Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments

van der Put, C.E.; Assink, M.; Boekhout van Solinge, N.F.

DOI
10.1016/j.chiabu.2017.09.016

Publication date
2017

Document Version
Final published version

Published in
Child Abuse & Neglect

Child Abuse & Neglect 73 (2017) 71–88

Contents lists available at ScienceDirect

Child Abuse & Neglect

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

Full length article

Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments

Claudia E. van der Put*, Mark Assink, Noëlle F. Boekhout van Solinge

Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Risk assessment
Child maltreatment
Predictive validity
Meta-analysis
Child abuse
Neglect

ABSTRACT

Risk assessment is crucial in preventing child maltreatment since it can identify high-risk cases in need of child protection intervention. Despite widespread use of risk assessment instruments in child welfare, it is unknown how well these instruments predict maltreatment and what instrument characteristics are associated with higher levels of predictive validity. Therefore, a multilevel meta-analysis was conducted to examine the predictive accuracy of (characteristics of) risk assessment instruments. A literature search yielded 30 independent studies (N = 87,329) examining the predictive validity of 27 different risk assessment instruments. From these studies, 67 effect sizes could be extracted. Overall, a medium significant effect was found (AUC = 0.681), indicating a moderate predictive accuracy. Moderator analyses revealed that onset of maltreatment can be better predicted than recurrence of maltreatment, which is a promising finding for early detection and prevention of child maltreatment. In addition, actuarial instruments were found to outperform clinical instruments. To bring risk and needs assessment in child welfare to a higher level, actuarial instruments should be further developed and strengthened by distinguishing risk assessment from needs assessment and by integrating risk assessment with case management.

1. Introduction

Child maltreatment is a widespread phenomenon affecting the lives of millions of children all over the world (Stoltenborgh, Bakermans-Kranenburg, Alink, & IJzendoorn, 2015). In case of (suspected) child maltreatment, child welfare staff are asked to make extremely difficult decisions about whether, and how best, to intervene so that a child's welfare is safeguarded (Arad-Davidson & Benbenishty, 2008; DePanfilis & Girvin, 2005; Munro, 1999; Pfister & Böhm, 2008). Identifying risks of maltreatment is of paramount importance in these decisions. In recent years, there has been a shift from using mainly unstructured clinical risk assessment to the widespread use of standardized risk assessment instruments (Munro, 2004; Tatara, 1996). Despite this shift, the development and evaluation of risk assessment instruments in the field of child protection is in its infancy. Risk assessment instruments are frequently implemented without proper empirical evaluation, and thus limited knowledge is available about their validity and effectiveness (Barlow, Fisher, & Jones, 2012; Knoke & Trocmé, 2005). Moreover, the child protection field is currently engaged in an intense debate about the most effective approach to assessing risks. However, the average performance of (different approaches to) risk assessment instruments is unknown, because meta-analyses evaluating the predictive accuracy of these instruments have not yet been performed in the child protection field. Therefore, the aim of the current study was to examine the overall

* Corresponding author at: Research Institute of Child Development and Education, University of Amsterdam, P.O. Box 15780, 1001 NG, Amsterdam, The Netherlands.
E-mail address: NoteC.E.vanderPut@UvA.nl (C.E. van der Put).

http://dx.doi.org/10.1016/j.chiabu.2017.09.016
Received 27 February 2017; Received in revised form 6 July 2017; Accepted 11 September 2017
Available online 23 September 2017
0145-2134/

predictive validity of risk assessment instruments for child maltreatment and to examine whether the overall predictive validity is influenced by study and instrument characteristics.

1.1. Approaches to risk assessment

Currently, there are two main approaches to risk assessment in child welfare: the clinical and the actuarial (statistical) approach. In the actuarial approach, conclusions are based solely on empirically established relationships between risk factors and child maltreatment, whereas in the clinical approach, conclusions are based on the judgment of a professional who combines and weighs information in a subjective manner (Dawes, Faust, & Meehl, 1989). Clinical approaches can be further divided into consensus-based instruments and structured clinical judgment (SCJ) instruments. With consensus-based instruments, clinical professionals rate characteristics that are deemed relevant because of consensus among experts. Next, the professionals process these ratings in a subjective manner and come to a conclusion using their own judgment. Structured clinical judgment is a more recently developed method in which variables identified as risk factors in empirical research are assessed, but in which the weighting of risk factors as well as coming to the final decision is left to the professional. Several validation studies indicate that many implemented instruments perform questionably, especially instruments that are based on the clinical approach to risk assessment (see for example, Barlow, Fisher, & Jones, 2012; D'Andrade, Austin, & Benton, 2008; Knoke & Trocmé, 2005). Some studies have even shown that clinical methods, which are widely used in practice, do not perform better than chance, meaning that in half of the cases an incorrect risk estimate is made (Baird & Wagner, 2000; Barber, Shlonsky, Black, Goodman, & Trocmé, 2008; Van der Put, Assink, & Stams, 2016b). This leads to many inappropriate clinical decisions, resulting in unjustified out-of-home placements or recurrence of maltreatment. Therefore, it is essential to gain insight into which types of instruments perform well and which instrumental characteristics influence the predictive validity either positively or negatively.

The development of risk assessment instruments in the field of child welfare lags behind other disciplines, such as the field of criminal (youth) justice. In criminal justice, the literature identifies four generations of risk assessment instruments (Andrews & Bonta, 2010). Clinical instruments are considered the first generation of instruments and actuarial instruments the second generation. Third generation actuarial instruments have been developed incorporating dynamic as well as static risk factors, so that risk assessment can be distinguished from needs assessment. The newest, fourth generation actuarial risk assessment instruments serve not only as a guide for the professional in determining appropriate goals for intervention, but also as a guide in case management planning by offering the possibility of linking re-assessments to the initial assessment, service plans, and service delivery (Andrews & Bonta, 2010). Instruments used in child welfare can be classified into either the first or the second generation of instruments. In most of these instruments, risk assessment is not discriminated from needs assessment. Moreover, the needs assessment instruments that are available have mainly been developed on the basis of expert consensus and have not been subjected to sound empirical validation (Schwalbe, 2008).

