., $0 \leq I^2 < 50\%$; Higgins et al., 2003) was calculated, a fixed effect model was used by giving relatively greater weight to individual studies with larger sample sizes in contrast to the random effect model that does not take into account the weight of samples sizes when calculating pooled effect size (Borenstein et al., 2009).

Grading the quality of evidence on responsiveness (Step 3.3)

The quality of the evidence (i.e., the entire body of evidence used for overall ratings on responsiveness per measure) was graded as *high*, *moderate*, *low*, and *very low* evidence, using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Mokkink, Prinsen, et al., 2018; see Online Supplemental Table S4). The modified GRADE approach assumes that the initial quality of evidence used for overall ratings is of high quality. Subsequently, the quality of evidence is downgraded by one to three levels (to moderate, low, or very low) when there are serious (-1: one level down), very serious (-2: two levels down), or extremely serious (-3: three levels down) concerns across the evidence. The quality ratings of evidence were determined taking into consideration the following four factors: (a) risk of bias (limitations in the methodological quality of studies (Step 2); (b) inconsistency (heterogeneity in pooled results of studies (Step 3.2); (c) indirectness (evidence from different populations other than the target population in the review); and (d) imprecision (a low total sample size included in the studies) (Mokkink, Prinsen, et al., 2018). Quality of evidence should not be graded if the overall rating was indeterminate (?) due to lack of robust evidence (Prinsen et al., 2018). More detailed information on grading quality of evidence can be found in the COSMIN manual for systematic reviews of measures (Mokkink, Prinsen, et al., 2018).

Results

Systematic Literature Searches (Step 1)

A total of 1,475 abstracts were identified from six electronic databases after removing duplicates: 273 records in CINAHL; 129 records in Embase; 77 records in ERIC; 1,085 records in PsycINFO; 165 records in PubMed; and 84 records in Sociological Abstracts. Figure 2 shows the flow chart of the studies identified during literature searching and study selection (Step 1.2) in accordance with PRISMA (Moher et al., 2009). A total of 229 full-text articles were assessed for eligibility, of which 58 journal articles met all inclusion criteria: 171 articles did not meet at least one of the inclusion criteria. Reference checking of the included 58 journal articles identified 11 additional articles meeting all inclusion criteria. As a result, 69 journal articles reporting on the responsiveness of 15 parent- or caregiver-reported CM measures, were included in this review. General characteristics of the included 69 articles are presented in Online Supplemental Table S5. Furthermore, as most included articles presented data on the responsiveness of more than one overall scale or unidimensional subscale, the included 69 articles contained 223 analyses at scale level for the quality assessment of the study (step 2) and the responsiveness (step 3). The interrater agreement for selection of articles between two reviewers was very good (Altman, 1991): weighted κ for abstract selection = 0.81 (95% confidence interval [CI] = [0.74, 0.88]); weighted κ for article selection = 0.83 (95% CI [0.75, 0.90]).

Insert Figure 2 about here

Methodological Quality of the Included Studies (Step 2)

The methodological quality of the 223analyses at scale level in 69 included articles on responsiveness was assessed using the COSMIN Risk of Bias checklist (Mokkink, de Vet, et al., 2018). Table 2 presents an overview of all methodological quality ratings for the 223 analyses at scale level on responsiveness of 15 measures. In total, 57% (127/223) of analyses at scale level reporting on responsiveness were scored as having good or adequate methodological quality, whereas 43% (96/223) were scored as having doubtful or inadequate quality. The inter-rater agreement for study quality assessment between both reviewers was very good: weighted κ = 0.83 (95% CI [0.77, 0.91]).

Insert Table 2 about here

Responsiveness and Quality of Evidence of Measures (Step 3)

Table 3 summarises ratings on responsiveness for analyses at scale level (Step 3.1); the results of analyses at scale level and their quality ratings are presented in detail in Online Supplemental Table S6. All extracted data on responsiveness from the 223 analyses at scale level (from 69 included articles) were evaluated against the criteria for good responsiveness (Prinsen et al., 2018; see Online Supplemental Table S3). Of all 223 ratings on responsiveness data of analyses at scale level, only four ratings received an indeterminate rating due to less robust data being reported on responsiveness (see Table 3). All other analyses at scale level results received either a sufficient (69/223) or an insufficient (150/223) rating on responsiveness.

Insert Table 3 about here

Table 4 summarises the overall responsiveness ratings (Step 3.2) and the quality of evidence (Step 3.3) for responsiveness per overall scale or subscale of all 15 measures. The pooled results of all analyses at scale level on responsiveness for each overall scale or subscale and detailed reasons for downgrading on quality of all evidence used for the overall ratings, are displayed in Online Supplemental Table S7. The overall rating for pooled results of analyses at scale level on responsiveness for each overall scale or subscale were evaluated using the same criteria for good responsiveness (Prinsen et al., 2018; see Online Supplemental Table S3). None of the overall scales and subscales for the 15 measures received an indeterminate overall rating for responsiveness (see Table 4). Almost half of all measures (7 out of 15) received 'not reported' (NR) as overall ratings because no data on responsiveness could be retrieved from the included studies. Of the remaining 8 measures, only three measures and one subscale received an overall sufficient responsiveness; all the others received an overall insufficient rating on responsiveness. In addition, the quality of evidence (confidence level for the overall rating per overall scale or subscale) was evaluated using the modified GRADE approach (Prinsen et al., 2018; see Online Supplemental Table S4). Again, measures (7 out of 15) that had not reported on responsiveness data, received 'not reported' (NR) as quality ratings of evidence (see Table 4). Of the remaining 8 measures, only one single subscale reported a high-quality evidence supporting its overall rating on responsiveness; all the others reported either moderate or low quality evidence for their overall ratings on responsiveness.

Insert Table 4 about here

Discussion

The aim of this systematic review was to evaluate quality of responsiveness (comparison before and after interventions) of all current parent- or caregiver-reported measures on CM by parents or caregivers using the recently revised COSMIN guidelines. This review identified 69 articles that reported on responsiveness of the fifteen parent- or caregiver-reported CM measures identified by Authors et al. (2020a; 2020b [reference blinded for review]). The identified individual articles contained 223 analyses at scale level for each overall scale and subscale of the 15 measures. The methodological quality of the included studies was generally adequate. However, responsiveness data were only retrieved from the literature for about half of the included measures (8/15). Moreover, there is lack of high-quality evidence to support that the responsiveness of the measures is either sufficient or insufficient to determine the effect of parenting interventions for preventing CM. Only one subscale (ICAST-Trial [physical abuse]) reported high-quality evidence that it is sufficiently responsive to change before and after intervention. Due to lack of high-quality evidence on the responsiveness of overall scales and subscales, all of the measures included in this review may still have the potential to be used in interventions. However, additional robust research focusing on their responsiveness is needed before these measures can be recommneded for use to determine the effectivenss of interventions (before and after measurment).

Methodological Quality of the Included Studies

In terms of quality of study design, most of analyses at scale level (81 of 96) reporting doubtful or inadequate methodological quality (see Online Supplemental Table S6), as they had a methodological shortcoming (i.e., most studies were not designed as randomised controlled trials [RCTs]). As RCT randomly allocates study samples either to an intervention or a control group, it can minimise selection bias and confounding variables such as different sample characteristics (Altman, 1991). For this reason, RCT is considered to be the most powerful study design to estimate unbiased effect size of an intervention (Altman, 1991). However, only few RCTs have been conducted on the effectiveness of interventions to prevent CM due to practical issues related to cost effectiveness and ethical issues related to this socially sensitive research topic (van der Put et al., 2018). For this reason, if only RCT studies were to be included in this review, much data on responsiveness of parent- or caregiver-reported CM measures would have been excluded. This reasoning is also in line with a meta-analysis carried out by Gubbels et al. (2019), which noted that RCTs are rare in the field of CM. Thus, although many analyses at scale level showed poor methodological quality due to shortcomings in their study designs, no limitations to study design were applied in this review when retrieving data on responsiveness from the literature.

In terms of robustness of statistical methods, most of the analyses at scale level (78 of 96) were rated as having doubtful or inadequate methodological quality because they used a less robust statistical analysis, such as a paired *t*-test or a repeated-measures analysis of variance (ANOVA) reporting only *p*-values (see Online Supplemental Table S6). The *p*-value is an inappropriate measure of responsiveness (Mokkink, de Vet, et al., 2018) for the following two reasons: (1) it is only a statistic to confirm whether the estimated mean difference in scores before and after an intervention is likely not caused by chance (i.e., statistical significance) and it does not reflect whether the magnitude of the estimated mean difference is large enough to detect a clinically important effect (i.e., clinical significance); and (2) it is dependent on sample size (Altman, 1991). To account for these limitations of a *p*-value, an effect size (e.g., Hedges' g, Hedges & Olkin, 2014) is preferred as an indicator of responsiveness in the COSMIN risk of bias checklist (Mokkink, Prinsen, et al., 2018), as it reflects the magnitude of mean difference before and after an intervention, regardless of sample sizes (Altman, 1991). However, most analyses at scale level only reported on *p*-values of paired *t*-tests or repeated-measures ANOVAs, resulting in doubtful or inadequate methodological study quality ratings.

For subscales, the methodological quality of studies was reported in only three out of eight measures reporting data on their responsiveness (AAPI-2, CTSPC, and ICAST-Trial). For the remaining five measures (APT, FM-CA, MCNS, POQ, and PRCM), the methodological quality of their subscales was not rated as the internal structure of their subscales was unclear and not confirmed by statistical analyses (i.e., by conducting statistical analysis to determine the factor structure and internal consistency). If a subscale has an unclear internal structure and unidimensionality cannot be confirmed (i.e., all items assess one underlying construct), then the construct of the subscale's responsiveness has no further value (Prinsen et al., 2016), regardless of whether or not the subscale can detect treatment effects following intervention. For example, when a subscale on parental neglect also contains items that assess sexual abuse, the subscale would be of no use for capturing changes in parental neglect as different constructs are combined within the same subscale. However, most parent- or caregiver-reported CM measures has not been tested to confirm the internal structure of their subscales (Authors et al., 2020b [reference blinded for review]), which could lead to either underestimating or overestimating the effectiveness of CM interventions (Meinck et al., 2018).

Responsiveness of Measures

In general, evidence on responsiveness of a total of 25 overall scales or subscales was rated as either *sufficient* (3 overall scales and 1 subscale), *not reported* (7 overall scales), or *insufficient* (5 overall scales or 9 subscales). Insufficient responsiveness was due to not meeting the minimum criterion for good responsiveness (i.e., estimated effect size smaller than medium; Cohen, 1988). This review is based on current evidence on responsiveness as retrieved from the literature. Due to overall low quality of evidence of data, the estimated small effect sizes as presented in this review may change if future intervention studies provide high-quality evidence (Mokkink, Prinsen, et al., 2018). Therefore, the 14 measures for which no high-quality evidence could be identified, may still have potential to be used for detecting changes in parental maltreating behaviours towards their children after intervention, if high-quality evidence are provided to support their responsiveness in future studies. Another important consideration in relation to the overall low to medium effect sizes is the quality of interventions. The findings suggest that new approaches to parent focussed CM interventions need to be considered to improve outcomes for both children and parents. For three overall scales (APT, FM-CA, and POQ) and one subscale (ICAST-Trial [physical Abuse]), evidence on responsiveness was sufficient with estimated effect sizes higher than medium (Cohen, 1988). However, as quality of evidence for sufficient responsiveness of all three overall scales were rated as either moderate or low, the three overall scales need more robust evidence to be recommended for use in CM intervention. Only one single subscale (ICAST-Trial [Physical Abuse]) demonstrated high-quality evidence for responsiveness. Therefore, considering the most robust current evidence supporting sufficient responsiveness, only the Physical Abuse subscale of ICAST-Trial can be recommended as the most suitable measure for use in parenting interventions for reducing CM by parents.

Overall quality of evidence to support the responsiveness of parent- or caregiver-reported measures on CM was weak with mainly moderate to low ratings. The low quality of evidence was due to very inconsistent results across studies (i.e., substantial heterogeneity in the pooled effect sizes of studies). This substantial heterogeneity is in line with the previous meta-analysis on effects of parenting interventions to prevent CM by Chen and Chan (2016). The authors found a wide variation of effect sizes within groups of studies using the same measures on CM and between individual studies regardless of measures. Examining the influence of moderator variables on the heterogeneity, the authors found that characteristics of both sample (e.g., country income level and gender) and intervention (e.g., dosage and timing) contribute to significant between-study variance. However, there is no research, including Chen and Chan (2016), that focused on what variables contribute to the heterogeneity of effect sizes across studies on parenting interventions per parent- or caregiver-reported CM measure. Also, additional reasons for the poor evidence quality were small total sample sizes included in the studies (e.g., APT [$n < 50$] and POQ [$n < 100$]) and poor methodological quality of studies (e.g., FM-CA [only one study of adequate quality available]). Therefore, the quality of evidence to support the responsiveness of included measures was overall low due to concerns on inconsistent results across studies, small sample sizes and poor study quality.

Limitations

This systematic review has some limitations. Firstly, only measures developed in English and studies published in English were included. Accordingly, some findings on responsiveness of CM measures published in languages other than English may have been missed. Secondly, this review reported only on one aspect of the construct approach for responsiveness (comparison before and after intervention; Mokkink et al., 2010b); the other two aspects (comparison with other outcome measures and comparison between subgroups) were beyond the scope of the present review due to the size, scope, and complexity of reporting. Lastly, feasibility of measures and interpretability of change scores were also outside the scope of this review as neither feasibility nor interpretability are considered psychometric properties according to the COSMIN taxonomy, even though they are important characteristics to consider when selecting the most suitable measures (Mokkink, Prinsen, et al., 2018; Prinsen et al., 2018). One aspect of feasibility (i.e., cost of a measure), however, was described in Table 1.

Implications for Future Research and Practice

From the findings on the methodological quality of the included studies in this systematic review, three implications for future research and practice arise. First, future studies on responsiveness to compare changes before and after parenting interventions using parent- or caregiver-reported CM measures are encouraged to calculate and report the effect sizes, in addition to *p*-values. This is also in line with the recommendations of *Reporting Standards for Research in Psychology* by the American Psychological Association (APA, 2008). Next, to estimate unbiased effect sizes on responsiveness, more RCT studies using parent- or caregiver-reported CM measures should be conducted. Lastly, for data on the responsiveness of a measure's subscales to be meaningful, the internal structure of the measure should be confirmed using appropriate statistical analyses (i.e., factor analysis and internal consistency using Cronbach's alpha per subscale) resulting in subscales measuring a single underlying construct. For five measures (APT, FM-CA, MCNS, POQ, and PRCM) in particular, the internal structure is yet to be confirmed before further assessment of study quality and responsiveness is meaningful.

From the findings on the responsiveness of the included measures in this systematic review, another three implications for future research and practice arise. First, all overall scales or subscales of the 15 included measures need additional responsiveness studies due to lacking or low quality evidence to support the quality of their responsiveness, with the exception of the Physical Abuse subscale of ICAST-Trial which demonstrated high-quality evidence. Next, because of high-quality evidence supporting its sufficient responsiveness, the Physical Abuse subscale of ICAST-Trial could be recommended for use in parenting interventions to reduce physical abuse to their children. Lastly, future research needs to perform subgroup analyses to investigate whether the characteristics of samples (e.g., level of income and gender) and intervention (e.g., dosage and timing) contribute to the substantial heterogeneity in effect sizes on responsiveness of parent- or caregiver-reported CM measures (e.g., AAPI-2, CTSPC, ICAST-Trial, and PRCM reporting moderate to high heterogeneity in responsiveness across studies). The sub-group analyses may contribute to the selection and use of more culturally and contextually appropriate measures on CM in parenting interventions to reduce CM by parents.

Conclusion

This systematic review evaluated the responsiveness of 15 parent- or caregiver-reported measures on CM using the COSMIN guidelines. Evidence concerning responsiveness was limited and mostly of lower quality. Based on current available evidence on responsiveness, only one subscale (Physical Abuse subscale of ICAST-Trial) of all included measures can be recommended as the most suitable measure of physical abuse in parenting interventions to reduce CM by parents. All other overall scales or subscales of the included measures were identified as promising, but would still need further studies on their responsiveness before their use in clinical practice and research can be recommended.

References

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

American Psychological Association. (2008). Reporting standards for research in psychology: why do we need them? What might they be? *The American psychologist*, *63*(9), 839-851. https://doi.org/10.1037/0003-066X.63.9.839

Authors et al. (2020a). Reference blinded for review.

Authors et al. (2020b). Reference blinded for review.

Azar, S. T., & Rohrbeck, C. A. (1986). Child abuse and unrealistic expectations: Further validation of the Parent Opinion Questionnaire. *Journal of Consulting and Clinical Psychology*, *54*(6), 867-868. https://doi.org/10.1037/0022-006X.54.6.867

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*(1), 202. https://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.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive meta-analysis. Version 3*. In Biostat. https://www.meta-analysis.com/downloads/Meta-Analysis%20Manual%20V3.pdf

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introductionto meta-analysis*. John Wiley&Sons Ltd.