1.2. Results from previous review studies

As mentioned, there is an intense debate about which risk assessment approach is most effective in assessing the risk of child maltreatment, also referred to as the "risk assessment wars" (Johnson, 2006a; Johnson, 2006b; Morton, 2003; White & Wash, 2006). Earlier review studies on the predictive validity of risk assessment instruments for child maltreatment showed mixed results. D'andrade et al. (2008) summarized findings of research on seven risk assessment instruments and concluded that actuarial instruments appear to have greater predictive validity and inter-rater reliability than consensus-based instruments. Barlow et al. (2012) conducted a systematic review on the accuracy of risk assessment instruments for child maltreatment and identified 13 different tools. These authors concluded that there is currently limited evidence for the effectiveness of risk assessment instruments in the field of child protection. However, there is evidence supporting the use of one specific actuarial tool, the California Family Risk Assessment, particularly at referral or during initial assessment (Barlow et al., 2012). Bartelink, Van Yperen, and Ten Berge (2015) conducted a review of studies in which a comparison was made between the predictive accuracy of a) different risk assessment instruments or b) a risk assessment instrument and unstructured clinical judgment (i.e., not using an instrument at all). Based on this review, the authors concluded that: (a) actuarial instruments performed slightly better than consensus-based instruments, and that (b) the predictive validity of actuarial instruments did not outperform unstructured clinical judgment. However, the review of Bartelink and colleagues has been criticized by Van der Put, Assink, and Stams (2016a) because their decision to exclude articles reporting on the performance of individual instruments seems too restrictive. After all, studies comparing the predictive accuracy of at least two instruments for risk assessment using the same populations and outcome criteria are hardly available, as are studies in which the performance of a risk assessment instrument is compared to unstructured clinical judgment.

1.3. Research aims

Until today, only qualitative reviews have examined the predictive accuracy of risk assessment instruments used in child protection. Because these reviews lack *meta*-analysis of quantitative data, it is not yet known how these instruments perform on average. Furthermore, some primary studies report very low predictive accuracies (see, for instance, Barber et al., 2008; Ondersma, Chaffin, Mullins, & LeBreton, 2005), whereas others report far better predictive accuracies (see, for instance, Loman & Siegel, 2004; De Ruiter, Hildebrand, & Van der Hoorn, 2012). Given this rather wide range, synthesizing data in a quantitative manner is essential to get insight in the overall predictive accuracy of risk assessment instruments. A second merit of a quantitative review is that it can reveal

variables (such as instrument characteristics) that increase or decrease the overall accuracy, and thus act as moderators. Identifying moderators yields important knowledge that can be used in developing and/or improving risk assessment instruments. Therefore, the aim of the present study was to conduct a meta-analysis, in which we estimate the average predictive accuracy and identify variables that may influence this accuracy, such as approach to and focus of risk assessment. We believe that such a meta-analysis contributes to improving decision-making strategies in child welfare, and thus more effective child protection practices.

The following instrument characteristics were examined: type of risk assessment approach (actuarial, consensus-based, structured clinical judgment), length of instrument (number of items), type of assessor (professional, client (i.e., self-report), researcher, or computer system (i.e., automatic risk calculation based on variables stored in a computer database)), focus of risk assessment (recurrence of child maltreatment, onset of maltreatment, both/not specified) and related to focus. In addition, the following study design characteristics were examined: study design (retrospective versus prospective design), type of sample (clinical or non-clinical sample), sample used for validation (validation versus construction sample), length of follow-up (in months) and type of follow-up (number of months after assessment, number of months after case closure, both/not specified), type of outcome measure (for which the categories were derived from outcomes reported in primary studies), type of maltreatment (multiple forms, physical abuse, neglect, maltreatment not specified), publication year, sample size, and percentage of cultural minorities in the sample. Below, we elaborate on the rationale for testing these specific characteristics.

1.3.1. Type of risk assessment approach

We expected actuarial methods to outperform clinical methods (both consensus-based and SCJ instruments) for two reasons. First, the mathematical features of actuarial methods ensure not only that solely variables with predictive value are part of the instrument, but also that these variables are weighted in accordance with their independent contribution to the outcome of interest (Dawes et al., 1989). Earlier studies showed that it is difficult for professionals to accurately predict an outcome of interest using their clinical judgment, because professionals are unable to focus on the most important factors nor to properly weigh the observed risk factors (Dawes, 1994; Dawes et al., 1989). Second, the reliability of actuarial instruments is higher than that of clinical methods and hence the actuarial prediction is more consistent and accurate (e.g., Dawes et al., 1989; Gambrill & Shlonsky, 2000). That is because risk factors in actuarial prediction are scored according to a fixed algorithm, meaning that professionals use the same objective scoring rules, regardless of the expertise of the professional. On the other hand, scoring risk factors in clinical methods is done subjectively (e.g., Dawes et al., 1989; Gambrill & Shonsky, 2000). Further, we expected SCJ instruments to outperform consensus-based instruments, because a sound empirical basis is lacking for the latter, whereas the former is partly based on empirical evidence.

1.3.2. Length of instrument

The number of items a risk assessment instrument is comprised of was examined because the predictive validity may vary with the length of the instrument. Schwalbe (2007) conducted a meta-analysis on juvenile justice risk assessment instruments and found that brief instruments yielded smaller effect sizes than other types of instruments. In line with this result, we expected to find a negatively moderating effect of the number of items risk assessment instruments comprise of, as briefer instruments may be less capable of assessing all relevant risk factors than instruments of longer length. After all, both juvenile delinquency (Loeber, Slot, & Stouthaer-Loeber, 2008) and child maltreatment (Belsky, 1993) are determined by the presence and absence of multiple and varying risk and protective factors in children and different environmental systems around children.