Chen, M., & Chan, K. L. (2016). Effects of Parenting Programs on Child Maltreatment Prevention: A Meta-Analysis. *Trauma, Violence, & Abuse*, *17*(1), 88-104. https://doi.org/10.1177/1524838014566718

Cluver, L., Meinck, F., Yakubovich, A., Doubt, J., Redfern, A., Ward, C., Salah, N., De Stone, S., Petersen, T., Mpimpilashe, P., Romero, R. H., Ncobo, L., Lachman, J., Tsoanyane, S., Shenderovich, Y., Loening, H., Byrne, J., Sherr, L., Kaplan, L., & Gardner, F. (2016). Reducing child abuse amongst adolescents in low- and middle-income countries: A pre-post trial in South Africa. *BMC Public Health*, *16*(1), 567. https://doi.org/10.1186/s12889-016-3262-z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, J., & Humphreys, L. H. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. https://doi.org/10.1037/h0026256

Coley, R. L., Kull, M. A., & Carrano, J. (2014). Parental endorsement of spanking and children's internalizing and externalizing problems in African American and Hispanic families. *Journal of Family Psychology*, *28*(1), 22-31. https://doi.org/10.1037/a0035272

Cordier, R., Speyer, R., Chen, Y. W., 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*(7), e0132299. https://doi.org/10.1371/journal.pone.0132299

Currie, J., & Widom, C. S. (2010). Long-term consequences of child abuse and neglect on adult economic well-being. *Child Maltreatment*, *15*(2), 111-120. https://doi.org/10.1177/1077559509355316

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., Butron Riveros, B. C., Watts, C., & Abrahams, N. (2018). Who perpetrates violence against children? A systematic analysis of age-specific and sex-specific data. *BMJ Paediatrics Open*, *2*(1), e000180. https://doi.org/10.1136/bmjpo-2017-000180

Fang, X., Brown, D. S., Florence, C. S., & Mercy, J. A. (2012). The economic burden of child maltreatment in the United States and implications for prevention. *Child Abuse & Neglect*, *36*(2), 156-165. https://doi.org/10.1016/j.chiabu.2011.10.006

Fluke, J. D., Tonmyr, L., Gray, J., Bettencourt Rodrigues, L., Bolter, F., Cash, S., Jud, A., Meinck, F., Casas Muñoz, A., O'Donnell, M., Pilkington, R., & Weaver, L. (2020). Child maltreatment data: A summary of progress, prospects and challenges. *Child Abuse & Neglect*, 104650. https://doi.org/10.1016/j.chiabu.2020.104650

Gilbert, R., Widom, C. S., Browne, K., Fergusson, D., Webb, E., & Janson, S. (2009). Burden and consequences of child maltreatment in high-income countries. *Lancet*, *373*(9657), 68-81. https://doi.org/10.1016/s0140-6736(08)61706-7

Gordon, D. A., Jones, R. H., & Nowicki, S. (1979). A Measure of Intensity of Parental Punishment. *Journal of Personality Assessment*, *43*(5), 485-496. https://doi.org/10.1207/s15327752jpa4305_9

Gubbels, J., van der Put, C. E., & Assink, M. (2019). The effectiveness of parent training programs for child maltreatment and their components: A meta-analysis. *International journal of environmental research and public health*, *16*(13), 2404. https://doi.org/10.3390/ijerph16132404

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Heyman, R. E., Snarr, J. D., Slep, A. M. S., Baucom, K. J. W., & Linkh, D. J. (2020). Self-reporting DSM–5/ICD-11 clinically significant intimate partner violence and child abuse: Convergent and response process validity. *Journal of Family Psychology*, *34*(1), 101-111. https://doi.org/10.1037/fam0000560

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

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*(3), e20154079. https://doi.org/10.1542/peds.2015-4079

Hinds, T. S., & Giardino, A. P. (2017). Policy Direction: Focus on Prevention. In T. S. Hinds & A. P. Giardino (Eds.), *Child Physical Abuse: Current Evidence, Clinical Practice, and Policy Directions*. Springer. https://doi.org/10.1007/978-3-319-61103-7

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

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., & Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. *Lancet Public Health*, *2*(8), e356-e366. https://doi.org/10.1016/s2468-2667(17)30118-4

Institute of Medicine, & National Research Council. (2014). *New Directions in Child Abuse and Neglect Research*. The National Academies Press. https://doi.org/10.17226/18331

Karanicolas, P. J., Bhandari, M., Kreder, H., Moroni, A., Richardson, M., Walter, S. D., Norman, G. R., Guyatt, G. H., & Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group. (2009). Evaluating Agreement: Conducting a Reliability Study. *Journal of Bone and Joint Surgery*, *91*(Supplement 3), 99-106. https://doi.org/10.2106/jbjs.H.01624

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*(4), 241-255. https://doi.org/10.1023/a:1013662132189

Knerr, W., Gardner, F., & Cluver, L. (2013). Improving positive parenting skills and reducing harsh and abusive parenting in low- and middle-income countries: a systematic review. *Prevention Science*, *14*(4), 352-363. https://doi.org/10.1007/s11121-012-0314-1

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*(3), 390-398. https://doi.org/10.1037/tra0000235

Lounds, J. J., Borkowski, J. G., & Whitman, T. L. (2004). Reliability and Validity of the Mother-Child Neglect Scale. *Child Maltreatment*, *9*(4), 371-381. https://doi.org/10.1177/1077559504269536

Louwers, E. C. F. M., Korfage, I. J., Affourtit, M. J., Scheewe, D. J. H., van de Merwe, M. H., Vooijs-Moulaert, F. A. F. S. R., Woltering, C. M. C., Jongejan, M. H. T. M., Ruige, M., Moll, H. A., & De Koning, H. J. (2011). Detection of child abuse in emergency departments: a multi-centre study. *Archives of Disease in Childhood*, *96*(5), 422-425. https://doi.org/10.1136/adc.2010.202358

MacKenzie, M. J., Nicklas, E., Brooks-Gunn, J., & Waldfogel, J. (2015). Spanking and children's externalizing behavior across the first decade of life: evidence for transactional processes. *J Youth Adolesc*, *44*(3), 658-669. https://doi.org/10.1007/s10964-014-0114-y

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. https://doi.org/10.1016/j.chiabu.2018.05.022

Mikton, C., & Butchart, A. (2009). Child maltreatment prevention: a systematic review of reviews. *Bulletin of the World Health Organization*, *87*(5), 353-361. https://doi.org/10.2471/blt.08.057075

Milner, J. S., & Crouch, J. L. (1997). Impact and detection of response distortions on parenting measures used to assess risk for child physical abuse. *Journal of Personality Assessment*, *69*(3), 633-650. https://doi.org/10.1207/s15327752jpa6903_15

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*(7), e1000097. https://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*(5), 1171-1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs)-User manual (version 1.0)*. https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

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*(4), 539-549. https://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*(7), 737-745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Morsbach, S. K., & Prinz, R. J. (2006). Understanding and Improving the Validity of Self-Report of Parenting. *Clinical Child and Family Psychology Review*, *9*(1), 1-21. https://doi.org/10.1007/s10567-006-0001-5

Naaktgeboren, C. A., Bertens, L. C. M., Smeden, M. v., Groot, J. A. H. d., Moons, K. G. M., & Reitsma, J. B. (2013). Value of composite reference standards in diagnostic research. *BMJ*, *347*, f5605. https://doi.org/10.1136/bmj.f5605

Presser, S., & Stinson, L. (1998). Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance. *American Sociological Review*, *63*(1), 137-145. https://doi.org/10.2307/2657486

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

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*(1), 449. https://doi.org/10.1186/s13063-016-1555-2

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*(3), 199-209. https://doi.org/10.1016/j.chiabu.2010.12.004

Russa, M. B., & Rodriguez, C. M. (2010). Physical discipline, escalation, and child abuse potential: psychometric evidence for the Analog Parenting Task. *Aggressive Behavior*, *36*(4), 251-260. https://doi.org/10.1002/ab.20345

Russell, B. S. (2010). Revisiting the Measurement of Shaken Baby Syndrome Awareness. *Child Abuse & Neglect*, *34*(9), 671-676. https://doi.org/10.1016/j.chiabu.2010.02.008

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. https://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*(3), 236-240. https://doi.org/https://doi.org/10.1016/j.injury.2010.11.042

Sedlak, A. J., Mettenburg, J., Basena, M., Peta, I., McPherson, K., & Greene, A. (2010). *Fourth national incidence study of child abuse and neglect (NIS-4): Report to Congress*. Administration for Children and Families.

Singer, E., Von Thurn, D. R., & Miller, E. R. (1995). CONFIDENTIALITY ASSURANCES AND RESPONSE: A QUANTITATIVE REVIEW OF THE EXPERIMENTAL LITERATURE. *Public Opinion Quarterly*, *59*(1), 66-77. https://doi.org/10.1086/269458

Slep, A. M. S., Heyman, R. E., & Foran, H. M. (2015). Child Maltreatment in DSM-5 and ICD-11. *Family Process*, *54*(1), 17-32. https://doi.org/10.1111/famp.12131

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*, 458678. https://doi.org/10.1155/2014/458678

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*(19), 3343-3366. https://doi.org/10.1177/0886260514563836

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*(4), 249-270. https://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): CTS: Parent-Child Version (CTSPC)*. Western Psychological Services.

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*(1), 34-42. https://doi.org/10.1016/j.jclinepi.2006.03.012

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*(5), 1159-1170. https://doi.org/10.1007/s11136-018-1829-0 van der Put, C. E., Assink, M., Gubbels, J., & Boekhout van Solinge, N. F. (2018). Identifying Effective Components of Child Maltreatment Interventions: A Meta-analysis. *Clinical Child and Family Psychology Review*, *21*(2), 171-202. https://doi.org/10.1007/s10567-017-0250-5

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*(6), 2055-2064. https://doi.org/10.1542/peds.2005-2204

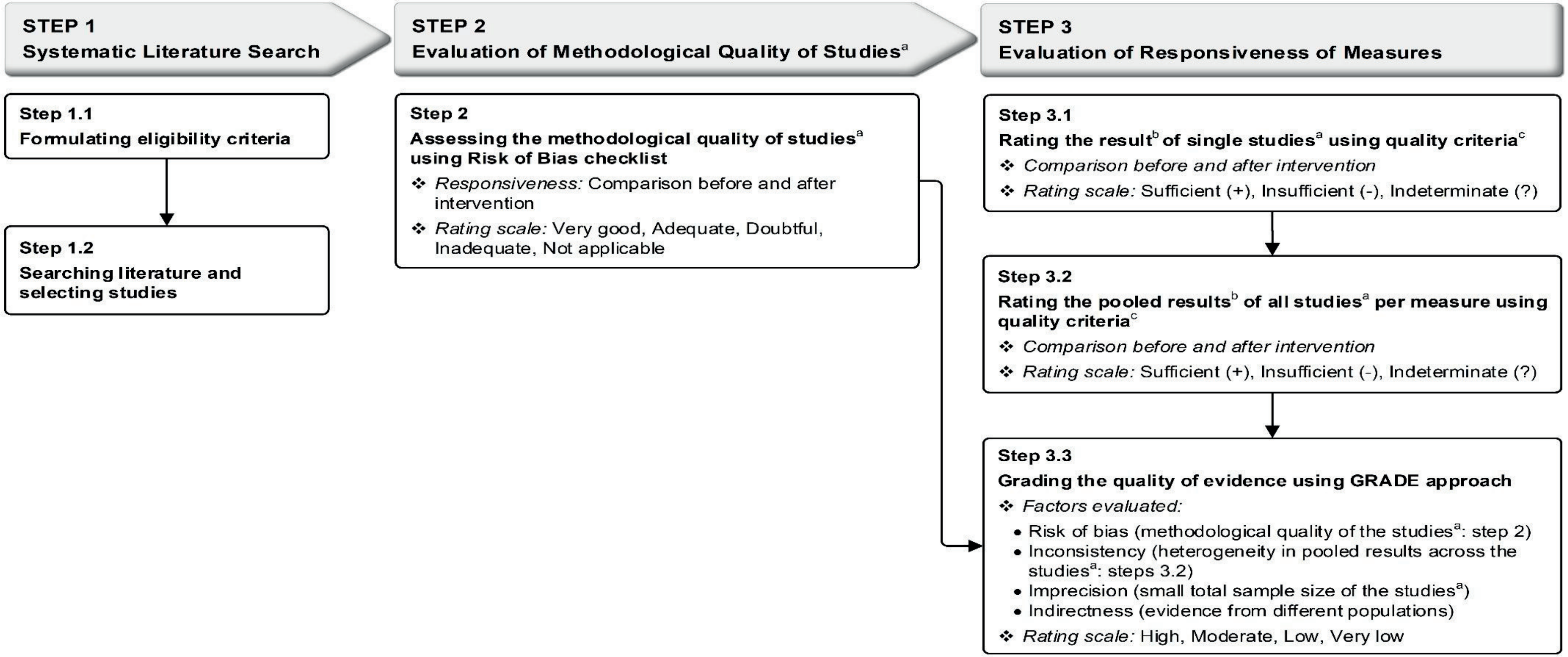
World Health Organization. (1999). *Report of the Consultation on Child Abuse Prevention*. Author. https://apps.who.int/iris/handle/10665/65900

World Health Organization. (2006). *Preventing child maltreatment: a guide to taking action and generating evidence*. Author. https://apps.who.int/iris/bitstream/handle/10665/43499/9241594365_eng.pdf

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*(2), 137-152. https://doi.org/10.1002/1098-2337(1989)15:2<137::AID-AB2480150202>3.0.CO;2-O

Figures


STEP 1
Systematic Literature Search
STEP 2
Evaluation of Methodological Quality of Studies[a]
STEP 3
Evaluation of Responsiveness of Measures
Step 1.1
Formulating eligibility criteria
Step 1.2
Searching literature and selecting studies
Step 2
Assessing the methodological quality of studies[a] using Risk of Bias checklist
❖ Responsiveness: Comparison before and after intervention
❖ Rating scale: Very good, Adequate, Doubtful, Inadequate, Not applicable
Step 3.1
Rating the result[b] of single studies[a] using quality criteria[c]
❖ Comparison before and after intervention
❖ Rating scale: Sufficient (+), Insufficient (-), Indeterminate (?)
Step 3.2
Rating the pooled results[b] of all studies[a] per measure using quality criteria[c]
❖ Comparison before and after intervention
❖ Rating scale: Sufficient (+), Insufficient (-), Indeterminate (?)
Step 3.3
Grading the quality of evidence using GRADE approach
❖ Factors evaluated:
• Risk of bias (methodological quality of the studies[a]: step 2)
• Inconsistency (heterogeneity in pooled results across the studies[a]: steps 3.2)
• Imprecision (small total sample size of the studies[a])
• Indirectness (evidence from different populations)
❖ Rating scale: High, Moderate, Low, Very low

Note. Content validity was evaluated in Authors et al. (2020a [reference blinded for review]); psychometric properties other than content validity and responsiveness were evaluated in Authors et al. (2020b [reference blinded for review]); for responsiveness, the criterion approach could not be assessed due to no existing gold standard and remaining aspects of construct approach, other than comparison before and after intervention, were outside the scope of this review.

[a] Studies reporting data on responsiveness of overall scales or unidimensional subscales (i.e., subscale[s] consisting of multiple items assessing a single underlying construct) were divided into separate 'analyses at scale level' (i.e., each assessment of responsiveness per scale or unidimensional subscale); each analysis at scale level served as a basic unit for evaluating methodological quality (step 2) and responsiveness (step 3; Mokkink, Prinsen, et al., 2018).

[b] Study result refers to a statistic to measure the effect size from change scores between before and after intervention (i.e., Hedges'g; Hedges & Olkin, 2014).

[c] Quality criteria refer to criteria for good responsiveness (Mokkink, Prinsen, et al., 2018).

Figure 1. Study design: Steps for Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Step 1) and COnsensus-based Standards for the selection of health Measurement INstruments processes (Step 2 and 3).

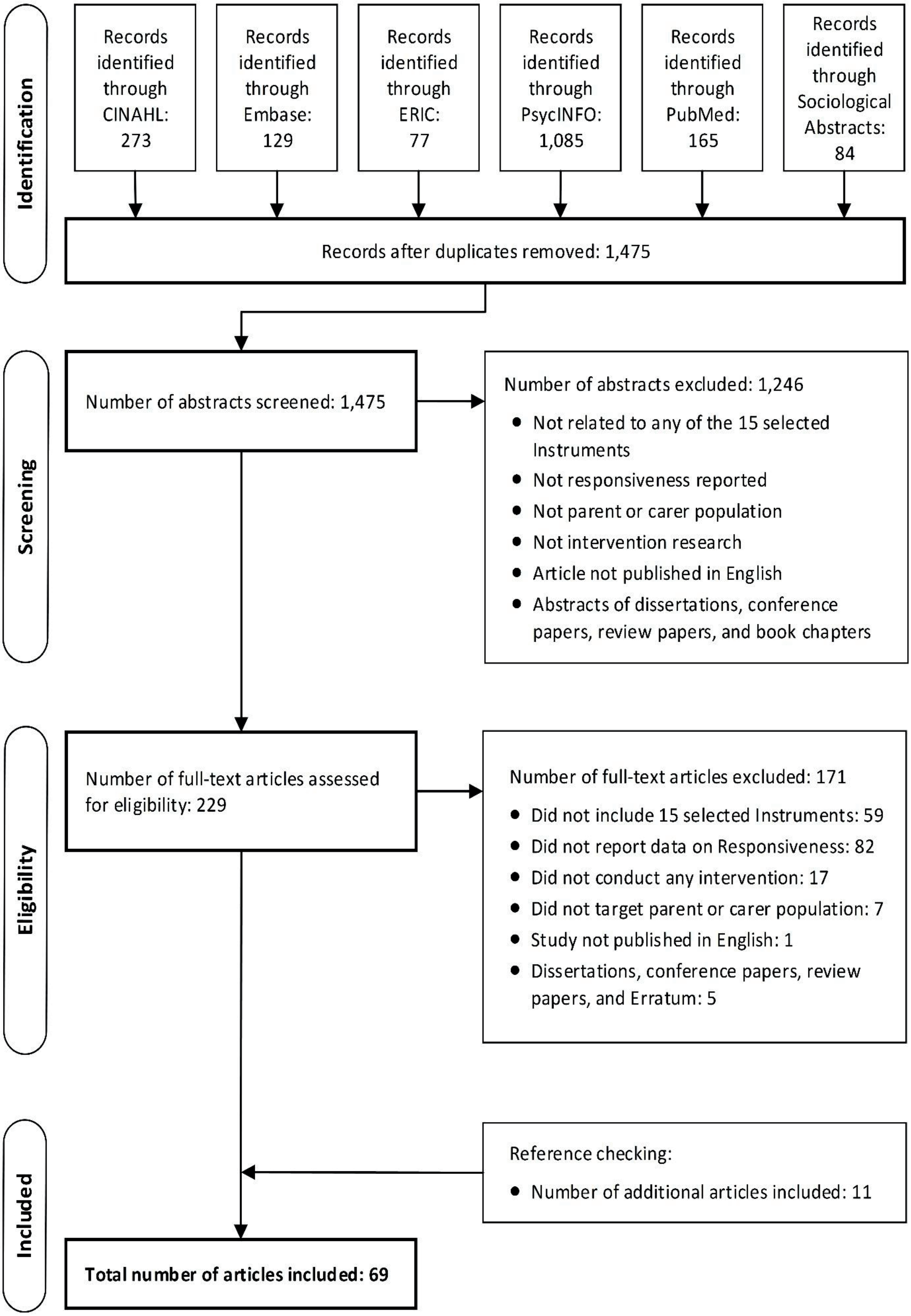

Identification
Records identified through CINAHL: 273
Records identified through Embase: 129
Records identified through ERIC: 77
Records identified through PsycINFO: 1,085
Records identified through PubMed: 165
Records identified through Sociological Abstracts: 84
Records after duplicates removed: 1,475
Screening
Number of abstracts screened: 1,475
Number of abstracts excluded: 1,246
Not related to any of the 15 selected Instruments
Not responsiveness reported
Not parent or carer population
Not intervention research
Article not published in English
Abstracts of dissertations, conference papers, review papers, and book chapters
Eligibility
Number of full-text articles assessed for eligibility: 229
Number of full-text articles excluded: 171
Did not include 15 selected Instruments: 59
Did not report data on Responsiveness: 82
Did not conduct any intervention: 17
Did not target parent or carer population: 7
Study not published in English: 1
Dissertations, conference papers, review papers, and Erratum: 5
Included
Reference checking:
Number of additional articles included: 11
Total number of articles included: 69

Figure 2. Flow diagram of the reviewing procedure based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Moher et al., 2009).

Tables

Table 1. *Characteristics of the Measures Assessing Child Maltreatment.*

Measure (abbreviation)	**Studies on development and validation**	**Main constructs**	**(Sub)scales**	**Target population**	**Purpose of use**	**No. of subscales (No. of items)**	**Range of score**	**Response Options**	**Recall period**	**Cost (mode of administration)**
Adult Adolescent Parenting Inventory-2 (AAPI-2)	Bavolek and Keene, 1999	Physical abuse; Emotional Abuse; Neglect	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	To identify maltreating parents/carers; To evaluate effectiveness of an intervention	5 (40)	0–50 (raw total scores per subscale are converted into standard scores: range 0–10)	5-point ordinal scale (*strongly disagree = 1 to strongly disagree = 5*)	Not specified	2 to 10 US dollars per administration (both paper- and web-based format)
Analog Parenting Task (APT)	Russa and Rodriguez, 2010; Zaidi et al. 1989	Physical abuse	Physical discipline; Escalation of physical discipline	Prospective parent populations	To identify maltreating parents/carers	2 (26)	0–26	10 nominal scale (from nonphysical discipline tactics to physical discipline tactics)	Not specified	Freely available (computer-based format)
Child Neglect Questionnaire (CNQ)	Stewart et al., 2015	Neglect	Physical neglect; Emotional neglect; Educational neglect; Supervision neglect	Parents with older children	To identify maltreating parents/carers	4 (46)	46–184	4-point ordinal scale (*always = 1 to never = 4*)	Past six months	Freely available (paper-based format)
Child Neglect Scales-Maternal Monitoring and Supervision Scale (CNS-MMS)	Kirisci et al., 2001	Neglect	Child neglect	Mothers	To evaluate effectiveness of an intervention	1 (11)	11–33	3-point ordinal scale (*hardly ever = 1 to often = 3*)	Past six months	Freely available (paper-based format)
Child Trauma Screen-Exposure Score (CTS-ES)	Lang and Connell, 2017	Physical abuse; Emotional abuse; Sexual abuse; Neglect	Potentially traumatic event	Caregivers	To identify maltreating parents/carers	1 (4)	0–4	Dichotomous scale (*no = 0 or yes = 1*)	Not specified	Freely available (paper-based format)

(continued)

Table 1. *(continued)*

Measure (abbreviation)	Studies on development and validation	Main constructs	(Sub)scales	Target population	Purpose of use	No. of subscales (No. of items)	Range of score	Response Options	Recall period	Cost (mode of administration)
Conflict Tactics Scales: Parent-Child version (CTSPC)	Straus et al., 1998; Straus et al., 2003	Physical abuse; Emotional abuse	Nonviolent discipline; Psychological aggression; Physical assault	Parents	To identify maltreating parents/carers; To evaluate effectiveness of an intervention	3 (22)	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)	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)	Past one year	62 US dollars per pack of 25 questionnaires (paper-based format)
Family Maltreatment-Child Abuse criteria (FM-CA)	Heyman et al., 2020	Physical abuse; Emotional Abuse	Physical child abuse; Psychological child abuse	Parents	To identify maltreating parents/carers; To evaluate effectiveness of an intervention	2 (27)	0–63	Dichotomous scale for physical child abuse subscale (I did = 0 or I never did = 1); 6-point ordinal scale for psychological child abuse subscale (never = 0 to more than once a day = 5)	Past one year	Freely available (computer-based format)
ISPCAN Child Abuse Screening Tool for use in Trials (ICAST-Trial)	Meinck et al., 2018	Physical abuse; Emotional abuse; Sexual abuse; Neglect	Physical abuse; Emotional abuse; Contact sexual abuse; Neglect	Caregivers	To evaluate effectiveness of an intervention	4 (14)	0-112	9-point ordinal scale (never = 0 to more than 8 times = 8)	Past one month	Freely available (both paper- and computer-based format)
Intensity of Parental Punishment Scale (IPPS)	Gordon et al., 1979	Physical abuse; Emotional Abuse	School misbehavior; Disobedience after a recent reminder; Public disobedience; Crying; Destructiveness	Parents	To identify maltreating parents/carers; To evaluate effectiveness of an intervention	5 (33)	33–231	7-point ordinal scale (no reaction = 1 to very strong punishment = 7)	Not specified	Freely available (paper-based format)
Mother-Child Neglect Scale (MCNS)	Lounds et al., 2004	Neglect	Emotional neglect; Cognitive neglect; Supervisory neglect; Physical needs neglect	Mothers	To identify maltreating parents/carers	4 (20)	20–80	4-point ordinal scale (strongly disagree = 1 to strongly agree = 4)	Past one year	Freely available (paper-based format)

(continued)

Table 1. *(continued)*

Measure (abbreviation)	Studies on development and validation	Main constructs	(Sub)scales	Target population	Purpose of use	No. of subscales (No. of items)	Range of score	Response Options	Recall period	Cost (mode of administration)
Mother-Child Neglect Scale-Short Form (MCNS-SF)	Lounds et al., 2004	Neglect	Emotional neglect; Cognitive neglect; Supervisory neglect; Physical needs neglect	Mothers	To identify maltreating parents/carers	4 (8)	8–32	4-point ordinal scale (strongly disagree = 1 to strongly agree = 4)	Past one year	Freely available (paper-based format)
Parent-Child Aggression Acceptability Movie Task (P-CAAM)	Rodriguez et al., 2011	Physical abuse	Physical discipline; Physical abuse	Current and prospective parent populations	To identify maltreating parents/carers; To evaluate effectiveness of an intervention	2 (8 video clips: 90 sec each)	0–NR	Clips builds towards 'initial physical contact between caregiver and child'; Respondents should identify that moment and stop video; Delay between actual physical contact and stop video = score (per video)	Not specified	Freely available (computer-based format)
Parent Opinion Questionnaire (POQ)	Azar and Rohrbeck, 1986	Physical abuse; Emotional abuse; Neglect	Self-care; Family responsibility and care of siblings; Help and affection to parents; Leaving children alone; Proper behavior and feelings; Punishment	Parents	To identify maltreating parents/carers	6 (60)	0–60	Dichotomous scale (disagree = 0 or agree = 1)	Not specified	Freely available (paper-based format)
Parental Response to Child Misbehavior questionnaire (PRCM)	Holden and Zambarano, 1992; Vittrup et al., 2006	Physical abuse; Emotional Abuse	Discipline techniques	Parents with young children	To identify maltreating parents/carers; To evaluate effectiveness of an intervention	1 (12)	0–72	6-point ordinal scale (never = 0 to 9 ≥ times per week = 6)	Past one week	Freely available (paper-based format)
Shaken Baby Syndrome awareness assessment-Short Version (SBS-SV)	Russell, 2010	Physical abuse; Emotional abuse; Neglect	Soothing techniques; Discipline techniques; Potential for injury	Parents and caregivers of young children	To evaluate effectiveness of an intervention	3 (36)	36–216	6-point ordinal scale (strongly disagree = 1 to strongly agree = 6)	Not specified	Freely available (paper-based format)

Notes. All information was derived from studies on development and validation of the measures; NR = Not Reported.

Table 2. *Methodological quality assessment on responsiveness of measures: Summary of findings for Step 2 in Figure 1.*

Measures	Overall scale / subscale[a]	Number of analyses at scale level on methodological quality[b]			
		Very good	Adequate	Doubtful	Inadequate
AAPI-2	Overall scale	13	10	16	4
	Inappropriate Expectations subscale	7	5	13	2
	Lack of Empathy subscale	8	6	13	2
	Oppressing Children's Power and Independence subscale	6	4	12	2
	Role Reversal subscale	6	6	13	2
	Value of Corporal Punishment subscale	7	6	11	2
APT	Overall scale	1	0	0	0
CNQ	Overall scale	NR			
CNS-MMS	Overall scale	NR			
CTS-ES	Overall scale	NR			
CTSPC	Overall scale	8	7	1	0
	Physical Assault subscale	6	4	0	0
FM-CA	Overall scale	0	1	0	0
ICAST-Trial	Overall scale	2	1	1	0
	Emotional Abuse subscale	1	1	0	0
	Neglect subscale	2	1	1	0
	Physical Abuse subscale	1	1	0	0
	Sexual Abuse subscale	1	1	0	0
IPPS	Overall scale	NR			
MCNS	Overall scale	1	0	0	0
MCNS-SF	Overall scale	NR			
P-CAAM	Overall scale	NR			
POQ	Overall scale	1	1	0	0
PRCM	Overall scale	1	0	1	0
SBS-SV	Overall scale	NR			

Notes. 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] Subscales were included if data on factor analysis and Cronbach's alpha determined per subscale could be retrieved from the literature, thus confirming the scale's multidimensional structure (Mokkink, Prinsen, et al., 2018).

[b] The methodological quality was rated using the COSMIN checklist (Mokkink, de Vet, et al., 2018): very good, adequate, doubtful, inadequate, and NR (not reported); Detailed rating results on methodological quality of single studies can be founded in Online Supplemental Table S6.

Table 3. *Ratings of single analysis at scale level results on responsiveness: Summary of findings for Step 3.1 in Figure 1.*

Measure	Overall scale / subscale[a]	Number of each rating on single scale analysis results on responsiveness[b]		
		+	-	?
AAPI-2	Overall scale	12	29	2
	Inappropriate Expectations subscale	5	22	0
	Lack of Empathy subscale	13	16	0
	Oppressing Children's Power and Independence subscale	5	19	0
	Role Reversal subscale	8	19	0
	Value of Corporal Punishment subscale	9	17	0
APT	Overall scale	1	0	0
CNQ	Overall scale		NR	
CNS-MMS	Overall scale		NR	
CTS-ES	Overall scale		NR	
CTSPC	Overall scale	5	9	2
	Physical Assault subscale	4	6	0
FM-CA	Overall scale	1	0	0
ICAST-Trial	Overall scale	1	3	0
	Emotional Abuse subscale	0	2	0
	Neglect subscale	0	4	0
	Physical Abuse subscale	2	0	0
	Sexual Abuse subscale	0	2	0
IPPS	Overall scale		NR	
MCNS	Overall scale	0	1	0
MCNS-SF	Overall scale		NR	
P-CAAM	Overall scale		NR	
POQ	Overall scale	2	0	0
PRCM	Overall scale	1	1	0
SBS-SV	Overall scale		NR	

Notes. 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; NR = not reported.

[a] Subscales were included if data on factor analysis and Cronbach's alpha determined per subscale could be retrieved from the literature, thus confirming the scale's multidimensional structure (Mokkink, Prinsen, et al., 2018).

[b] The single analysis at scale level results on responsiveness was rated in Step 3 of Figure 1, using the criteria for good responsiveness (Mokkink, Prinsen, et al., 2018): + = sufficient, - = insufficient, ? = indeterminate (due to less robust psychometric data), and NR = not reported (due to no data on responsiveness); Detailed single analysis at scale level results and ratings on each responsiveness are available in Online Supplemental Table S6.

Table 4. *Overall ratings on pooled study results and quality of evidence on responsiveness per measure: Summary of findings for Step 3.2 and 3.3 in Figure 1.*

Measure	Overall scale / subscale[a]	Overall rating[b]	Quality of evidence[c]
AAPI-2	Overall scale	-	Low
	Inappropriate Expectations subscale	-	Low
	Lack of Empathy subscale	-	Low
	Oppressing Children's Power and Independence subscale	-	Low
	Role Reversal subscale	-	Low
	Value of Corporal Punishment subscale	-	Low
APT	Overall scale	+	Low
CNQ	Overall scale	NR	NR
CNS-MMS	Overall scale	NR	NR
CTS-ES	Overall scale	NR	NR
CTSPC	Overall scale	-	Low
	Physical Assault subscale	-	Low
FM-CA	Overall scale	+	Moderate
ICAST-Trial	Overall scale	-	Low
	Emotional Abuse subscale	-	Low
	Neglect subscale	-	Low
	Physical Abuse subscale	+	High
	Sexual Abuse subscale	-	Moderate
IPPS	Overall scale	NR	NR
MCNS	Overall scale	-	Moderate
MCNS-SF	Overall scale	NR	NR
P-CAAM	Overall scale	NR	NR
POQ	Overall scale	+	Moderate
PRCM	Overall scale	-	Moderate
SBS-SV	Overall scale	NR	NR

Notes. 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] Subscales were included if data on factor analysis and Cronbach's alpha determined per subscale could be retrieved from the literature, thus confirming the scale's multidimensional structure (Mokkink, Prinsen, et al., 2018).

[b] Overall ratings of pooled study results on responsiveness was rated in Step 3.2 of Figure 1, using the criteria for good responsiveness (Mokkink, Prinsen, et al., 2018); + = Sufficient rating, - = Insufficient rating, and NR = not reported (due to no data on responsiveness); If the overall rating of an measure is sufficient, the measure is considered to be sufficiently responsive or sensitive to detect effects of interventions; Detailed pooled results on responsiveness per measure are available in Online Supplemental Table S7.

[c] Level of quality of evidence (i.e., a degree of confidence on overall rating of responsiveness) was graded in Step 3.3 of Figure 1, using the modified GRADE approach for grading the quality of summarized evidence on responsiveness (Mokkink, Prinsen, 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, NR = not reported (due to not reported overall rating of responsiveness); If the evidence quality is very low, we should be concerned about using the overall ratings alone to recommend good measures; Reasons for each grading on quality of evidence are available in Online Supplemental Table S7.

Online Supplemental Materials

Table S1. *Database Search Strategies.*

Database	Search Terms (Free text words)	Number of records
CINAHL	(((Adult Adolescent Parenting Inventory) OR (Adult-Adolescent Parenting Inventory)) AND Time limit 1999-Current) OR (Analog Parenting Task) OR (Child Neglect Questionnaire) OR (Child Neglect Scales) OR (Child Trauma Screen) OR ((Conflict Tactics Scales) AND (child and (parent or parents))) OR ((Family Maltreatment) AND Time limit 2019-Current) OR ((Child Abuse Screening) AND Time limit 2018-Current) OR (Intensity of Parental Punishment Scale) OR ((Mother-Child Neglect Scale) or (Mother Child Neglect Scale)) OR ((Parent-Child Aggression Acceptability Movie) OR (Parent Child Aggression Acceptability Movie)) OR (Parent Opinion Questionnaire) OR (Parental Response to Child Misbehavior) OR (Shaken Baby Syndrome awareness assessment))	195
Embase	*As per CINAHL*	116
ERIC	*As per CINAHL*	50
PsycINFO	*As per CINAHL*	1,031
PubMed[a]	*As per CINAHL*	129
Sociological Abstracts[a]	*As per CINAHL*	63

Notes. All searches performed on the 15th and 16th of January 2020 with an update on the 23rd of March 2021.

[a] Search terms in PubMed and Sociological Abstracts are same as in CINAHL except using double quotation marks before and after name of measures.

Table S2. *Risk of Bias checklist for assessing the methodological quality of studies adapted from the COSMIN manual for systematic reviews of measures (Mokkink et al., 2018).*

Psychometric property	Aspect	Standard[a]	Item description
Responsiveness	Comparison before and after an intervention	Design requirements	Was an adequate description provided of the intervention given?
		Statistical methods	Was the statistical method appropriate for the hypotheses to be tested?
		Other flaws	Were there any other important flaws in the design or statistical methods of the study?

Note. AUC = Area Under the Curve; The Risk of Bias checklist was used for assessing the methodological quality of studies (Step 2 in Figure 1).

[a] Each standard on methodological quality was rated using a four-point rating scale: inadequate, doubtful, adequate, and very good; The overall methodological quality per study was determined calculating a percentage of the ratings (Cordier et al., 2015): inadequate = 0–25%, doubtful = 25.1–50%, adequate = 50.1–75%, and very good = 75.1–100%.