1.3.3. Type of assessor

Predictive validity may vary depending on the type of assessor (by a professional, by self-reporting, by a researcher, or automatic risk calculation based on variables stored in a computer database). We exploratively examined whether there was an effect of assessor type on predictieve validity, because no clear moderating evidence was found in previous studies.

1.3.4. Focus of risk assessment

Two types of risk assessment instruments can be distinguished: 1) instruments *screening* for maltreatment in the general population (onset of maltreatment); and 2) instruments *assessing the risk of recurrence* of maltreatment in populations already investigated by child protection services. The predictive validity may vary depending on the focus of an instrument, since the populations assessed, their risk of maltreatment, and (effects of) risk factors within populations may differ (Cash, 2001). Screening aims to assess the risk of child maltreatment in the general population in which the risk of child maltreatment is relatively small, whereas risk assessment aims to assess the risk of (repeated) child maltreatment in high-risk groups, such as families involved in child protection services. In scientific literature, there is particular emphasis on instruments assessing the risk of recurrence of child maltreatment, whereas screening instruments for assessing the risk of child maltreatment in the general population get far less attention (Barlow et al., 2012). The reason is that assessing the risk of recurrence of child maltreatment is the most commonly employed prognostic process in child welfare services. However, screenings instruments can be of great value for early prevention of child maltreatment. Related to this, we also tested whether estimates of predictive validity obtained in clinical samples differ from estimated obtained in non-clinical samples.

1.3.5. Study design

Whether a study has a prospective or retrospective design may influence predictive validity. Some researchers have argued that risk assessment instruments can be examined retrospectively, using file information from sources such as institutional files, psychological reports, and/or court reports (e.g., De Vogel, De Ruiter, Hildebrand, Bos, & Van de Ven, 2004). On the contrary, other

researchers have argued that prospective research is required to adequately examine the predictive validity of a risk assessment tool (Caldwell, Bogat, & Davidson, 1988). Therefore, we examined the effect of study design on predictive validity.

1.3.6. Sample used for validation

In some studies, the predictive validity of an instrument is examined in the same sample that was used to construct the instrument, whereas in other studies, the predictive validity is examined in a sample independent of the construction sample. We expected the predictive validity to be lower in validation samples than in construction samples, because random sampling error arising from testing an instrument in a sample that differs from a construction sample, results in reduced predictive validity estimates. In fact, models built in a construction (or training) sample tend to "overfit" the data (i.e., capitalizing on random variation). Thus, predictive validity estimates reported for construction samples are commonly inflated.

1.3.7. Length and type of follow-up

The potential moderating effect of the follow-up length was examined, because the predictive validity may vary over time and differences in follow-up length are frequently observed between studies. As studies also use different types of follow-up (assessing the time after assessment, the time after case closure, or both/not specified), we also examined follow-up type as a potential moderator.

1.3.8. Type of outcome measure

Studies on the predictive validity of risk assessment instruments vary in the outcome that is predicted. We examined whether the predictive validity of instruments is influenced by type of outcome (new reports, investigations, substantiated maltreatment, supervision orders, out-of-home placements, recidivism/relapse), and type of abuse that is predicted (physical abuse, neglect, sexual abuse, and child abuse in general).

A number of additional variables were exploratively tested as potential moderating variables of the predictive validity of risk assessment instruments for child maltreatment. These variables were: the type of maltreatment assessed in primary validity studies, the publication year of primary studies, the size of the samples used in primary studies, and the percentage of cultural minorities in samples of primary studies.

In summary, despite the widespread use of risk assessment instruments in child welfare, it is unclear how well these instruments generally perform and whether the predictive validity is influenced by study and instrument characteristics. This knowledge is not only scientifically important, but also for clinical practice, as it provides guidance on implementing the most effective risk assessment tools. Consequently, this review may contribute to decreasing the number of inappropriate decisions in child protection, resulting in less unjustified out-of-home placements and les recurrences of maltreatment. A three-level random-effects meta-analysis was performed to estimate the overall predictive validity of risk assessment instruments for child maltreatment and to identify variables that moderate this predictive validity.

2. Method

2.1. Review protocol

The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Statement (Moher, Liberati, Tetzlaff, & Altman, 2009) was followed in the present meta-analysis.

2.2. Sample of studies

For selecting relevant studies, several criteria were formulated. First, we selected studies that examined the predictive validity of risk assessment instruments that were specifically developed for the prediction of one or more forms of child maltreatment (physical abuse, sexual abuse, and neglect) in the (near) future. We excluded studies examining instruments that predict more general parenting problems (such as attachment problems between mother and child), since protecting children from different forms of maltreatment and neglect is the primary and most urgent task of child welfare systems. Therefore, child welfare agencies have widely implemented risk assessment instruments for maltreatment, so that practitioners can best identify at-risk children that are the most in need of care directed on reducing this risk. Instruments predicting general and less severe parenting problems are scarcely used for this purpose, and therefore not within the scope of this review.

We also did not search for studies describing instruments that can only be used for assessing the immediate child safety, since these instruments serve a different purpose than risk assessment. Second, studies examining the predictive validity of risk assessment instruments administered to clinical and/or general populations were included. Third, both prospective (longitudinal) and retrospective studies were included. Fourth, studies had to report either an actual effect size of the predictive validity of an instrument (e.g., an Area Under the receiver operating characteristics Curve (AUC) value, a correlation (r), or Cohen's d), or sufficient statistical information for manually calculating an effect size. Fifth, studies had to be written in English or Dutch. Finally, both published and unpublished studies (also known as "gray literature" such as doctoral dissertations, Master's theses, conference presentations, and government reports) were considered for inclusion.

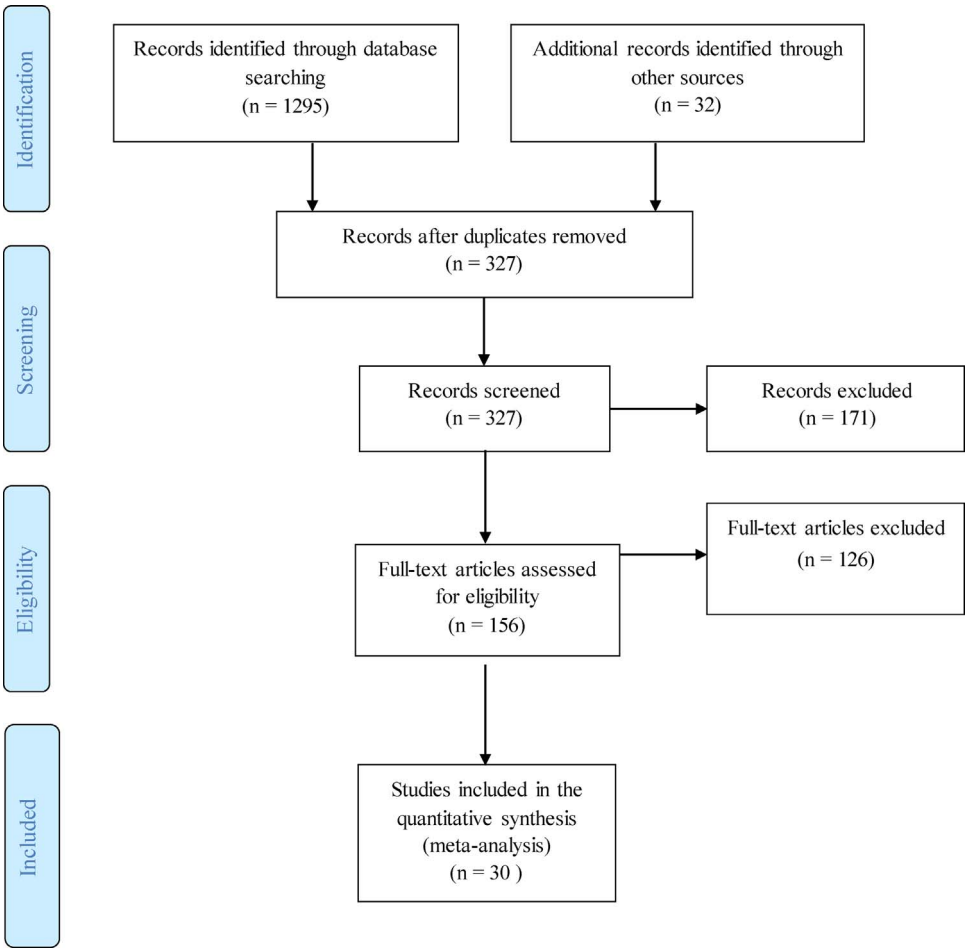

Fig. 1. Flow chart of the search procedure.

2.3. Search strategy

An electronic search was conducted in the databases PubMed, ERIC, Medline, Sociological Abstracts, and Google Scholar. The following combination of four syntax components was used in this search method: ("child abuse" OR "child maltreatment" OR "abus*" OR "maltreat*" OR "physical abuse" OR "sex* abuse" OR "psychological abuse" OR "neglect" OR "harm" OR "child protect*") AND ("child*" OR "infant*" OR "baby" OR "babies" OR "toddler*" OR "teen*" OR "adolesc*" OR "minor") AND ("risk assessment" OR "risk tool" OR "risk measure" OR "risk evaluat*" OR "risk analys*" OR "screen*") AND ("AUC" OR "ROC" OR "sensitivity" OR "specificity" OR "predictive validity" OR "predictive accuracy"). In this syntax, an asterisk (*) represents a wildcard character. To assess the retrieved studies against the inclusion criteria of the present meta-analysis, we read titles, abstracts and, if necessary, full article texts.

Additional studies were searched for by screening reference lists of reviews on risk assessment instruments for child maltreatment (i.e., Barlow et al., 2012; Bartelink, De Kwaadsteniet, Ten Berge, Witteman, & Van Gastel, 2015; D'andrade et al., 2008; Doueck, English, DePanfilis, & Moote, 1992; Gershater-Molko, Lutzker, & Sherman, 2003; Johnson et al., 2008; Knoke & Trocmé, 2005; Lyons, Doueck, & Wodarski, 1996; Pecora, 1991; Peters & Barlow, 2003; Stowman & Donohue, 2005; Walker & Davies, 2010). Further, we screened for potential relevant studies in the reference list of each primary study that was eligible for inclusion in the present meta-analysis. Finally, we contacted multiple scholars who either conducted validation research in the field of child welfare, and/or developed instruments for assessing the risk for child maltreatment.

We searched for studies until September 2016 and the above described search procedure yielded 327 studies. After thoroughly screening these studies, we could include 30 studies that met our inclusion criteria. A flow chart of the search procedure is presented in Fig. 1.