Table S3. *Criteria for good responsiveness adapted from the COSMIN manual for systematic reviews of measures (Mokkink et al., 2018).*

Psychometric property	Aspect	Rating[a]	Quality criteria[b]
Responsiveness	Comparison before and after an intervention	+	Meaningful changes in scores before and after intervention (e.g., Hedges' g ≥0.50)
		?	Not all information for '+' reported (e.g., lack of information to calculate Hedges' g)
		-	Criteria for '+' not met (e.g., Hedges' g < 0.50)
		NR	No information found on responsiveness

Note. AUC = Area Under the Curve; The criteria for good responsiveness was used for rating the results of single studies on responsiveness (Step 3.1 of Figure 1) and rating the pooled results of all studies per measure (Step 3.2 of Figure1).

[a] + = Sufficient, - = Insufficient, ? = Indeterminate, and NR = Not Reported.

[b] The quality criterion for good responsiveness on comparison of change scores before and after intervention was determined as medium effect size (Hedges' g = 0.5) using (Cohen, 1988) conventions to interpret effect size, which was decided by the review team for this current review as suggested by the COSMIN manual (Mokkink et al., 2018).

Table S4. *Modified GRADE approach for grading the quality of evidence on responsiveness per measure adapted from the COSMIN manual for systematic reviews of measures (Mokkink et al., 2018).*

Level of evidence quality (sum of scores per factor)	Factor	Score	Criteria
High (0)	*Risk of bias*	0	Multiple studies of at least adequate methodological quality *OR* One study of very good methodological quality
Moderate (-1)		-1	Multiple studies of doubtful methodological quality *OR* Only one study of adequate methodological quality
Low (-2)		-2	Multiple studies of inadequate methodological quality *OR* Only one study of doubtful methodological quality
Very low (< -3)		-3	Only one study of inadequate methodological quality
	Inconsistency[a]	0	Low heterogeneity in results across studies ($0\% \le I^2 < 50\%$)
		-1	Moderate heterogeneity in results across studies ($50\% \le I^2 < 75\%$)
		-2	High heterogeneity in results across studies ($75\% \le I^2$)
	Imprecision	0	Pooled sample sizes of all individual studies > 100
		-1	Pooled sample sizes of all individual studies = 50–100
		-2	Pooled sample sizes of all individual studies = n < 50
	Indirectness	0	All studies addressing construct or target population of the review
		-1	At least one study not addressing construct or target population of the review, but not all
		-2	All studies not addressing construct or target population of the review

Note. The modified GRADE approach was used for grading the quality of summarized evidence on responsiveness (Step 3.3 of Figure 1); The starting point of evidence quality is 'high' quality of evidence; the level of evidence quality is downgraded by the sum of scores per factor.

[a] The criterion for inconsistency was determined by the review team for this current review as suggested by the COSMIN manual (Mokkink et al., 2018), et al., 2018); The review team decided to evaluate inconsistency or heterogeneity in results across studies using *I-squared (I^2)* statistic that is the percentage of the total variability in a set of effect sizes across the studies due to heterogeneity; Values of less than 50%, 50% to 74%, and higher than 75% denote low, moderate, and high heterogeneity, respectively (Higgins et al., 2003).

Table S5. *Descriptions of included articles on responsiveness of measures for the assessment of child maltreatment.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation^a	Sample size^b	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
AAPI-2	Akai et al. (2008)	To evaluate the effectiveness of an intervention designed to improve early parenting	My Baby and Me	Random	23	Mothers at risk	100	15–38	22.81	5.07
	Alvarez et al. (2018)	To examine the components affecting the quality of the implementation and their impact on the outcomes of a parenting program	Growing Up Happily in the Family	Non-random	133	Parents with children aged 0 to 5 years	90.3	NR	32.85	8.36
	Axford et al. (2020)	To evaluate effectiveness of a therapeutic parenting program for parents of children with behavioural and emotional difficulties	Inspiring Futures	Random	134	Parents of children aged 6 to 11 years with behavioural and emotional difficulties	45.1	NR	NR	NR
	Barden et al. (2015)	To investigate the effectiveness of a relationship education on increasing positive parenting attitudes	Becoming Parents Program	Non-random	140	Economically strained couples with children	50	NR	NR	NR
	Barnes et al. (2017)	To determine the effectiveness and cost-effectiveness of a group-based parenting program in reducing risk factors for maltreatment	Group Family Nurse Partnership	Random	75	Mothers from pregnancy to the first year postpartum	100	NR	21	1.8
	Barnet et al. (2007)	To evaluate the impact of a community-based home-visiting program on poor parenting and other risk factors in pregnant adolescents	Home-Visiting Program	Random	31	Pregnant adolescents aged 12 to 18 years	100	NR	16.4	1.4
	Benzies et al. (2011)	To examine the effects of a two-generation, multi-cultural preschool program on children of Aboriginal heritage and their caregivers	One World	Non-random	23	Caregivers of aboriginal preschool children	NR	NR	30	5.76
	Benzies et al. (2014)	To evaluate a single-site, two-generation preschool demonstration program for low-income families in Canada	Nobody's Perfect; 1-2-3 Magic	Non-random	67	Low-income parents of preschool children	NR	8–46	30.82	6.3
	Berry et al. (2007)	To evaluate the effectiveness of a reunification program in increasing rates of reunification for foster children with their birth parents	Intensive Reunification Program	Non-random	4	Biological parents served a program for reunification with their children in foster care	NR	NR	NR	NR
	Burton et al. (2018)	To evaluate the impact of a parenting program for parents of children with developmental disabilities on nurturing parenting skills	Nurturing Program for Parents and Their Children with Special Needs and Health Challenges	Random	20	Parents of children with development disabilities	97.6	NR	NR	NR
	Clark et al. (2013)	To examine the effect of Love's Cradle relationship enhancement intervention on positive non-abusive parenting attitudes	Love's Cradle	Non-random	69	Low-income pregnant or postpartum (maximum of 3 months post-delivery) adult couples	50	NR	28.60	7.27
	Conn et al. (2018)	To assess the impacts of Incredible Years intervention on child behaviour, foster parent stress and attitudes, and perceived effect on parenting	Incredible Years	Random	16	Foster parents of children aged 2–7 years	81.3	NR	NR	NR

(Continued)

Table S5. *Continued.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation[a]	Sample size[b]	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
AAPI-2	Conners et al. (2006)	To examine the impact of a comprehensive, residential substance abuse treatment program for pregnant and parenting women on substance use, employment, legal involvement, mental health symptoms, risky sexual behaviour, and parenting attitudes	Residential treatment for substance abuse problems	Non-random	200	Pregnant women and mothers with substance abuse	100	NR	29.8	7.2
	Cullen et al. (2010)	To examine the effects of a home visitation program on the parenting attitudes and practices of at-risk parents	Healthy Families America home visitation program	Non-random	55	At-risk parents	100	NR	NR	NR
	Estefan et al. (2013)	To explore the family stressors in parents involved in the child welfare system who have been referred to an intensive therapeutic parenting program, and the relationship of those stressors to change in parenting attitudes	Nurturing Parenting Program	Non-random	94	Parents involved in the child welfare system	52.1	NR	NR	NR
	Farber (2009)	To assess the effects of parent mentoring and guidance programs on changes in parenting and child outcomes	Well-Baby Care; Brazelton Touchpoints Training	Non-random	30	Low-income Latino and African American mothers	100	NR	23	5.6
	Galanter et al. (2012)	To evaluate the effect of a parent–child interaction therapy delivered in-home by community agency therapists on changes in parenting behaviour and attitudes of parents	In-Home Parent–Child Interaction Therapy	Non-random	48	Parents at risk for child maltreatmen	88.0	NR	NR	NR
	Gibbs et al. (2008)	To evaluate the impact of a health camp psychosocial intervention on children with behavioural and emotional problems and the impact of a parenting programme of their parents.	Health camp intervention	Non-random	100	Parents of children with emotional and behaviour problems	89.0	NR	34.6	6.4
	Lavi et al. (2015)	To examine the potential impact of an evidence-based treatment for traumatized mother–child dyads on maternal functioning 6 months post-partum	Child-Parent Psychotherapy	Non-random	64	Pregnant women at risk for intimate partner violence	100	NR	27.5	8.9
	Lawson et al. (2012)	To examine the extent to which participation in a county-wide prevention program leads to improvements in protective factors associated with child abuse prevention	safe families	Non-random	1184	Mothers living in economically and socially vulnerable communities in urban region	100	NR	NR	NR

(Continued)

Table S5. *Continued.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation[a]	Sample size[b]	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
AAPI-2	LeCroy and Judy (2011)	To examine the effectiveness of home visiting on improving parental, child, and maternal outcomes and preventing child abuse and neglect	Healthy Families Program	Random	92	Mothers at risk	100	NR	23.5	NR
	Maher et al. (2011)	To examine the relationship between program dosage and subsequent child maltreatment.	Nurturing Parenting Program	Random	442	Parents of young children (from infant to pre-school) referred to child welfare services	74	12–60	27.38	7.35
	Marcynyszyn et al. (2011)	To exam the effectiveness of an evidence-based parent education program in the context of a child welfare population, as well as implementation challenges and recommendations.	Incredible Years	Non-random	24	Caregivers involved in child welfare agencies	71	NR	36	NR
	McKelvey et al. (2012)	To examine the impact of a home visiting intervention on adolescent mothers’ parenting attitudes	Healthy Families America intervention	Non-random	93	Low income Adolescent mothers	100	NR	17.3	1.4
	Miller et al. (2014)	To assess mothers' needs and interests regarding parenting while they were incarcerated, adapt the program to address those needs, and establish intervention delivery and evaluation methods in collaboration with a community-based agency	Parenting While Incarcerated	Non-random	22	Mothers of children under 18 at the local county jail	100	21–48	31	6.8
	Palusci et al. (2008)	to measure the effects of a formal parenting education program offered to several high-risk groups, including incarcerated and residential substance abuse treatment populations, before maltreatment occurred	Helping Your Child Succeed based on Family Nurturing Program	Non-random	781	Parents enrolled in diverse rehabilitation services to reduce substance abuse, violence, and mental health problems in county jail and community	44	NR	33.2	NR
	Renzaho and Sonia (2011)	To evaluate the impact of a culturally appropriate parenting program to reduce intergenerational conflicts and enhance family cohesion and wellbeing among sub-Saharan African refugees and migrants living in Australia.	African Migrant Parenting Program	Non-random	39	African migrant and refugee parents in Australia	54	19–55	33.4	10.9
	Robbers (2008)	To evaluate the effect of a multifaceted intervention program operating on improving parenting skills of teenage mothers and their male partners	Caring Equation	Non-random	194	Adolescent parents	73	14–20	16.72	2.18
	Rodriguez et al. (2010)	To examine the effectiveness of a home visiting program in promoting parenting competencies and preventing maladaptive parenting behaviours in mothers at risk for child abuse and neglect	Healthy Families New York	Random	255	Mothers at risk for child abuse and neglect	100	NR	22.5	5.8

(Continued)

Table S5. *Continued.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation[a]	Sample size[b]	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
AAPI-2	Sangalang and Kathleen (2005)	To examine the effect of parenting case management program on substance use, contraceptive behaviour, and parenting knowledge	North Carolina's Adolescent Parenting Program	Non-random	91	Pregnant and parenting adolescents	100	12–18	15.9	1.2
	Sawasdipanich et al. (2010)	To examine the effects of a cognitive adjustment program on parental attitudes toward child rearing and the potential for this abuse	Full Love in the Family Protects Your Kids	Random	53	Thai parents of children aged 1 to 6 years	79	NR	NR	NR
	Schilling et al. (2017)	To measure impact of group parent training designed to teach positive parenting skills on child behaviour and parenting attitudes	Child–Adult Relationship Enhancement in Primary Care (PriCARE)	Random	80	Parents of children 2 to 6 years old with behaviour difficulties	95	NR	NR	NR
	Scudder et al. (2014)	To explore the effectiveness of two facility-based group parenting models in enhancing parent-reported and observed parenting outcomes	Parent–Child Interaction Therapy	Random	39	Mothers (of a child aged 2 to 12) incarcerated at a state correctional facility	100	NR	31.31	4.69
	Strickler et al. (2018)	To compare the effect of an enhanced pre-service training developed for treatment parents on their parenting attitudes, personal dedication and willingness to provide foster care, and licensing rates	Pressley Ridge's Treatment Foster Care	Non-random	66	Prospective treatment foster parents	63	NR	48.34	13.00
	Stover et al. (2019)	To evaluate a residential substance misuse treatment program for fathers, integrated treatment for intimate partner violence and child maltreatment	Fathers for Change (F4C)	Random	34	Fathers registered in residential substance use treatment programs	0	23-62	36.82	9.07
	Suess et al. (2016)	To examine the effect of an attachment-based early intervention program on attachment security, parental stress, attitudes, and depression of German mothers	Steps Toward Effective and Enjoyable Parenting	Non-random	54	Young high-risk mothers	100	NR	18.08	NR
	Thomas and Stephen (2004)	To examine the effectiveness of a comprehensive psychoeducational intervention on depression, self-esteem, and parenting attitudes/beliefs of at-risk pregnant and parenting adolescents	Residential Treatment Facility	Non-random	5	Pregnant and parenting Adolescents	100	14-20	16.8	1.3
	Twomey et al. (2010)	To exam the impact of maternal participation in a multidisciplinary therapeutic approach for perinatal substance users on maternal functioning, infant developmental, and permanency outcomes	Family Treatment Drug Court	Non-random	52	Perinatal substance user mothers participating in family treatment drug court	100	19–45	29.2	6.3

(Continued)

Table S5. *Continued.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation[a]	Sample size[b]	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
AAPI-2	Waters et al. (2015)	To investigate maternal sensitivity in a treatment-seeking sample of predominately Latina, low-income pregnant women with histories of interpersonal trauma exposure	Perinatal Child–Parent Psychotherapy (P-CPP)	Non-random	51	Latina, low-income pregnant women	100	18–40	27.08	5.66
	Waterston et al. (2009)	To evaluate the effect of a parenting newsletter, sent monthly to the parents' home from birth to 1 year, on maternal well-being and parenting style	Baby Express	Random	81	First-time mothers	100	NR	29.4	5.8
	Wood et al. (2020)	To measure the impact of a primary care program on disruptive child behaviours, parenting stress, and parenting attitudes	Primary Care (PriCARE)	Random	105	Caregivers of children ages 2 to 6 years with behaviour difficulties	96	NR	NR	NR
	Zajicek-Farber (2010)	To evaluate the impact of an individualized parent mentoring on parenting practices and knowledge of age-appropriate nurturing and emotionally sensitive caregiving	Parent Mentoring Intervention	Non-random	35	Pregnant mothers with high-risk in urban settings	100	NR	23	5.6
	Zolnoski et al. (2012)	To assess the effect of a mixed home visitation parenting program on addressing family need and the risk for child maltreatment	Healthy Families America; Parents as Teachers	Non-random	13	Parents registered in home visiting program	82	21–62	32.5	11.1
APT	Holland and Holden (2016)	To evaluate the efficacy of a motivational interviewing approach in changing positive attitudes toward corporal punishment behavioural intentions, and behaviour	Motivational Interviewing	Random	21	Mothers of children ages 3 to 5	100	22–44	32.37	6.3
CNQ	No study included	NR	NR	NR	NR	NR	NR	NR	NR	NR
CNS-MMS	No study included	NR	NR	NR	NR	NR	NR	NR	NR	NR
CTS-ES	No study included	NR	NR	NR	NR	NR	NR	NR	NR	NR
CTSPC	Dobowitz et al. (2012)	To examine the effectiveness of a paediatric primary care program on reducing child maltreatment	Safe Environment for Every Kid	Random	583	Mothers of children ages 0 to 5 years	100	NR	33.4	5.7
	Feinberg et al. (2016)	To test the short-term efficacy of a brief, universal, transition-to-parenthood intervention	Family Foundations	Random	169	Couples expecting their first child	50	NR	30.10	4.93
	Fowler and Michael (2017)	To test whether permanent housing plus housing case management reduces child maltreatment among families at risk of out-of-home placement compared to housing case management alone	Family Unification Program; Housing Advocacy Program	Random	68	Homeless and child welfare-involved parents	92	NR	32.0	8.5

(Continued)

Table S5. *Continued.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation[a]	Sample size[b]	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
CTSPC	Guterman et al. (2013)	To examine the benefits of home-based paraprofessional parent aide services in reducing physical abuse and neglect risk in high-risk parents	Parent Aide; Case Management	Random	73	High-risk parents	100	NR	29.2	0.9
	Guterman et al. (2018)	To assess the feasibility, acceptability, and preliminary outcomes of an evidence-based perinatal home visitation program	Dads Matter	Non-random	23	Biological parents in vulnerable families	50	NR	22.50	6.29
	Knox and Burkhart (2014)	To examine the factors related to attrition and treatment outcomes in a family violence and child abuse prevention program for parents and caregivers of young children.	ACT-Raising Safe Kids	Non-random	60	Parents and caregivers of young children	75	NR	36.41	8.93
	Lindhiem et al. (2014)	To compare changes in two different assessments (the absolute frequency method and the relative frequency method for quantifying parenting practices) in response to treatment	Parent-Management Training	Non-random	139	Parents of children with disruptive behaviour problems	NR	NR	NR	NR
	McDonell et al. (2015)	To evaluate the effect of a multi-year comprehensive community-based initiative on preventing child maltreatment and improve children's safety.	Strong Communities for Children	Random	229	Parents or caregivers of a child aged 10 or younger	72.5	NR	35.9	8.7
	Ondersma et al. (2017)	To test the effectiveness of a multicomponent computer-based parenting program to prevent child maltreatment	e-Parenting Program; Early home visitation	Random	112	At-risk mothers	100	NR	23.8	4.8
	Oveisi et al. (2010)	To assess whether primary health care settings can be used to engage and provide a preventive intervention to mothers of young children	SOS (helps for parents) Program	Random	108	Iranian mothers of young children (age 2 to 6)	100	NR	29.8	4.49
	Portnoy et al. (2018)	To evaluating the effect of omega-3 supplementation to reducing intimate partner violence and child maltreatment among adult caregivers	Omega-3	Random	94	Caregivers of young children in Maritius	89	NR	38.21	2.99
	Self-brown et al. (2017)	To examine the acceptability and initial efficacy of an augmented version of the evidence-based child maltreatment prevention program, SafeCare, for improving father parenting skills and reducing maltreatment risk	SafeCare® Dad to Kids	Random	50	At-risk fathers	0	NR	30.05	7.75

(Continued)

Table S5. *Continued.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation[a]	Sample size[b]	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
CTSPC	Shaffer et al. (2013)	To examine pre–post treatment changes for a modular intervention that has previously demonstrated significant clinical improvements in child behaviour and maintenance of these effects	Modular Intervention	Random	137	Parents of children ages 6 to11 with disruptive behaviour problems	15	NR	NR	NR
	Swenson et al. (2010)	To evaluate effectiveness of Multisystemic Therapy for Child Abuse and Neglect for physically abused youth and their families	Multisystemic Therapy for Child Abuse and Neglect	Random	43	Parents of physically abused youth	65.9	NR	40.82	11.15
	Wieling et al. (2015)	To assess the feasibility of providing a parenting intervention for war-affected families in Uganda	Enhancing Family Connection	Non-random	14	War-affected mothers in Northern Uganda	100	23–48	33.5	7.0
	Zoysa et al. (2015)	To exam the impact of an awareness raising program to reduce parental use of aversive disciplinary practices	Awareness Raising Program	Non-random	157	Sri Lankan parents	87.6	20–70	39.8	8.86
FM-CA	Slep et al. (2020)	To evaluate the effectiveness of a community-based framework to reduce adult substance misuse, intimate partner violence, child abuse, suicidality, and cumulative risk	NORTH STAR	Random	11,377	Military parents with children at US Air Force base	42	NR	32.61	7.65
ICAST-Trial	Meinick et al. (2018)	To evaluate the adaptation and the psychometric properties of the ISPCAN child abuse screening tool for use in trials (ICAST-Trial) among South African adolescents and their primary caregivers	Parenting for Lifelong Health	Random	240	Primary caregivers of South African adolescents	94.7	NR	49.4	14.69
	Shenderovich et al. (2019)	To examine whether the implementation measures in this study predict participant outcomes on child maltreatment and parenting behaviour	Sinovuyo Teen	Random	270	Caregivers of South African adolescents aged 10-18	97	NR	49	15.2
	Cluver et al. (2018)	To assess the impact of a parenting programme for adolescents in low-income and middle-income countries, on abuse and parenting practices	Sinovuyo Teen	Random	270	Caregivers reporting conflict with their adolescent children (aged 10–18)	97	NR	48.79	
	Lachman et al. (2020)	To evaluate the effectiveness of an intervention combining parenting and economic strengthening programmes to reduce violence against children for caregivers in rural Tanzania	Skillful Parenting & Agribusiness	Random	248	Parents with children aged 0–18 years in farming communities in Tanzania	55	NR	41.65	
IPPS	No study included	NR	NR	NR	NR	NR	NR	NR	NR	NR

(Continued)

Table S5. *Continued.*

Measure	Study	Purpose of study	Name of Intervention	Sample allocation[a]	Sample size[b]	Study population	Percentage of Female	Age		
								Range	Mean	Standard Deviation
MCNS	Gallitto et al. (2020)	To investigate the impact of the SafeCare program on parenting behaviours in child welfare-involved families	SafeCare®	Non-random	68	Child welfare-involved caregivers from Ontario	82.9	NR	28.1	8.9
MCNS-SF	No study included	NR	NR	NR	NR	NR	NR	NR	NR	NR
P-CAAM	No study included	NR	NR	NR	NR	NR	NR	NR	NR	NR
POQ	Sanders et al. (2004)	To exam whether parental attributional retraining and anger management enhance the effects of the Triple P-Positive Parenting Program with parents at risk of child maltreatment	Triple P-Positive Parenting Program; Enhanced group Behavioral Family Intervention	Non-random	35	Parents at risk for child maltreatment	NR	NR	33.33	5.37
	Vorhies et al. (2009)	To evaluate the effectiveness of a residential program with comprehensive wrap-around services for pregnant and parenting foster care youth with severe mental illness or severe emotional disturbance who are preparing to transition to independent living	Thresholds Mothers' Project	Non-random	17	Pregnant and parenting foster care female youth with severe mental illness	100	NR	19.31	1.23
PRCM	Holland et al. (2016)	To evaluate the efficacy of a motivational interviewing approach in changing positive attitudes toward corporal punishment behavioural intentions, and behaviour	Motivational Interviewing	Random	21	Mothers of children ages 3 to 5	100	22–44	32.37	6.3
	Caughy et al. (2003)	To exam the effects of Healthy Steps on discipline strategies of parents of young children	Healthy Steps	Random	134	Parents of children aged 16 to 37 months	NR	NR	NR	NR
SBS-SF	No study included	NR	NR	NR	NR	NR	NR	NR	NR	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; NR = Not Reported.

[a] Random sample allocation indicates that the sample is randomly allocated to an intervention or control group; Non-random sample allocation indicates that the sample is not randomly allocated to an intervention or control group (Altman, 1991).

[b] Sample size is the total number of sample completing the measures both before and after intervention in treatment group.

Table S6. *Single analysis at scale level results and ratings on responsiveness: Detailed findings for Step 3.1 in Figure 1.*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges' g effect size[e] (95% CI)	Rating[f] on result per study
AAPI-2: Overall scale	Akai et al. (2008)	Adequate	*P*-value	Random	23	Mothers	**1.501** (0.417 – 2.584)	+
	Alvarez et al. (2018)	Very good	Effect size	Non-random	133	Parents	**0.303** (0.121 – 0.485)	-
	Axford et al. (2020)	Very good	*P*-value	Random	134	Parents	**-0.205** (-0.375 – -0.034)	-
	Barden et al. (2015)	Doubtful	Effect size	Non-random	140	Couples with children	**0.116** (-0.05 – 0.281)	-
	Barneset al. (2017)	Very good	Effect size	Random	75	Mothers	**0.36** (0.058 – 0.663)	-
	Barnetet al. (2007)	Adequate	*P*-value	Random	31	Pregnant adolescents	**0.492** (0.03 – 0.954)	-
	Benzies et al. (2011)	Doubtful	Effect size	Non-random	23	Caregivers	**0.111** (-0.259 – 0.481)	-
	Benzies et al. (2014)	Very good	*P*-value	Non-random	67	Parents	**0.136** (-0.103 – 0.375)	-
	Berry et al. (2007)	Doubtful	Effect size	Non-random	4	Parents	**NR**	?
	Burton et al. (2018)	Very good	Effect size	Random	20	Parents	**0.226** (-0.219 – 0.671)	-
	Clark et al. (2013)	Doubtful	*P*-value	Non-random	69	Couples with babies	**1.024** (0.718 – 1.329)	+
	Conn et al. (2018)	Doubtful	*P*-value	Random	16	Foster parents	**0.005** (-0.497 – 0.507)	-
	Conners et al. (2006)	Doubtful	*P*-value	Non-random	200	Mothers	**0.187** (0.048 – 0.326)	-
	Cullen et al. (2010)	Doubtful	*P*-value	Non-random	55	Mothers	**1.804** (1.378 – 2.23)	+
	Estefan et al. (2013)	Doubtful	Effect size	Non-random	94	Parents	**1.005** (0.829 – 1.18)	+
	Farber (2009)	Very good	Effect size	Non-random	30	Mothers	**0.774** (0.31 – 1.238)	+
	Galanter et al. (2012)	Adequate	*P*-value	Non-random	48	Parents	**0.476** (0.18 – 0.772)	-
	Gibbs et al. (2008)	Inadequate	Effect size	Non-random	100	Parents	**-0.001** (-0.2 – 0.199)	-
	Lavi et al. (2015)	Very good	*P*-value	Non-random	64	Pregnant women	**0.91** (0.607 – 1.213)	+
	Lawson et al. (2012)	Adequate	*P*-value	Non-random	1184	Mothers	**0.383** (0.324 – 0.442)	-
	LeCroy and Judy (2011)	Doubtful	Effect size	Random	92	Mothers	**-0.35** (-0.672 – -0.027)	-
	Maher et al. (2011)	Very good	*P*-value	Random	442	Parents	**-0.005** (-0.098 – 0.088)	-
	Marcynyszyn et al. (2011)	Doubtful	*P*-value	Non-random	24	Caregivers	**0.275** (-0.13 – 0.679)	-
	McKelvey et al. (2012)	Inadequate	*P*-value	Non-random	93	Adolescent mothers	**0.124** (-0.131 – 0.379)	-
	Miller et al. (2014)	Doubtful	*P*-value	Non-random	22	Mother	**0.162** (-0.243 – 0.568)	-
	Palusci et al. (2008)	Doubtful	*P*-value	Non-random	773	Parents	**NR**	?
	Renzaho and Sonia (2011)	Doubtful	*P*-value	Non-random	39	Parents	**0.732** (0.388 – 1.077)	+
	Robbers (2008)	Doubtful	*P*-value	Non-random	194	Adolescent parents	**0.655** (0.529 – 0.781)	+
	Rodriguez et al. (2010)	Adequate	Effect size	Random	255	Mothers	**0.049** (-0.098 – 0.196)	-
	Sangalang and Kathleen (2005)	Doubtful	*P*-value	Non-random	91	Adolescent parents	**0.297** (0.09 – 0.504)	-
	Sawasdipanich et al. (2010)	Very good	Effect size	Random	53	Parents	**0.539** (0.254 – 0.823)	+
	Schilling et al. (2017)	Adequate	*P*-value	Random	80	Parents	**0.37** (0.144 – 0.596)	-

(Continued)

Table S6. *(Continued).*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges' g effect size[e] (95% CI)	Rating[f] on result per study
AAPI-2: Overall scale	Scudder et al. (2014)	Very good	Effect size	Random	39	Mothers	**0.463** (0.022 – 0.904)	-
	Stover et al. (2019)	Adequate	*P*-value	Non-random	34	Fathers	**0.446** (0.101 – 0.791)	-
	Strickler et al. (2018)	Very good	*P*-value	Random	66	Foster parents	**0.332** (0.084 – 0.579)	-
	Suess et al. (2016)	Inadequate	*P*-value	Non-random	54	Young mothers	**0.42** (0.067 – 0.774)	-
	Thomas and Stephen (2004)	Inadequate	Effect size	Non-random	5	Adolescent parents	**1.135** (0.142 – 2.128)	+
	Twomey et al. (2010)	Doubtful	Effect size	Non-random	52	Mothers	**-0.092** (-0.376 – 0.193)	-
	Waters et al. (2015)	Adequate	*P*-value	Non-random	51	Pregnant women	**0.895** (0.574 – 1.217)	+
	Waterston et al. (2009)	Very good	Effect size	Random	81	First-time mothers	**0.213** (-0.006 – 0.433)	-
	Wood et al. (2020)	Adequate	*P*-value	Random	105	Caregivers	**0.222** (0.028 – 0.415)	-
	Zajicek-Farber (2010)	Very good	Effect size	Non-random	35	Pregnant mothers	**1.411** (1.087 – 1.734)	+
	Zolnoski et al. (2012)	Adequate	*P*-value	Non-random	13	Parents	**0.037** (-0.479 – 0.553)	-
AAPI-2: Inappropriate Expectations subscale	Alvarez et al. (2018)	Very good	Effect size	Non-random	133	Parents	**0.14** (-0.03 – 0.31)	-
	Benzies et al. (2011)	Doubtful	*P*-value	Non-random	23	Caregivers	**0.217** (-0.082 – 0.516)	-
	Benzies et al. (2014)	Very good	Effect size	Non-random	67	Parents	**0.157** (-0.082 – 0.395)	-
	Burton et al. (2018)	Very good	Effect size	Random	20	Parents	**0.384** (-0.053 – 0.821)	-
	Clark et al. (2013)	Doubtful	Effect size	Non-random	69	Couples with babies	**0.126** (-0.108 – 0.36)	-
	Conn et al. (2018)	Doubtful	*P*-value	Random	16	Foster parents	**0.074** (-0.392 – 0.54)	-
	Conners et al. (2006)	Doubtful	*P*-value	Non-random	200	Mothers	**0.284** (0.144 – 0.425)	-
	Cullen et al. (2010)	Doubtful	*P*-value	Non-random	55	Mothers	**1.203** (0.859 – 1.547)	+
	Estefan et al. (2013)	Doubtful	*P*-value	Non-random	94	Parents	**1.145** (0.962 – 1.328)	+
	Galanter et al. (2012)	Adequate	Effect size	Non-random	48	Parents	**0.518** (0.221 – 0.815)	+
	Gibbs et al. (2008)	Inadequate	*P*-value	Non-random	100	Parents	**-0.392** (-0.594 – -0.19)	-
	LeCroy and Judy (2011)	Doubtful	*P*-value	Random	92	Mothers	**0.415** (0.129 – 0.702)	-
	Maher et al. (2011)	Very good	Effect size	Random	442	Parents	**-0.004** (-0.097 – 0.089)	-
	Marcynyszyn et al. (2011)	Doubtful	*P*-value	Non-random	24	Caregivers	**0.282** (-0.113 – 0.677)	-
	McKelvey et al. (2012)	Inadequate	*P*-value	Non-random	93	Adolescent mothers	**0.152** (-0.103 – 0.407)	-
	Miller et al. (2014)	Doubtful	*P*-value	Non-random	22	Mother	**0.405** (-0.015 – 0.825)	-
	Renzaho and Sonia (2011)	Doubtful	*P*-value	Non-random	39	Parents	**0.846** (0.491 – 1.201)	+
	Robbers (2008)	Doubtful	*P*-value	Non-random	194	Adolescent parents	**0.826** (0.711 – 0.941)	+
	Rodriguez et al. (2010)	Adequate	*P*-value	Random	255	Mothers	**0** (-0.147 – 0.147)	-
	Sangalang and Kathleen (2005)	Doubtful	*P*-value	Non-random	91	Adolescent parents	**0.26** (0.054 – 0.466)	-
	Schilling et al. (2017)	Adequate	Effect size	Random	80	Parents	**0.282** (0.061 – 0.504)	-

(Continued)

Table S6. *(Continued).*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges’ g effect size[e] (95% CI)	Rating[f] on result per study
AAPI-2: Inappropriate Expectations subscale	Scudder et al. (2014)	Very good	Effect size	Random	39	Mothers	**0.086** (-0.346 – 0.518)	-
	Strickler et al. (2018)	Very good	*P*-value	Non-random	66	Foster parents	**0.449** (0.199 – 0.7)	-
	Twomey et al. (2010)	Doubtful	Effect size	Non-random	52	Mothers	**-0.445** (-0.726 – -0.164)	-
	Waters et al. (2015)	Adequate	Effect size	Non-random	51	Pregnant women	**0.339** (0.117 – 0.561)	-
	Wood et al. (2020)	Adequate	*P*-value	Random	105	Caregivers	**0.211** (0.019 – 0.403)	-
	Zolnoski et al. (2012)	Adequate	*P*-value	Non-random	13	Parents	**0** (-0.509 – 0.509)	-
AAPI-2: Lack of Empathy subscale	Akai et al. (2008)	Adequate	*P*-value	Random	23	Mothers	**0.971** (0.025 – 1.917)	+
	Alvarez et al. (2018)	Very good	Effect size	Non-random	133	Parents	**1.204** (0.982 – 1.427)	+
	Axford et al. (2020)	Very good	*P*-value	Random	134	Parents	**-0.205** (-0.375 – -0.034)	-
	Barden et al. (2015)	Doubtful	*P*-value	Non-random	140	Couples with children	**0.042** (-0.122 – 0.207)	-
	Benzies et al. (2011)	Doubtful	Effect size	Non-random	23	Caregivers	**-0.079** (-0.437 – 0.278) 3735	-
	Benzies et al. (2014)	Very good	Effect size	Non-random	67	Parents	**0.971** (0.025 – 1.917)	+
	Burton et al. (2018)	Very good	Effect size	Random	20	Parents	**1.205** (0.982 – 1.427)	+
	Clark et al. (2013)	Doubtful	*P*-value	Non-random	69	Couples with babies	**0.043** (-0.122 – 0.207)	-
	Conn et al. (2018)	Doubtful	*P*-value	Random	16	Foster parents	**-0.08** (-0.437 – 0.278)	-
	Conners et al. (2006)	Doubtful	*P*-value	Non-random	200	Mothers	**0.127** (-0.111 – 0.364)	-
	Cullen et al. (2010)	Doubtful	*P*-value	Non-random	55	Mothers	**0.806** (0.317 – 1.296)	+
	Estefan et al. (2013)	Doubtful	Effect size	Non-random	94	Parents	**1.619** (1.262 – 1.976)	+
	Galanter et al. (2012)	Adequate	*P*-value	Non-random	48	Parents	**0.712** (0.185 – 1.238)	+
	Gibbs et al. (2008)	Inadequate	*P*-value	Non-random	100	Parents	**-0.061** (-0.199 – 0.077)	-
	LeCroy and Judy (2011)	Doubtful	Effect size	Random	92	Mothers	**1.492** (1.111 – 1.874)	+
	Maher et al. (2011)	Very good	*P*-value	Random	442	Parents	**0.945** (0.774 – 1.115)	+
	Marcynyszyn et al. (2011)	Doubtful	*P*-value	Non-random	24	Caregivers	**0.171** (-0.109 – 0.452)	-
	McKelvey et al. (2012)	Inadequate	*P*-value	Non-random	93	Adolescent mothers	**0.063** (-0.132 – 0.257)	-
	Miller et al. (2014)	Doubtful	*P*-value	Non-random	22	Mother	**0.343** (0.057 – 0.628)	-
	Renzaho and Sonia (2011)	Doubtful	*P*-value	Non-random	39	Parents	**0.006** (-0.087 – 0.099)	-
	Robbers (2008)	Doubtful	*P*-value	Non-random	194	Adolescent parents	**0.843** (0.388 – 1.297)	+
	Rodriguez et al. (2010)	Adequate	*P*-value	Random	255	Mothers	**0.043** (-0.103 – 0.190)	-
	Sangalang and Kathleen (2005)	Doubtful	*P*-value	Non-random	91	Adolescent parents	**0.506** (0.248 – 0.764)	+
	Schilling et al. (2017)	Adequate	Effect size	Random	80	Parents	**-0.39** (-0.809 – 0.029)	-
	Scudder et al. (2014)	Very good	Effect size	Random	39	Mothers	**0.749** (0.401 – 1.097)	+
	Strickler et al. (2018)	Very good	Effect size	Non-random	66	Foster parents	**0.543** (0.287 – 0.799)	+