2.4. Assessment of bias

A common phenomenon in meta-analytic research referred to as the "file drawer problem" (Rosenthal, 1979) is that non-

significant findings are less likely to be published than significant findings (Rosenthal, 1979), and this is also referred to as publication bias. The presence of other forms of bias, such as selection or reporting bias, may also influence results of meta-analytic research. To determine whether results of the present meta-analysis were affected by the presence of (different forms of) bias, we conducted the trim-and-fill method (Duval & Tweedie, 2000a, 2000b) by using the function "trimfill" of the metafor package (Viechtbauer, 2010) in the R environment (Version 3.2.0; R Core Team, 2015). This method is built on the assumption that effect sizes are symmetrically distributed (in the form of a funnel) around the "true" effect size, if bias in the results is absent. In case of asymmetry in the funnel plot, the trim-and-fill method restores symmetry by imputing effect sizes that are derived from the effect sizes present in the data set. An "adjusted" overall effect can then be estimated using the data set to which the imputed effect sizes have been added. Of the available techniques for evaluating missing data and its implications for the results in meta-analysis, the trim and fill method is a conceptually easy method to adjust for the impact of missing effect sizes (Nakagawa & Santos, 2012).

2.5. Coding of studies

The guidelines proposed by Lipsey and Wilson (2001) were followed in developing a

- Conversion(M.E.)

coding form. The following instrument characteristics were examined: risk assessment approach (actuarial, consensus-based, structured clinical judgment), number of items, type of assessor (a professional, self-report, a researcher, or automatic risk calculation based on variables stored in a computer database), focus of risk assessment (onset of child maltreatment, recurrence of child maltreatment, both/not specified) and related to focus, type of sample (clinical or non-clinical sample). In addition, the following study design characteristics were examined: study design (prospective or retrospective design), sample used for validation (validation or construction), length of follow up (in months), type of follow-up (time after assessment, time after case closure or both/not specified), type of outcome measure (maltreatment substantiated, number of new reports, number of investigations, relapse, supervision order, out of home placement), and type of abuse predicted (multiple forms, physical abuse, neglect, sexual abuse).

Primarily for descriptive purposes, the following general aspects of included studies were coded: sample size, publication year, publication status, percentage cultural minority in the sample, name of the instrument that was examined, and the country in which the research was conducted.

2.6. Calculation of effect sizes

In the present meta-analysis, we chose the AUC value as effect size, since it is the most common performance indicator for the predictive validity of risk assessment instruments. According to some experts, it is even the preferred measure of predictive accuracy (e.g., Swets, Dawes, & Monahan, 2000). The AUC value is a global and base rate resistant index of discriminatory accuracy of a risk assessment instrument (or, more general, a statistical predictive model) (see also Altman & Bland, 1994Singh, 2013). In the context of risk assessment of child maltreatment, this value represents the probability that a randomly selected child who was exposed to (forms of) maltreatment was assigned a higher risk classification than a randomly selected child who was not exposed to (forms of) maltreatment. More simply stated, the AUC value is an index of how well a risk assessment instrument discriminates between maltreated and non-maltreated children across all possible cut-off scores of the instrument. AUC values range from 0.500 indicating a discriminative accuracy not better than random, to 1.000 indicating a perfect discrimination. AUC values between 0.556 and 0.639 correspond with a small effect size, AUC values between 0.639 and 0.714 with a medium effect size, and AUC values of.714 and higher correspond with a large effect size (Rice & Harris, 2005).

Important to note is that AUC values are not informative on the calibration performance of risk assessment instruments, which is a different − but important − aspect of predictive validity. For inferences about how well risk predictions agree with observed risks, other performance indicators need to be calculated (see, for instance, Singh, 2013 for an overview of calibration and discrimination performance indicators). In this meta-analysis, we only assessed the discriminative accuracy of instruments assessing the risk for child maltreatment.

For calculating AUC values several methods were used. In transforming Cohen's *d* values into AUC values, the formulas of Ruscio (2008) were applied. Converting Pearson's correlations into Cohen's *d* values was done using formulas as given by Rosenthal (1994). If primary studies only reported on sensitivity (i.e., the proportion of maltreated children who were classified as high risk) and specificity (i.e., the proportion of non-maltreated children who were classified as low risk), we calculated an AUC value using the formula [sensitivity + specificity]/2. In this formula, it is assumed that an instrument has only one cut-off threshold. In case of multiple cut-off thresholds, we used the formula to calculate an AUC value for each threshold, after which we selected the highest AUC value as the effect size that was to be included in the meta-analysis. If, instead of sensitivity and specificity, a 2 × 2 contingency table was given on true and false positives as well as true and false negatives (see Fig. 2), we used the information in this table to calculate sensitivity and specificity. We calculated sensitivity as [[number of true positives]/[number of true positives + number of false negatives]] and specificity as [[number of true negatives]/[number of true negatives + number of false positives]] (Singh, 2013).

In the study of Barber et al. (2008), the authors reported that the predictive validity of the Ontario Risk Assessment Tool was not significant, without providing sufficient statistical information to calculate the actual AUC value. Therefore, for this particular study, we chose to include an AUC value of 0.500 representing random discriminative accuracy. Although this value is most probably an

		Outcome: Positive	Outcome: Negative
Test Result	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Fig. 2. 2 × 2 contingency table comparing risk assessment tool predictions and outcomes.

underestimation of the true discriminative accuracy, we preferred this method above excluding the study because of insufficient information for calculating an effect size (see also Mullen (1989) for this procedure).

2.7. Interrater agreement

All included studies were double coded by the first and second author and any disagreement on the coding of a variable was resolved through discussion between the two authors. In some instances the last author was involved in the discussion to arrive at a single coding of a variable. When the coding ended, there were no discrepancies between the two authors, and therefore, there was a perfect interrater reliability of all variables.

2.8. Data analyses

After all AUC values were calculated, we first converted them into Pearson's correlations using the formulas of Ruscio (2008) and Rosenthal (1994). This was needed because an effect size in which the value zero is indicative for a random discriminative accuracy (i.e., "no" effect) was required for properly fitting the multilevel meta-analytic models in R (see further details below). Next, the correlations were transformed into Fisher's *z* values, because correlations are not normally distributed (Mullen, 1989). After the statistical analyses were conducted, the Fisher's *z*-scores were converted back into correlations for ease of interpretation.