(Continued)

Table S6. *(Continued).*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges' g effect size[e] (95% CI)	Rating[f] on result per study
AAPI-2: Lack of Empathy subscale	Waterston et al. (2009)	Very good	Effect size	Random	81	First-time mothers	**-0.548** (-0.654 – -0.441)	-
	Wood et al. (2020)	Adequate	Effect size	Random	105	Caregivers	**0.145** (-0.046 – 0.336)	-
	Zolnoski et al. (2012)	Adequate	*P*-value	Non-random	13	Parents	**0.043** (-0.104 – 0.191)	-
AAPI-2: Oppressing Children's Power and Independence subscale	Alvarez et al. (2018)	Very good	Effect size	Non-random	133	Parents	**-0.205** (-0.375 – -0.034)	-
	Benzies et al. (2011)	Doubtful	*P*-value	Non-random	23	Caregivers	**0.269** (-0.117 – 0.654)	-
	Benzies et al. (2014)	Very good	Effect size	Non-random	67	Parents	**-0.087** (-0.324 – 0.151)	-
	Burton et al. (2018)	Very good	Effect size	Random	20	Parents	**-0.297** (-0.727 – 0.134)	-
	Clark et al. (2013)	Doubtful	Effect size	Non-random	69	Couples with babies	**0.546** (0.296 – 0.797)	+
	Conn et al. (2018)	Doubtful	*P*-value	Random	16	Foster parents	**0.027** (-0.438 – 0.492)	-
	Conners et al. (2006)	Doubtful	*P*-value	Non-random	200	Mothers	**0** (-0.138 – 0.138)	-
	Cullen et al. (2010)	Doubtful	*P*-value	Non-random	55	Mothers	**0.948** (0.633 – 1.264)	+
	Estefan et al. (2013)	Doubtful	*P*-value	Non-random	94	Parents	**0.831** (0.668 – 0.995)	+
	Gibbs et al. (2008)	Inadequate	*P*-value	Non-random	100	Parents	**-0.332** (-0.532 – -0.132)	-
	LeCroy and Judy (2011)	Doubtful	*P*-value	Random	92	Mothers	**-3.323** (-3.761 – -2.885)	-
	Maher et al. (2011)	Very good	Effect size	Random	442	Parents	**0.003** (-0.09 – 0.096)	-
	Marcynyszyn et al. (2011)	Doubtful	*P*-value	Non-random	24	Caregivers	**0.059** (-0.328 – 0.446)	-
	McKelvey et al. (2012)	Inadequate	*P*-value	Non-random	93	Adolescent mothers	**-0.301** (-0.557 – -0.045)	-
	Miller et al. (2014)	Doubtful	*P*-value	Non-random	22	Mother	**-0.201** (-0.609 – 0.206)	-
	Renzaho and Sonia (2011)	Doubtful	*P*-value	Non-random	39	Parents	**0.424** (0.109 – 0.74)	-
	Robbers (2008)	Doubtful	*P*-value	Non-random	194	Adolescent parents	**0.92** (0.795 – 1.045)	+
	Rodriguez et al. (2010)	Adequate	*P*-value	Random	255	Mothers	**0.083** (-0.064 – 0.23)	-
	Schilling et al. (2017)	Adequate	*P*-value	Random	80	Parents	**0.206** (-0.013 – 0.425)	-
	Strickler et al. (2018)	Very good	Effect size	Non-random	66	Foster parents	**0.072** (-0.166 – 0.311)	-
	Twomey et al. (2010)	Doubtful	*P*-value	Non-random	52	Mothers	**-0.443** (-0.724 – -0.162)	-
	Waterston et al. (2009)	Very good	Effect size	Random	81	First-time mothers	**0.131** (-0.085 – 0.348)	-
	Wood et al. (2020)	Adequate	Effect size	Random	105	Caregivers	**0.528** (0.325 – 0.731)	+
	Zolnoski et al. (2012)	Adequate	*P*-value	Non-random	13	Parents	**0.18** (-0.333 – 0.694)	-
AAPI-2: Role Reversal subscale	Akai et al. (2008)	Adequate	*P*-value	Random	23	Mothers	**1.838** (0.67 – 3.005)	+
	Alvarez et al. (2018)	Very good	Effect size	Non-random	133	Parents	**0.25** (0.078 – 0.421)	-
	Barden et al. (2015)	Doubtful	*P*-value	Non-random	140	Couples with children	**0.189** (0.022 – 0.355)	-
	Benzies et al. (2011)	Doubtful	*P*-value	Non-random	23	Caregivers	**0.22** (-0.151 – 0.591)	-
	Benzies et al. (2014)	Very good	Effect size	Non-random	67	Parents	**0.249** (0.006 – 0.491)	-

(Continued)

Table S6. *(Continued).*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges' g effect size[e] (95% CI)	Rating[f] on result per study
AAPI-2: Role Reversal subscale	Burton et al. (2018)	Very good	Effect size	Random	20	Parents	**-0.175** (-0.6 – 0.249)	-
	Clark et al. (2013)	Doubtful	Effect size	Non-random	69	Couples with babies	**1.263** (0.949 – 1.578)	+
	Conn et al. (2018)	Doubtful	*P*-value	Random	16	Foster parents	**-0.966** (-1.539 – -0.393)	-
	Conners et al. (2006)	Doubtful	*P*-value	Non-random	200	Mothers	**0.397** (0.254 – 0.541)	-
	Cullen et al. (2010)	Doubtful	*P*-value	Non-random	55	Mothers	**1.847** (1.415 – 2.28)	+
	Estefan et al. (2013)	Doubtful	*P*-value	Non-random	94	Parents	**0.849** (0.679 – 1.018)	+
	Galanter et al. (2012)	Adequate	Effect size	Non-random	48	Parents	**0.623** (0.318 – 0.928)	+
	Gibbs et al. (2008)	Inadequate	*P*-value	Non-random	100	Parents	**0.425** (0.222 – 0.628)	-
	LeCroy and Judy (2011)	Doubtful	*P*-value	Random	92	Mothers	**0.448** (0.161 – 0.735)	-
	Maher et al. (2011)	Very good	Effect size	Random	442	Parents	**-0.031** (-0.124 – 0.063)	-
	Marcynyszyn et al. (2011)	Doubtful	*P*-value	Non-random	24	Caregivers	**-0.071** (-0.459 – 0.316)	-
	McKelvey et al. (2012)	Inadequate	*P*-value	Non-random	93	Adolescent mothers	**0.172** (-0.083 – 0.427)	-
	Miller et al. (2014)	Doubtful	*P*-value	Non-random	22	Mother	**0.027** (-0.376 – 0.43)	-
	Renzaho and Sonia (2011)	Doubtful	*P*-value	Non-random	39	Parents	**0.759** (0.414 – 1.104)	+
	Robbers (2008)	Doubtful	*P*-value	Non-random	194	Adolescent parents	**0.169** (0.069 – 0.27)	-
	Rodriguez et al. (2010)	Adequate	*P*-value	Random	255	Mothers	**0.024** (-0.124 – 0.171)	-
	Schilling et al. (2017)	Adequate	*P*-value	Random	80	Parents	**0.583** (0.348 – 0.819)	+
	Strickler et al. (2018)	Very good	Effect size	Non-random	66	Foster parents	**0.082** (-0.157 – 0.321)	-
	Twomey et al. (2010)	Doubtful	*P*-value	Non-random	52	Mothers	**0.614** (0.321 – 0.906)	+
	Waterston et al. (2009)	Very good	Effect size	Random	81	First-time mothers	**0.367** (0.144 – 0.286)	-
	Wood et al. (2020)	Adequate	Effect size	Random	105	Caregivers	**0.096** (-0.094 – 0.591)	-
	Zolnoski et al. (2012)	Adequate	*P*-value	Non-random	13	Parents	**0.06** (-0.45 – 0.569)	-
	Burton et al. (2018)	Very good	Effect size	Random	20	Parents	**-0.175** (-0.6 – 0.249)	-
	Clark et al. (2013)	Doubtful	Effect size	Non-random	69	Couples with babies	**1.263** (0.949 – 1.578)	+
	Conn et al. (2018)	Doubtful	*P*-value	Random	16	Foster parents	**-0.966** (-1.539 – -0.393)	-
	Conners et al. (2006)	Doubtful	*P*-value	Non-random	200	Mothers	**0.397** (0.254 – 0.541)	-
	Cullen et al. (2010)	Doubtful	*P*-value	Non-random	55	Mothers	**1.847** (1.415 – 2.28)	+
	Estefan et al. (2013)	Doubtful	*P*-value	Non-random	94	Parents	**0.849** (0.679 – 1.018)	+
	Galanter et al. (2012)	Adequate	Effect size	Non-random	48	Parents	**0.623** (0.318 – 0.928)	+
	Gibbs et al. (2008)	Inadequate	*P*-value	Non-random	100	Parents	**0.425** (0.222 – 0.628)	-
	LeCroy and Judy (2011)	Doubtful	*P*-value	Random	92	Mothers	**0.448** (0.161 – 0.735)	-
	Maher et al. (2011)	Very good	Effect size	Random	442	Parents	**-0.031** (-0.124 – 0.063)	-

(Continued)

Table S6. *(Continued).*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges' g effect size[e] (95% CI)	Rating[f] on result per study
AAPI-2: Value of Corporal Punishment subscale	Akai et al. (2008)	Adequate	*P*-value	Random	23	Mothers	**1.694** (0.569 – 2.818)	+
	Alvarez et al. (2018)	Very good	Effect size	Non-random	133	Parents	**0.125** (-0.044 – 0.295)	-
	Benzies et al. (2011)	Doubtful	*P*-value	Non-random	23	Caregivers	**-0.072** (-0.498 – 0.353)	-
	Benzies et al. (2014)	Very good	Effect size	Non-random	67	Parents	**0.233** (-0.007 – 0.473)	-
	Burton et al. (2018)	Very good	Effect size	Random	20	Parents	**0.411** (-0.029 – 0.851)	-
	Clark et al. (2013)	Doubtful	Effect size	Non-random	69	Couples with babies	**1.564** (1.214 – 1.914)	+
	Conn et al. (2018)	Doubtful	*P*-value	Random	16	Foster parents	**0.18** (-0.289 – 0.649)	-
	Conners et al. (2006)	Doubtful	*P*-value	Non-random	200	Mothers	**-0.014** (-0.152 – 0.124)	-
	Cullen et al. (2010)	Doubtful	*P*-value	Non-random	55	Mothers	**1.3** (0.944 – 1.656)	+
	Estefan et al. (2013)	Doubtful	*P*-value	Non-random	94	Parents	**1.043** (0.867 – 1.219)	+
	Galanter et al. (2012)	Adequate	Effect size	Non-random	48	Parents	**0.591** (0.288 – 0.893)	+
	Gibbs et al. (2008)	Inadequate	*P*-value	Non-random	100	Parents	**0.234** (0.036 – 0.431)	-
	LeCroy and Judy (2011)	Doubtful	*P*-value	Random	92	Mothers	**0.367** (0.082 – 0.653)	-
	Maher et al. (2011)	Very good	Effect size	Random	442	Parents	**0.001** (-0.092 – 0.094)	-
	Marcynyszyn et al. (2011)	Doubtful	*P*-value	Non-random	24	Caregivers	**0.261** (-0.132 – 0.655)	-
	McKelvey et al. (2012)	Inadequate	*P*-value	Non-random	93	Adolescent mothers	**-0.175** (-0.43 – 0.08)	-
	Miller et al. (2014)	Doubtful	*P*-value	Non-random	22	Mother	**0.588** (0.149 – 1.026)	+
	Renzaho and Sonia (2011)	Doubtful	*P*-value	Non-random	39	Parents	**0.846** (0.491 – 1.201)	+
	Robbers (2008)	Doubtful	*P*-value	Non-random	194	Adolescent parents	**1.758** (1.596 – 1.92)	+
	Rodriguez et al. (2010)	Adequate	*P*-value	Random	255	Mothers	**0.094** (-0.053 – 0.241)	-
	Schilling et al. (2017)	Adequate	*P*-value	Random	80	Parents	**0.212** (-0.007 – 0.432)	-
	Scudder et al. (2014)	Very good	Effect size	Random	39	Mothers	**0.365** (-0.07 – 0.8)	-
	Strickler et al. (2018)	Very good	Effect size	Non-random	66	Foster parents	**0.511** (0.257 – 0.765)	+
	Waterston et al. (2009)	Very good	Effect size	Random	81	First-time mothers	**0.306** (0.085 – 0.527)	-
	Wood et al. (2020)	Adequate	Effect size	Random	105	Caregivers	**0.128** (-0.063 – 0.319)	-
	Zolnoski et al. (2012)	Adequate	*P*-value	Non-random	13	Parents	**-0.352** (-0.878 – 0.175)	-
APT: Overall scale	Holland and Holden (2016)	Very good	Effect size	Random	21	Mothers of young children	**1.078** (0.448 – 1.708)	+
CNQ: Overall scale	No study included	NE	NE	NE	NE	NE	**NE**	**NE**
CNS-MMS: Overall scale	No study included	NE	NE	NE	NE	NE	**NE**	**NE**
CTS-ES: Overall scale	No study included	NE	NE	NE	NE	NE	**NE**	**NE**
CTSPC: Overall scale	Dubowitz et al. (2012)	Very good	Effect size	Random	583	Mothers	**0.12** (0 – 0.24)	-
	Feinberg et al. (2016)	Very good	*P*-value	Random	169	Couples expecting their first child	**0.688** (0.469 – 0.908)	+

(Continued)

Table S6. *(Continued).*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges' g effect size[e] (95% CI)	Rating[f] on result per study
CTSPC: Overall scale	Fowler and Michael (2017)	Adequate	Effect size	Random	68	Parents	**NR**	**?**
	Guterman et al. (2013)	Adequate	Effect size	Random	73	Parents	**0.28** (-0.044 – 0.604)	**-**
	Guterman et al. (2018)	Adequate	Effect size	Non-random	23	Parents	**0.229** (-0.156 – 0.614)	**-**
	Knox and Burkhart (2014)	Very good	Effect size	Non-random	60	Parents and caregivers	**0.368** (0.11 – 0.627)	**-**
	Lindhiem et al. (2014)	Very good	Effect size	Non-random	139	Parents	**0.588** (0.348 – 0.828)	**+**
	McDonell et al. (2015)	Very good	*P*-value	Random	229	Parents or caregivers	**NR**	**?**
	Ondersma et al. (2017)	Adequate	Effect size	Random	112	Mothers	**0.026** (-0.149 – 0.201)	**-**
	Oveisi et al. (2010)	Adequate	Effect size	Random	108	Mothers	**0.407** (0.212 – 0.602)	**-**
	Portnoy et al. (2018)	Very good	*P*-value	Random	94	Caregivers	**0.138** (-0.144 – 0.42)	**-**
	Self-brown et al. (2017)	Adequate	Effect size	Random	50	Fathers	**0.777** (0.408 – 1.147)	**+**
	Shaffer et al. (2013)	Very good	Effect size	Random	137	Parents	**0.689** (0.446 – 0.931)	**+**
	Swenson et al. (2010)	Very good	*P*-value	Random	43	Parents	**0.469** (0.044 – 0.894)	**-**
	Wieling et al. (2015)	Doubtful	Effect size	Non-random	14	Mothers	**0.737** (0.173 – 1.301)	**+**
	Zoysa et al. (2015)	Adequate	*P*-value	Non-random	157	Parents	**0.372** (0.149 – 0.594)	**-**
CTSPC: Physical Assault subscale	Dubowitz et al. (2012)	Very good	Effect size	Random	583	Mothers	**0.154** (0.034 – 0.274)	**-**
	Feinberg et al. (2016)	Very good	Effect size	Random	169	Couples expecting their first child	**0.619** (0.401 – 0.836)	**+**
	Guterman et al. (2013)	Adequate	Effect size	Random	73	Parents	**0.302** (-0.022 – 0.627)	**-**
	Guterman et al. (2018)	Adequate	Effect size	Non-random	23	Parents	**0.276** (-0.111 – 0.663)	**-**
	Lindhiem et al. (2014)	Very good	Effect size	Non-random	139	Parents	**0.848** (0.603 – 1.093)	**+**
	Portnoy et al. (2018)	Very good	*P*-value	Random	94	Parents	**0.331** (0.048 – 0.614)	**-**
	Self-brown et al. (2017)	Adequate	Effect size	Random	50	Fathers	**0.31** (0.031 – 0.59)	**-**
	Shaffer et al. (2013)	Very good	Effect size	Random	137	Parents	**0.683** (0.441 – 0.925)	**+**
	Swenson et al. (2010)	Very good	Effect size	Random	43	Parents	**0.565** (0.138 – 0.992)	**+**
	Zoysa et al. (2015)	Adequate	*P*-value	Non-random	157	Parents	**0.349** (0.127 – 0.572)	**-**
FM-CA: Overall scale	Slep et al. (2020)	Adequate	*P*-value	Random	11377	Parents	**0.603** (0.582 – 0.624)	**+**
ICAST-Trial: Overall scale	Cluver et al. (2018)	Very good	*P*-value	Random	270	Caregivers	**0.392** (0.268 – 0.516)	**-**
	Lachman et al. (2020)	Very good	*P*-value	Random	248	Parents	**0.536** (0.442 – 0.63)	**+**
	Meinick et al. (2018)	Adequate	Effect size	Random	240	Primary caregivers	**0.31** (0.181 – 0.44)	**-**
	Shenderovich et al. (2019)	Doubtful	Effect size	Random	270	Caregivers	**0.303** (0.181 – 0.425)	**-**
ICAST-Trial: Emotional Abuse subscale	Lachman et al. (2020)	Very good	*P*-value	Random	248	Parents	**0.485** (0.392 – 0.578)	**-**
	Meinick et al. (2018)	Adequate	Effect size	Random	240	Primary caregivers	**0.32** (0.191 – 0.45)	**-**

(Continued)

Table S6. *(Continued).*

Measure: Overall scale / subscale[a]	Reference	Methodological quality of study[b]	Statistical method of study[c]	Sample allocation[d]	Sample size	Study population	Result of each study Hedges' g effect size[e] (95% CI)	Rating[f] on result per study
ICAST-Trial: Neglect subscale	Cluver et al. (2018)	Very good	Effect size	Random	270	Caregivers	**0.245** (0.124 – 0.366)	-
	Lachman et al. (2020)	Very good	Effect size	Random	248	Parents	**-0.02** (-0.108 – 0.069)	-
	Meinck et al. (2018)	Adequate	*P*-value	Random	240	Primary caregivers	**0.229** (0.101 – 0.357)	-
	Shenderovich et al. (2019)	Doubtful	*P*-value	Random	270	Caregivers	**0.21** (0.09 – 0.331)	-
ICAST-Trial: Physical Abuse subscale	Lachman et al. (2020)	Very good	Effect size	Random	248	Parents	**0.552** (0.458 – 0.647)	+
	Meinck et al. (2018)	Adequate	*P*-value	Random	240	Primary caregivers	**0.512** (0.378 – 0.647)	+
ICAST-Trial: Sexual Abuse subscale	Lachman et al. (2020)	Very good	Effect size	Random	248	Parents	**0.039** (-0.049 – 0.128)	-
	Meinck et al. (2018)	Adequate	*P*-value	Random	240	Primary caregivers	**0.179** (0.052 – 0.306)	-
IPPS: Overall scale	No study included	NE	NE	NE	NE	NE	**NE**	**NE**
MCNS: Overall scale	Gallitto et al. (2020)	Very good	Effect size	Non-random	68	Caregivers	**0.231** (-0.089 – 0.551)	-
MCNS-SF: Overall scale	No study included	NE	NE	NE	NE	NE	**NE**	**NE**
P-CAAM: Overall scale	No study included	NE	NE	NE	NE	NE	**NE**	**NE**
POQ: Overall scale	Sanders et al. (2004)	Very good	Effect size	Non-random	35	Parents	**0.866** (0.484 – 1.248)	+
	Vorhies et al. (2009)	Adequate	*P*-value	Non-random	17	Adolescent mothers	**0.86** (-0.088 – 1.492)	+
PRCM: Overall scale	Holland et al. (2016)	Very good	Effect size	Random	21	Mothers	**0.509** (-0.176 – 1.106)	-
	Caughy et al. (2003)	Doubtful	*P*-value	Random	134	Parents	**0.039** (-0.088 – 0.254)	+
SBS-SV: Overall scale	No study included	NE	NE	NE	NE	NE	**NE**	**NE**

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; NE = Not Evaluated due to no intervention study assessing responsiveness, NR = Not Reported due to no relevant data found to calculate effect size.

[a] Subscales were included if data on factor analysis and Cronbach's alpha determined per subscale could be retrieved from the literature, thus confirming the scale's multidimensional structure (Mokkink et al., 2018).

[b] Methodological quality was evaluated using the Risk of Bias checklist for assessing the methodological quality of studies on responsiveness (Online Supplemental Table S2) in Step 2 of Figure 1.

[c] Statistical method for mean difference before and after intervention was used either to calculate *p*-values or to estimate effect sizes in the included studies. *P*-values were calculated through paired t-tests or repeated measures ANOVAs in most cases; effect size was estimated through calculating standardized mean differences (SMD) such as Cohen's d or Hedges' g (Hedges & Olkin, 2014).

[d] Random sample allocation indicates that the sample is randomly allocated to an intervention or control group; Non-random sample allocation indicates that the sample is not randomly allocated to an intervention or control group (Altman, 1991).

[e] Effect size was calculated using the formulas presented by Borenstein et al. (2009); Hedges' g = a statistic to measure the effect size from change scores between before and after intervention (Hedges & Olkin, 2014), CI = Confidence Interval.

[f] Rating on result of each study was determined using the criteria for good responsiveness (Online Supplemental Table S3) in Step 3.1 of Figure 1; + = Sufficient, ? = Indeterminate, - = Insufficient, ± = Inconsistent.

Table S7. *Pooled results, overall ratings, and quality of evidence on responsiveness per measure: Detailed findings for Step 3.2 and 3.3 in Figure 1.*

Measure	Overall scale / subscale[a]	Quality of evidence[b]				Pooled results	Overall Rating[d] on pooled results	Overall quality of evidence[e] (reasons)
		Risk of bias	Inconsistency	Imprecision	Indirectness	Hedges' g effect size[c] (95% CI; I^2)		
AAPI-2	**Overall scale**	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 90%)	**No concern:** Pooled sample size = 4,430	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.397**(0.287 – 0.506; 90%)	-	**Low** (totally inconsistent results across studies)
	Inappropriate Expectations subscale	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 92.4%)	**No concern:** Pooled sample size = 2,513	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.295** (0.143 – 0.447; 92.4%)	-	**Low** (totally inconsistent results across studies)
	Lack of Empathy subscale	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 95.2%)	**No concern:** Pooled sample size = 2,758	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.392** (0.208 – 0.577; 95.2%)	-	**Low** (totally inconsistent results across studies)
	Oppressing Children's Power and Independence subscale	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 96.2%)	**No concern:** Pooled sample size = 2,335	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.017** (-0.204 – 0.238; 96.2%)	-	**Low** (totally inconsistent results across studies)
	Role Reversal subscale	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 90.1%)	**No concern:** Pooled sample size = 2,546	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.351** (0.216 – 0.486; 90.1%)	-	**Low** (totally inconsistent results across studies)
	Value of Corporal Punishment subscale	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 95.7%)	**No concern:** Pooled sample size = 2,393	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.448** (0.231 – 0.665; 95.7%)	-	**Low** (totally inconsistent results across studies)
APT	**Overall scale**	**No concern**: One study of very good methodological quality	**No concern**: Low heterogeneity in results across studies (I^2 = 0%)	**Very serious concern:** Pooled sample size = 21	**No concern:** All studies addressing target population of this review (caregiver or parent)	**1.078** (0.448 – 1.708; 0%)	+	**Low** (very small total sample size)
CNQ	**Overall scale**	NE	NE	NE	NE	NE	NE	NE
CNS-MMS	**Overall scale**	NE	NE	NE	NE	NE	NE	NE
CTS-ES	**Overall scale**	NE	NE	NE	NE	NE	NE	NE
CTSPC	**Overall scale**	**No concern**: Multiple studies of adequate methodological quality	**Serious concern**: High heterogeneity in results across studies (I^2 = 77.4%)	**No concern:** Pooled sample size = 1,812	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.400** (0.260 – 0.539; 77.4%)	-	**Low** (totally inconsistent results across studies)

(Continued)

Table S7. *(Continued).*

Measure	Overall scale / subscale[a]	Quality of evidence[b]				Pooled results	Overall Rating[d] on pooled results	Overall quality of evidence[e] (reasons)
		Risk of bias	Inconsistency	Imprecision	Indirectness	Hedges' g effect size[c] (95% CI; I^2)		
CTSPC	**Physical Assault subscale**	**No concern**: Multiple studies of adequate methodological quality	**Serious concern**: High heterogeneity in results across studies (I^2 = 77.4%)	**No concern:** Pooled sample size = 885	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.442** (0.277 – 0.607; 77.4%)	-	**Low** (totally inconsistent results across studies)
FM-CA	**Overall scale**	**Serious concern**: Only one study of adequate methodological quality available	**No concern**: Low heterogeneity in results across studies (I^2 = 0%)	**No concern:** Pooled sample size = 11,377	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.603** (0.582 – 0.624; 57.4%)	+	**Moderate** (only one study of adequate methodological quality available)
ICAST-Trial	**Overall scale**	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 75.5%)	**No concern:** Pooled sample size = 1,028	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.390 (**0.273 – 0.508; 75.5%)	-	**Low** (totally inconsistent results across studies)
	Emotional Abuse subscale	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 75.8%)	**No concern:** Pooled sample size = 488	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.409** (0.248 – 0.571; 75.8%)	-	**Low** (totally inconsistent results across studies)
	Neglect subscale	**No concern**: Multiple studies of adequate methodological quality	**Very serious concern**: High heterogeneity in results across studies (I^2 = 83.9%)	**No concern:** Pooled sample size = 1,028	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.162** (0.021 – 0.302; 83.9%)	-	**Low** (totally inconsistent results across studies)
	Physical Abuse subscale	**No concern**: Multiple studies of adequate methodological quality	**No concern**: High heterogeneity in results across studies (I^2 = 0%)	**No concern:** Pooled sample size = 488	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.539** (0.462 – 0.616; 0%)	+	**High** (no concern)
	Sexual Abuse subscale	**No concern**: Multiple studies of adequate methodological quality	**Serious concern**: Moderate heterogeneity in results across studies (I^2 = 68.0%)	**No concern:** Pooled sample size = 488	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.102** (-0.034 – 0.238; 68.0%)	-	**Moderate** (partly inconsistent results across studies)
IPPS	**Overall scale**	NE	NE	NE	NE	NE	NE	NE
MCNS	**Overall scale**	**No concern**: One study of very good methodological quality	**No concern**: Low heterogeneity in results across studies (I^2 = 0%)	**Serious concern:** Pooled sample size = 68	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.231** (-0.089 – 0.551; 0%)	-	**Moderate** (small total sample size)
MCNS-SF	**Overall scale**	NE	NE	NE	NE	NE	NE	NE
P-CAAM	**Overall scale**	NE	NE	NE	NE	NE	NE	NE

(Continued)

Table S7. *(Continued).*

Measure	Overall scale / subscale[a]	Quality of evidence[b]				Pooled results	Overall Rating[d] on pooled results	Overall quality of evidence[e] (reasons)
		Risk of bias	Inconsistency	Imprecision	Indirectness	Hedges' g effect size[c] (95% CI; I^2)		
POQ	**Overall scale**	**No concern**: One study of very good methodological quality	**No concern**: Low heterogeneity in results across studies (I^2 = 0%)	**Serious concern:** Pooled sample size = 52	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.864** (0.537 – 1.191; 0%)	**+**	**Moderate** (small total sample size)
PRCM	**Overall scale**	**No concern**: One study of very good methodological quality	**Serious concern**: Moderate heterogeneity in results across studies (I^2 = 52.7%)	**No concern:** Pooled sample size = 155	**No concern:** All studies addressing target population of this review (caregiver or parent)	**0.188** (-0.241 – 0.618; 52.7%)	**-**	**Moderate** (partly inconsistent results across studies)
SBS-SV	**Overall scale**	NE	NE	NE	NE	NE	NE	NE

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; NE = Not Evaluated due to no intervention study assessing responsiveness.

[a] Subscales were included if data on factor analysis and Cronbach's alpha determined per subscale could be retrieved from the literature, thus confirming the scale's multidimensional structure (Mokkink et al., 2018).

[b] Quality of evidence consists of four factors: risk of bias (methodological quality of the studies: step 2 in Figure 1), inconsistency (inconsistent results across the studies: final pooled results from step 3.2 in Figure 1), Imprecision (small pooled sample size of the studies resulting in wide confidence intervals), and indirectness (evidence from different populations other than the ones of interest in the review).

[c] Effect size was calculated using the formulas presented by Borenstein et al. (2009); Hedges' g = a statistic to measure the effect size from change scores between before and after intervention (Hedges & Olkin, 2014), CI = Confidence Interval, I^2 = *I-squared* as measure of inconsistency (the percentage of total variability across studies due to heterogeneity; Higgins et al., 2003).

[d] Overall rating on pooled result of all studies was determined using the criteria for good responsiveness (Online Supplemental Table S3) in the step 3.2 of Figure 1; + = Sufficient, ? = Indeterminate, - = Insufficient, ± = Inconsistent.

[e] Overall quality of evidence was downgraded using the modified GRADE approach (Online Supplemental Table S4) for grading the quality of summarized evidence on responsiveness (Step 3.3 of Figure 1) when there were concerns regarding each factor on quality of evidence: High = high level of confidence in overall ratings, Moderate = moderate level of confidence in overall ratings, Low = low level of confidence in overall ratings, Very Low = very low level of confidence in overall ratings.

References for Online Supplemental Materials

Akai, C. E., Guttentag, C. L., Baggett, K. M., & Noria, C. C. (2008). Enhancing parenting practices of at-risk mothers. *The Journal of Primary Prevention*, *29*(3), 223-242. https://doi.org/10.1007/s10935-008-0134-z

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

Alvarez, M., Rodrigo, M. J., & Byrne, S. (2018). What implementation components predict positive outcomes in a parenting program?. *Research on Social Work Practice*, *28*(2), 173-187. https://doi.org/10.1177/1049731516640903

Axford, N., Bjornstad, G., Matthews, J., Heilmann, S., Raja, A., Ukoumunne, O. C., Berry, V., Wilkinson, T., Timmons, L., Hobbs, T., Eames, T., Kallitsoglou, A., Blower, S., & Warner, G. (2020). The effectiveness of a therapeutic parenting program for children aged 6-11 years with behavioral or emotional difficulties: Results from a randomized controlled trial. *Children and Youth Services Review, 117*. https://doi.org/10.1016/j.childyouth.2020.105245

Barden, S. M., Carlson, R. G., Daire, A. P., Finnell, L. R., Christopher, K., & Young, E. (2015). Investigating the influence of relationship education on parental attitudes. *Marriage & Family Review*, *51*(3), 246-263. https://doi.org/10.1080/01494929.2015.1031422

Barnes, J., Stuart, J., Allen, E., Petrou, S., Sturgess, J., Barlow, J., Macdonald, G., Spiby, H., Aistrop, D., Melhuish, E., Kim, S. W., & Elbourne, D. (2017). Randomized controlled trial and economic evaluation of nurse-led group support for young mothers during pregnancy and the first year postpartum versus usual care. *Trials*, *18*(1), 508. https://doi.org/10.1186/s13063-017-2259-y

Barnet, B., Liu, J., DeVoe, M., Alperovitz-Bichell, K., & Duggan, A. K. (2007). Home visiting for adolescent mothers: Effects on parenting, maternal life course, and primary care linkage. *Annals of Family Medicine*, *5*(3), 224-232. https://doi.org/10.1370/afm.629

Benzies, K., Mychasiuk, R., Kurilova, J., Tough, S., Edwards, N., & Donnelly, C. (2014). Two-generation preschool programme: Immediate and 7-year-old outcomes for low-income children and their parents. *Child & Family Social Work*, *19*(2), 203-214. https://doi.org/10.1111/j.1365-2206.2012.00894.x

Benzies, K., Tough, S., Edwards, N., Mychasiuk, R., & Donnelly, C. (2011). Aboriginal Children and Their Caregivers Living with Low Income: Outcomes from a Two-Generation Preschool Program. *Journal of Child and Family Studies*, *20*(3), 311-318. https://doi.org/10.1007/s10826-010-9394-3

Berry, M., McCauley, K., & Lansing, T. (2007). Permanency through group work: A pilot intensive reunification program. *Child & Adolescent Social Work Journal*, *24*(5), 477-493. https://doi.org/10.1007/s10560-007-0102-0

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). Introductionto meta-analysis. *West Sussex, England: John Wiley&Sons Ltd*.

Burton, R. S., Zwahr-Castro, J., Magrane, C. L., Hernandez, H., Farley, L. G., & Amodei, N. (2018). The Nurturing Program: An Intervention for Parents of Children with Special Needs. *Journal of Child and Family Studies*, *27*(4), 1137-1149. https://doi.org/10.1007/s10826-017-0966-3

Caughy, M. O. B., Miller, T. L., Genevro, J. L., Huang, K.-Y., & Nautiyal, C. (2003). The effects of Healthy Steps on discipline strategies of parents of young children. *Journal of Applied Developmental Psychology*, *24*(5), 517-534. https://doi.org/10.1016/j.appdev.2003.08.004

Clark, C., Young, M., & Dow, M. G. (2013). Can strengthening parenting couples' relationships reduce at-risk parenting attitudes?. *The Family Journal*, *21*(3), 306-312. https://doi.org/10.1177/1066480713476841

Cluver, L. D., Meinck, F., Steinert, J. I., Shenderovich, Y., Doubt, J., Herrero Romero, R., Lombard, C. J., Redfern, A., Ward, C. L., Tsoanyane, S., Nzima, D., Sibanda, N., Wittesaele, C., De Stone, S., Boyes, M. E., Catanho, R., Lachman, J. M., Salah, N., Nocuza, M., & Gardner, F. (2018). Parenting for Lifelong Health: a pragmatic cluster randomised controlled trial of a non-commercialised parenting programme for adolescents and their families in South Africa. *BMJ Global Health*, *3*(1), e000539. https://doi.org/10.1136/bmjgh-2017-000539

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic Press.