Prior to the analyses, we checked for outliers. Since extreme effect sizes may have a disproportionate influence on inferences derived from statistical analyses, we checked for outliers by searching for effect sizes with standardized scores larger than 3.29 or smaller than −3.29 (Tabachnik & Fidell, 2013). One outlier was detected with a Fisher's *z* value of 0.936 (corresponding to an AUC value of 0.937). To reduce the impact of this outlier, this *z* value was substituted by a new *z* value equaling the highest possible effect falling within the normal range. In this way, a disproportionate influence of this particular outlying discriminatory accuracy on the overall effect and (possible) moderating effects was reduced.

In the present study we aimed for maximal statistical power in the analyses as well as minimal information loss, so that (overall) effects could be estimated as accurate as possible. Consequently, we extracted all possible effect sizes from each included primary study, implying that from most primary studies more than one effect size was extracted (see Assink & Wibbelink, 2016 for this procedure). However, a key assumption in traditional meta-analytic approaches is that included effect sizes are independent, so including multiple effect sizes based on the same sample violates this assumption (Lipsey & Wilson, 2001). Following scholars reporting on recent meta-analyses (e.g., Assink, Van der Put, Hoeve, De Vries, Stams, & Oort, 2015; Houben, Van den Noortgate, & Kuppens, 2015; Kuppens, Laurent, Heyvaert, & Onghena, 2013; Rapp, Van den Noortgate, Broekaert, & Vanderplasschen, 2014; Weisz et al., 2013), a multilevel random effects model was used for the calculation of combined effect sizes and for the moderator analyses in order to deal with dependency of effect sizes (Hox, 2002; Van den Noortgate and Onghena, 2003). Van den Noortgate and Onghena (2003) compared this multilevel approach to traditional meta-analytical methods, and concluded that the (maximum likelihood) multilevel approach is in general superior to the fixed-effects approaches used in traditional meta-analysis, and that for models without moderators, the results of the multilevel approach are not substantially different from results of the traditional random-effects approaches.

In the present study, a three-level meta-analytic model was used to analyze all effect sizes, in which three sources of variance were modeled: sampling variance of the observed effect sizes (Level 1), variance between effect sizes extracted from the same primary study (Level 2), and variance between primary studies (Level 3) (see also Cheung, 2014; Houben et al., 2015; Van den Noortgate, López-López, Marin-Martinez, & Sánchez-Meca, 2013,2015). By building this model without covariates (i.e., an intercept-only model), an overall effect can be estimated that is represented by the intercept. Subsequently, in case of significant variation between effect sizes extracted from the same primary study and/or between primary studies, the intercept-only model can be extended with covariates (i.e., potential moderating variables) to determine whether level-2 and level-3 variance can be explained by characteristics of studies, samples, and/or instruments.

The statistical analyses were conducted using the function "rma.mv" of the metafor package (Viechtbauer, 2010) in the R environment (version 3.2.0; R Core Team, 2015). We used the R syntax as given by Assink and Wibbelink (2016), so that the three sources of variance as described by for instance Van den Noortgate et al. (2013, 2015) were modeled. We applied the Knapp and Hartung (2003) adjustment in estimating coefficients of the multilevel meta-analytic model, meaning that the *t*-distribution (instead of the *z*-distribution) was used for testing individual regression coefficients and for calculating the corresponding confidence intervals. Further, when models were extended with categorical moderators comprising three or more categories, the omnibus test of the null hypothesis that all group mean effect sizes are equal, followed an *F*-distribution.

To determine whether there was significant variance between effect sizes extracted from the same primary study (Level 2), and

Table 1
Included Studies and their Characteristics.

Author(s)/ Pub. year	*N*	Name of instrument	Type of instrument	Maltreatment type	Start of follow-up	Outcome	AUC
Altemeier et al. (1984)	1400	Maternal History Interview (MHI)	SCJ	Physical	After assessment	Substantiated	0.5849
	1400	Maternal History Interview 2 (MHI 2)	Actuarial	Physical	After assessment	Substantiated	0.7620
Assink,Van der Put, Hoeve et al. (2015)	1651	Y-ACNAT-NO	Actuarial	General	After assessment	Supervision order	0.7700
Ayoub and Milner (1985)	42	Child Abuse Potential Inventory (CAPI)	Actuarial	Neglect	After case closure	Substantiated	0.6100
Baird and Wagner (2000)	929	Michigan Family Risk Assessment	Actuarial	Multiple forms	After assessment	Substantiated	0.6000
				Multiple forms	After assessment	Investigation	0.5763
	908	Washington Risk Assessment Matrix (WRAM)	Consensus	Multiple forms	After assessment	Substantiated	0.5331
				Multiple forms	After assessment	Investigation	0.5437
	876	California Family Assessment Factor Analysis	Consensus	Multiple forms	After assessment	Substantiated	0.5272
				Multiple forms	After assessment	Investigation	0.5565
Barber et al. (2008)	1118	Ontario Risk Assessment Tool	Consensus	Multiple forms	After case closure	Substantiated	0.5000
Bartelink, Van Yperen et al. (2015)	278	Check List of Child Safety (CLCS)	SCJ	General	After assessment	New reports	0.6542
				General	After assessment	Supervision order	0.6895
				General	After assessment	Out of home placement	0.6021
Brayden et al. (1993)	527	Maternal History Interview 2 (MHI 2)	Actuarial	Physical	After assessment	New reports	0.6470
				Neglect	After assessment	New reports	0.5385
Camasso and Jagannathan (1995)	239	Washington Risk Assessment Matrix (WRAM)	Consensus	Multiple forms	After assessment	Substantiated	0.6800
				Multiple forms	After assessment	Recidivism/relapse	0.6800
	239	CANTS 17B	Consensus	Multiple forms	After assessment	Substantiated	0.6600
				Multiple forms	After assessment	Recidivism/relapse	0.5800
Chaffin and Valle (2003)	459	Child Abuse Potential Inventory (CAPI)	Actuarial	Multiple forms	After assessment	New reports	0.6610
				Multiple forms	After case closure	New reports	0.6460
Coohey, Johnson, K., Renner, and Easton (2013)	6832	Colorado Family Risk Assessment Abuse scale	Actuarial	Multiple forms	Not specified	Substantiated	0.6400
				Multiple forms	Not specified	Substantiated	0.6000
		Colorado Family Risk Assessment Revised − Abuse scale	Actuarial	Multiple forms	Not specified	Substantiated	0.6600
				Multiple forms	Not specified	Substantiated	0.6800
Dankert and Johnson (2014)	11444	California Family Risk Assessment (CFRA)	Actuarial	Physical	After assessment	Substantiated	0.6000
				Physical	After assessment	Investigation	0.5900
				Neglect	After assessment	Substantiated	0.6100
				Neglect	After assessment	Investigation	0.6200
				Multiple forms	After assessment	Substantiated	0.6000
				Multiple forms	After assessment	Investigation	0.5900
De Ruiter et al. (2012)	211	Child Abuse Risk Evaluation (CARE-NL)	SCJ	Multiple forms	After case closure	Supervision order	0.7300
				Multiple forms	After case closure	Out of home placement	0.7700
Flaherty (2001)	5612	Regression model	Actuarial	Physical	After assessment	Substantiated	0.6500
		Neural network model	Actuarial	Physical	After assessment	Substantiated	0.6440
Hamilton and Browne (1999)	400	Screening Checklist for Risk of Referral	Actuarial	Multiple forms	After case closure	New reports	0.7465
Horikawa et al. (2016)	716	No name given	Actuarial	Multiple forms	After assessment	Substantiated	0.6600
Hunter, Kilstrom, Kraybill, and Loda (1978)	255	Family Psychosocial Risk Inventory	Actuarial	Multiple forms	After assessment	Substantiated	0.9365
Johnson (2011)	6543	California Family Risk Assessment (CFRA)	Actuarial	Multiple forms	After case closure	Substantiated	0.6300
	114	California Family Risk Assessment (CFRA)		Multiple forms	After case closure	Substantiated	0.6800
	114	CFRA with possibility to overrule	SCJ	Multiple forms	After case closure	Substantiated	0.5100
Johnson, Clancy, and Bastian (2015)	236	California Family Risk Assessment (CFRA)	Actuarial	Multiple forms	After case closure	Substantiated	0.7400
Lealman, Phillips, Haigh, Stone, and Ord-Smith (1983)	2802	No name given	Actuarial	Multiple forms	After assessment	Substantiated	0.7445

(*continued on next page*)

Table 1 (*continued*)

Author(s)/ Pub. year	*N*	Name of instrument	Type of instrument	Maltreatment type	Start of follow-up	Outcome	AUC
Loman and Siegel (2004)	15100	Minnesota Family Risk Assessment (MFRA)	Actuarial	Multiple forms	After case closure	Recidivism/relapse	0.8345
Milner, Gold, Ayoub, and Jacewitz (1984)	190	Child Abuse Potential Inventory (CAPI)	Actuarial	Neglect	After assessment	Substantiated	0.6078
				General	After assessment	Substantiated	0.6895
Murphy, Orkow, and Nicola (1985)	587	Family Stress Checklist	Actuarial	Multiple forms	After assessment	Substantiated	0.8470
Ondersma et al. (2005)	713	Child Abuse Potential Inventory (CAPI)	Actuarial	Physical	After assessment	New reports	0.5565
				Neglect	After assessment	New reports	0.5907
				Multiple forms	After assessment	New reports	0.5565
		Brief CAPI (CAPI shortened version)	Actuarial	Physical	After assessment	New reports	0.5226
				Neglect	After assessment	New reports	0.5508
				Multiple forms	After assessment	New reports	0.5282
Sledjeski, Dierker, Brigham, and Breslin (2008)	244	Connecticut Risk Assessment–regression model	Actuarial	Multiple forms	After case closure	Substantiated	0.6200
		Connecticut Risk Assessment–CART model	Actuarial	Multiple forms	After case closure	Substantiated	0.6700
Staal, Hermanns, Schrijvers, and van Stel(2013)	1850	Structured Problem Analysis of Raising Kids (SPARK)	SCJ	Multiple forms	After assessment	New reports	0.7450
Vaithianathan, Maloney, Putnam-Hornstein and Jiang (2013)	17396	Predictive Risk Model	Actuarial	Multiple forms	After assessment	Substantiated	0.7600
Van der Put et al. (2016)	3963	Actuarial Risk ass. Instrument Youth (ARIJ)	Actuarial	Multiple forms	After assessment	Recidivism/relapse	0.6300
		Check List of Child Safety (CLCS)	SCJ	Multiple forms	After assessment	Recidivism/relapse	0.5300
van der Put et al.(2017)	4962	Instrument for early identification of Parents At Risk for child Abuse and Neglect (IPARAN)	Actuarial	Multiple forms	After assessment	New reports	0.7450
Van der Put, Hermanns, Van Rijn-van Gelderen, and Sondeijker (2016)	491	California Family Risk Assessment (CFRA)	Actuarial	Multiple forms	After assessment	New reports	0.6930
Hermanns et al.,		CFRA Abuse scale		Multiple forms	After assessment	New reports	0.7190
		CFRA Neglect scale		Multiple forms	After assessment	New reports	0.6530
		Detection of Unsafety in Families (DUF)	Actuarial	Multiple forms	After assessment	New reports	0.7990
Wood (1997)	409	NCCD Risk Assessment Tools	Actuarial	Multiple forms	Not specified	Substantiated	0.6195
				Multiple forms	Not specified	New reports	0.6221