Conn, A.-M., Szilagyi, M. A., Alpert-Gillis, L., Webster-Stratton, C., Manly, J. T., Goldstein, N., & Jee, S. H. (2018). Pilot randomized controlled trial of foster parent training: A mixed-methods evaluation of parent and child outcomes. *Children and Youth Services Review*, *89*, 188-197. https://doi.org/10.1016/j.childyouth.2018.04.035

Conners, N. A., Grant, A., Crone, C. C., & Whiteside-Mansell, L. (2006). Substance abuse treatment for mothers: Treatment outcomes and the impact of length of stay. *Journal of Substance Abuse Treatment*, *31*(4), 447-456. https://doi.org/10.1016/j.jsat.2006.06.001

Cordier, R., Speyer, R., Chen, Y. W., 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*(7), e0132299. https://doi.org/10.1371/journal.pone.0132299

Cullen, J. P., Ownbey, J. B., & Ownbey, M. A. (2010). The effects of the Healthy Families America home visitation program on parenting attitudes and practices and child social and emotional competence. *Child & Adolescent Social Work Journal*, *27*(5), 335-354. https://doi.org/10.1007/s10560-010-0206-9

Dubowitz, H., Lane, W. G., Semiatin, J. N., & Magder, L. S. (2012). The SEEK model of pediatric primary care: can child maltreatment be prevented in a low-risk population? *Academic Pediatrics, 12*(4), 259-268. https://doi.org/10.1016/j.acap.2012.03.005

Estefan, L. F., Coulter, M. L., VandeWeerd, C. L., Armstrong, M., & Gorski, P. (2013). Relationships between stressors and parenting attitudes in a child welfare parenting program. *Journal of Child and Family Studies*, *22*(2), 199-208. https://doi.org/10.1007/s10826-012-9569-1

Farber, M. L. (2009). Parent mentoring and child anticipatory guidance with Latino and African American families. *Health & Social Work*, *34*(3), 179-189. https://doi.org/10.1093/hsw/34.3.179

Feinberg, M. E., Jones, D. E., Hostetler, M. L., Roettger, M. E., Paul, I. M., & Ehrenthal, D. B. (2016). Couple-focused prevention at the transition to parenthood, a randomized trial: Effects on coparenting, parenting, family violence, and parent and child adjustment. *Prevention Science*, *17*(6), 751-764. https://doi.org/10.1007/s11121-016-0674-z

Fowler, P. J., & Schoeny, M. (2017). Permanent housing for child welfare-involved families: Impact on child maltreatment overview. *American Journal of Community Psychology*, *60*(1-2), 91-102. https://doi.org/10.1002/ajcp.12146

Galanter, R., Self-Brown, S., Valente, J. R., Dorsey, S., Whitaker, D. J., Bertuglia-Haley, M., & Prieto, M. (2012). Effectiveness of parent-child interaction therapy delivered to at-risk families in the home setting. *Child & Family Behavior Therapy*, *34*(3), 177-196. https://doi.org/10.1080/07317107.2012.707079

Gallitto, E., Romano, E., & Whitaker, D. (2020). Investigating the Impact of the SafeCare Program on Parenting Behaviours in Child Welfare-Involved Families. *Child and Adolescent Social Work Journal*. https://doi.org/10.1007/s10560-020-00672-6

Gibbs, A., Moor, S., Frampton, C., & Watkins, W. (2008). Impact of psychosocial interventions on children with disruptive and emotional disorders treated in a health camp. *Australian & New Zealand Journal of Psychiatry*, *42*(9), 789-799. https://doi.org/10.1080/00048670802277248

Guterman, N. B., Bellamy, J. L., & Banman, A. (2018). Promoting father involvement in early home visiting services for vulnerable families: Findings from a pilot study of "Dads Matter". *Child Abuse & Neglect*, *76*, 261-272. https://doi.org/10.1016/j.chiabu.2017.10.017

Guterman, N. B., Tabone, J. K., Bryan, G. M., Taylor, C. A., Napoleon-Hanger, C., & Banman, A. (2013). Examining the effectiveness of home-based parent aide services to reduce risk for physical child abuse and neglect: Six-month findings from a randomized clinical trial. *Child Abuse & Neglect*, *37*(8), 566-577. https://doi.org/10.1016/j.chiabu.2013.03.006

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

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

Holland, G. W., & Holden, G. W. (2016). Changing orientations to corporal punishment: A randomized, control trial of the efficacy of a motivational approach to psycho-education. *Psychology of Violence*, *6*(2), 233-242. https://doi.org/10.1037/a0039606

Knox, M., & Burkhart, K. (2014). A multi-site study of the ACT Raising Safe Kids program: Predictors of outcomes and attrition. *Children and Youth Services Review*, *39*, 20-24. https://doi.org/10.1016/j.childyouth.2014.01.006

Lachman, J., Wamoyi, J., Spreckelsen, T., Wight, D., Maganga, J., & Gardner, F. (2020). Combining parenting and economic strengthening programmes to reduce violence against children: a cluster randomised controlled trial with predominantly male caregivers in rural Tanzania. *BMJ Glob Health*, *5*(7). https://doi.org/10.1136/bmjgh-2020-002349

Lavi, I., Gard, A. M., Hagan, M., Van Horn, P., & Lieberman, A. F. (2015). Child-Parent Psychotherapy examined in a perinatal sample: Depression, posttraumatic stress symptoms and child-rearing attitudes. *Journal of Social and Clinical Psychology*, *34*(1), 64-82. https://doi.org/10.1521/jscp.2015.34.1.64

Lawson, M. A., Alameda-Lawson, T., & Byrnes, E. C. (2012). A multilevel evaluation of a comprehensive child abuse prevention program. *Research on Social Work Practice*, *22*(5), 553-566. https://doi.org/10.1177/1049731512444165

LeCroy, C. W., & Krysik, J. (2011). Randomized trial of the healthy families Arizona home visiting program. *Children and Youth Services Review*, *33*(10), 1761-1766. https://doi.org/10.1016/j.childyouth.2011.04.036

Lindhiem, O., Shaffer, A., & Kolko, D. J. (2014). Quantifying discipline practices using absolute versus relative frequencies: Clinical and research implications for child welfare. *Journal of Interpersonal Violence*, *29*(1), 66-81. https://doi.org/10.1177/0886260513504650

Maher, E. J., Marcynyszyn, L. A., Corwin, T. W., & Hodnett, R. (2011). Dosage matters: The relationship between participation in the Nurturing Parenting Program for infants, toddlers, and preschoolers and subsequent child maltreatment. *Children and*

Youth Services Review, *33*(8), 1426-1434. https://doi.org/10.1016/j.childyouth.2011.04.014

Marcynyszyn, L. A., Maher, E. J., & Corwin, T. W. (2011). Getting with the (evidence-based) program: An evaluation of the Incredible Years Parenting Training Program in child welfare. *Children and Youth Services Review*, *33*(5), 747-757. https://doi.org/10.1016/j.childyouth.2010.11.021

McDonell, J. R., Ben-Arieh, A., & Melton, G. B. (2015). Strong Communities for Children: Results of a multi-year community-based initiative to protect children from harm. *Child Abuse & Neglect*, *41*, 79-96. https://doi.org/10.1016/j.chiabu.2014.11.016

McKelvey, L. M., Burrow, N. A., Balamurugan, A., Whiteside-Mansell, L., & Plummer, P. (2012). Effects of home visiting on adolescent mothers' parenting attitudes. *American Journal of Public Health*, *102*(10), 1860-1862. https://doi.org/10.2105/AJPH.2012.300934

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. https://doi.org/10.1016/j.chiabu.2018.05.022

Miller, A. L., Weston, L. E., Perryman, J., Horwitz, T., Franzen, S., & Cochran, S. (2014). Parenting while incarcerated: Tailoring the Strengthening Families Program for use with jailed mothers. *Children and Youth Services Review*, *44*, 163-170. https://doi.org/10.1016/j.childyouth.2014.06.013

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs)-User manual (version 1.0)*.

https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

Ondersma, S. J., Martin, J., Fortson, B., Whitaker, D. J., Self-Brown, S., Beatty, J., Loree, A., Bard, D., & Chaffin, M. (2017). Technology to augment early home visitation for child maltreatment prevention: A pragmatic randomized trial. *Child Maltreatment*, *22*(4), 334-343. https://doi.org/10.1177/1077559517729890

Oveisi, S., Ardabili, H. E., Dadds, M. R., Majdzadeh, R., Mohammadkhani, P., Rad, J. A., & Shahrivar, Z. (2010). Primary prevention of parent-child conflict and abuse in Iranian mothers: a randomized-controlled trial. *Child Abuse Negl*, *34*(3), 206-213. https://doi.org/10.1016/j.chiabu.2009.05.008

Palusci, V. J., Crum, P., Bliss, R., & Bavolek, S. J. (2008). Changes in parenting attitudes and knowledge among inmates and other at-risk populations after a family nurturing program. *Children and Youth Services Review*, *30*(1), 79-89. https://doi.org/10.1016/j.childyouth.2007.06.006

Portnoy, J., Raine, A., Liu, J., & Hibbeln, J. R. (2018). Reductions of intimate partner violence resulting from supplementing children with omega-3 fatty acids: A randomized, double-blind, placebo-controlled, stratified, parallel-group trial. *Aggressive Behavior*, *44*(5), 491-500. https://doi.org/10.1002/ab.21769

Renzaho, A. M. N., & Vignjevic, S. (2011). The impact of a parenting intervention in Australia among migrants and refugees from Liberia, Sierra Leone, Congo, and Burundi: Results from the African Migrant Parenting Program. *Journal of Family Studies*, *17*(1), 71-79. https://doi.org/10.5172/jfs.2011.17.1.71

Robbers, M. L. (2008). The caring equation: An intervention program for teenage mothers and their male partners. *Children & Schools*, *30*(1), 37-47. https://doi.org/10.1093/cs/30.1.37

232

Rodriguez, M. L., Dumont, K., Mitchell-Herzfeld, S. D., Walden, N. J., & Greene, R. (2010). Effects of Healthy Families New York on the promotion of maternal parenting competencies and the prevention of harsh parenting. *Child Abuse & Neglect*, *34*(10), 711-723. https://doi.org/10.1016/j.chiabu.2010.03.004

Sanders, M. R., Pidgeon, A. M., Gravestock, F., Connors, M. D., Brown, S., & Young, R. W. (2004). Does parental attributional retraining and anger management enhance the effects of the triple P-positive parenting program with parents at risk of child maltreatment? *Behavior Therapy*, *35*(3), 513-535. https://doi.org/10.1016/S0005-7894(04)80030-3

Sangalang, B. B., & Rounds, K. (2005). Differences in health behaviors and parenting knowledge between pregnant adolescents and parenting adolescents. *Social Work in Health Care*, *42*(2), 1-22. https://doi.org/10.1300/J010v42n02_01

Sawasdipanich, N., Srisuphan, W., Yenbut, J., Tiansawad, S., & Humphreys, J. C. (2010). Effects of a cognitive adjustment program for Thai parents. *Nursing & Health Sciences*, *12*(3), 306-313. https://doi.org/10.1111/j.1442-2018.2010.00531.x

Schilling, S., French, B., Berkowitz, S. J., Dougherty, S. L., Scribano, P. V., & Wood, J. N. (2017). Child-Adult Relationship Enhancement in Primary Care (PriCARE): A Randomized Trial of a Parent Training for Child Behavior Problems. *Academic Pediatrics*, *17*(1), 53-60. https://doi.org/10.1016/j.acap.2016.06.009

Scudder, A. T., McNeil, C. B., Chengappa, K., & Costello, A. H. (2014). Evaluation of an existing parenting class within a women's state correctional facility and a parenting class modeled from Parent-Child Interaction Therapy. *Children and Youth Services Review*, *46*, 238-247. https://doi.org/10.1016/j.childyouth.2014.08.015

Self-brown, S., Osborne, M. C., Lai, B. S., De Veauuse Brown, N., Glasheen, T. L., & Adams, M. C. (2017). Initial Findings from a Feasibility Trial Examining the

SafeCare Dad to Kids Program with Marginalized Fathers. *Journal of Family Violence*, *32*(8), 751-766. https://doi.org/10.1007/s10896-017-9940-5

Shaffer, A., Lindhiem, O., & Kolko, D. J. (2013). Treatment effects of a modular intervention for early-onset child behavior problems on family contextual outcomes. *Journal of Emotional and Behavioral Disorders*, *21*(4), 277-288. https://doi.org/10.1177/1063426612462742

Shenderovich, Y., Eisner, M., Cluver, L., Doubt, J., Berezin, M., Majokweni, S., & Murray, A. L. (2019). Delivering a parenting program in South Africa: The impact of implementation on outcomes. *Journal of Child and Family Studies*, *28*(4), 1005-1017. https://doi.org/10.1007/s10826-018-01319-y

Slep, A. M. S., Heyman, R. E., Lorber, M. F., Baucom, K. J. W., & Linkh, D. J. (2020). Evaluating the Effectiveness of NORTH STAR: a Community-Based Framework to Reduce Adult Substance Misuse, Intimate Partner Violence, Child Abuse, Suicidality, and Cumulative Risk. *Prevention Science*, *21*(7), 949-959. https://doi.org/10.1007/s11121-020-01156-w

Stover, C. S., McMahon, T. J., & Moore, K. (2019). A randomized pilot trial of two parenting interventions for fathers in residential substance use disorder treatment. *Journal of Substance Abuse Treatment*, *104*, 116-127. https://doi.org/10.1016/j.jsat.2019.07.003

Strickler, A., Trunzo, A. C., & Kaelin, M. S. (2018). Treatment foster care pre-service trainings: Changes in parenting attitudes and fostering readiness. *Child & Youth Care Forum*, *47*(1), 61-79. https://doi.org/10.1007/s10566-017-9418-x

Suess, G., Bohlen, U., Carlson, E., Spangler, G., & Frumentia Maier, M. (2016). Effectiveness of attachment based STEEPTM intervention in a German high-risk sample. *Attachment & Human Development*, *18*(5), 443-460. https://doi.org/10.1080/14616734.2016.1165265

Swenson, C. C., Schaeffer, C. M., Henggeler, S. W., Faldowski, R., & Mayhew, A. M. (2010). Multisystemic therapy for child abuse and neglect: A randomized effectiveness trial. *Journal of Family Psychology*, *24*(4), 497-507. https://doi.org/10.1037/a0020324

Thomas, D. V., & Looney, S. W. (2004). Effectiveness of a Comprehensive Psychoeducational Intervention With Pregnant and Parenting Adolescents: A Pilot Study. *Journal of Child and Adolescent Psychiatric Nursing*, *17*(2), 66-77. https://doi.org/10.1111/j.1744-6171.2004.00066.x

Twomey, J. E., Miller-Loncar, C., Hinckley, M., & Lester, B. M. (2010). After family treatment drug court: Maternal, infant, and permanency outcomes. *Child Welfare: Journal of Policy, Practice, and Program*, *89*(6), 23-41. https://pubmed.ncbi.nlm.nih.gov/21877562/

Vorhies, V., Glover, C. M., Davis, K., Hardin, T., Krzyzanowski, A., Harris, M., Fagan, M., & Wilkniss, S. (2009). Improving outcomes for pregnant and parenting foster care youth with severe mental illness: An evaluation of a transitional living program. *Psychiatric Rehabilitation Journal*, *33*(2), 115-124. https://doi.org/10.2975/33.2.2009.115.124

Waters, S. F., Hagan, M. J., Rivera, L., & Lieberman, A. F. (2015). Improvements in the child-rearing attitudes of Latina mothers exposed to interpersonal trauma predict greater maternal sensitivity toward their 6-month-old infants. *Journal of Traumatic Stress*, *28*(5), 426-433. https://doi.org/10.1002/jts.22043

Waterston, T., Welsh, B., Keane, B., Cook, M., Hammal, D., Parker, L., & McConachie, H. (2009). Improving early relationships: A randomized, controlled trial of an age-paced parenting newsletter. *Pediatrics*, *123*(1), 241-247. https://doi.org/10.1542/peds.2007-1872

Wieling, E., Mehus, C., Mollerherm, J., Neuner, F., Achan, L., & Catani, C. (2015). Assessing the feasibility of providing a parenting intervention for war-affected families in Northern Uganda. *Family & Community Health: The Journal of Health Promotion & Maintenance*, *38*(3), 252-267. https://doi.org/10.1097/FCH.0000000000000064

Wood, J. N., Kratchman, D., Scribano, P. V., Berkowitz, S., & Schilling, S. (2020). Improving Child Behaviors and Parental Stress: A Randomized Trial of Child Adult Relationship Enhancement in Primary Care. *Academic Pediatrics*. https://doi.org/10.1016/j.acap.2020.08.002

Zajicek-Farber, M. L. (2010). Building practice evidence for parent mentoring home visiting in early childhood. *Research on Social Work Practice*, *20*(1), 46-64. https://doi.org/10.1177/1049731509333172

Zolnoski, S., Stacks, A. M., Kohl-Hanlon, A., & Dykehouse, T. A. (2012). Lessons learned from the first-year evaluation of a small-scale home visitation program. *Journal of Social Service Research*, *38*(4), 515-528. https://doi.org/10.1080/01488376.2012.699407

Zoysa, P., Siriwardhana, C., Samaranayake, M., Athukorala, S., Kumari, S., & Fernando, D. (2015). The Impact of an Awareness Raising Program to Reduce Parental use of Aversive Disciplinary Practices. *Journal of Family Violence*, *30*(5), 651-659. https://doi.org/10.1007/s10896-015-9701-2