Note: pub. year = year of publication; *N* = total sample size; maltreatment type = type of maltreatment predicted with the instrument; start of follow-up = the moment at which follow-up started; outcome = type of outcome used in assessing the predictive accuracy of the instrument; AUC = Area Under the ROC Curve; Y-ACNAT-NO = Youth Actuarial Care Needs Assessment Tool for Non-Offenders; CANTS 17B = Child Abuse and Neglect Tracking System − 17B; CARE-NL = Child Abuse Risk Evaluation − *Nederland* [the Netherlands]; CART = Classification and Regression Tree; NCCD = National Council on Crime and Delinquency; SCJ = Structured Clinical Judgment; consensus = consensus-based; multiple forms = instrument was designed to predict multiple forms of child maltreatment (including neglect); general = general maltreatment (type not further specified); after assessment = follow-up started directly after the risk assessment; after case closure = follow-up started directly after case closure; substantiated = child maltreatment substantiated by child protective services; investigation = child maltreatment under investigation by child protective services; recidivism/relapse = relapse of the child (and the family) in child protective services; new reports = new official reports of suspected maltreatment or neglect.

significant variance between primary studies (Level 3), two separate one-tailed log-likelihood-ratio-tests were performed in which the deviance of the full model was compared to the deviance of a model excluding one of the two variance parameters. The sampling variance of observed effect sizes (Level 1) was estimated by using a formula as given by Cheung (2014, pg. 2015). All model parameters were estimated using the restricted maximum likelihood estimation method and before moderator analyses were conducted, each continuous variable was centered around its mean and dichotomous dummy variables were created for all categories of discrete variables. The log-likelihood-ratio-tests were performed one-tailed whereas all other tests were performed two-tailed. We considered p-values < 0.05 as statistically significant, and < 0.10 as trend significant.

3. Results

3.1. Descriptive characteristics, central tendency, and variability

The present study included 30 studies (k) published between 1978 and 2016 (median publication year is 2005). In total, these studies reported on validation research of 27 different risk assessment instruments, from which 67 effect sizes could be extracted. Each effect size represented the discriminative accuracy of a particular risk assessment instrument or a statistical predictive model that was used for the purpose of risk assessment. An overview of all risk assessment instruments that have been examined in the primary studies can be found in Table 1.

The total sample size consisted of $N = 87.329$ children and their families for whom the risk for child maltreatment was assessed using one of the risk assessment instruments as listed in Table 1. Sample sizes in the primary studies ranged from 42 to 17,396 participants. The included studies were conducted in the USA ($k = 18$), Europe ($k = 9$), Canada ($k = 1$), New Zealand ($k = 1$), and Japan ($k = 1$).

The statistical analyses yielded an overall effect of $z = 0.328$ (SE = 0.033), $t(66) = 9.961$, p .0001 < 0.0001, which equals an AUC value of 0.681 (see Table 2). The results of the trim-and-fill-analysis suggested that bias was present in the data set, because of an asymmetric distribution of effect sizes. From the funnel plot in Fig. 3 can be derived that effect sizes were missing on the right sight of the funnel, and consequently, 12 effect sizes (from 6 studies) were added to the dataset, so that a "corrected" overall effect could be estimated. The results showed a "corrected" overall effect of $z = 0.370$ (SE = 0.032), $t(78) = 11.631$, $p < 0.001$, equaling an AUC value of AUC = 0.704 (see Table 2).

As for heterogeneity in effect sizes, the one-sided likelihood-ratio tests showed significant variance both on the second level $\chi^2(1) = 208.445$, $p < 0.0001$ and the third level $\chi^2(1) = 30.618$, $p < 0.0001$ of the *meta*-analytic model. Consequently, we proceeded to moderator analyses to examine whether characteristics of the risk assessment instruments, the study, and/or the sample could (partly) explain level 2 and/or level 3 variance.

3.2. Univariate moderator analyses

First, each potential moderator of interest was examined in a bivariate model. The results of these analyses can be found in Table 3 in which potential moderators are classified into characteristics of the instrument, the study, and the sample. Below, the same classification is used to describe the results.

3.3. Instrument characteristics

The results of the moderator analyses showed a significant effect for type of risk assessment approach. The mean effect size of actuarial instruments (AUC = 0.704) was higher than consensus-based instruments (AUC = 0.644) and instruments used for making a structured clinical judgment (AUC = 0.592). No significant difference was found between the mean effect of consensus-based instruments and the mean effect of instruments used for making a structured clinical judgment. Further, a significant moderating effect was found for the focus of the instrument. The mean effect size of instruments predicting the onset of maltreatment (AUC = 0.744) was higher than the mean effect size of instruments predicting the recurrence of maltreatment (AUC = 0.659). No significant moderating effects were found for the number of items of a risk assessment instrument, nor for type of assessor.

Table 2
Overall Effects Before and After Trim-and-Fill Analyses.

	Mean z (SE)	95% CI	Sig. mean z (p)	% var. at level 1	Level 2 variance	% Var. at level 2	Level 3 variance	% Var. at level 3	AUC-value
Overall effect before trim-and-fill	.328 (.033)	.262, 0.393	< 0.001***	1.1	0.005***	15.3	0.028***	83.6	0.681
Overall effect after trim-and-fill	.370 (.032)	.307, 0.433	< 0.001***	1.1	0.004***	11.3	0.032***	87.6	0.704

Note: Mean z = mean effect size (Fisher's z); SE = standard error; CI = confidence interval; Sig = significance; Var = variance; Level 1 variance = sampling variance of observed effect sizes; Level 2 variance = variance between effect sizes extracted from the same study; Level 3 variance = variance between studies; AUC = Area under the ROC curve.
*** $p < 0.001$.

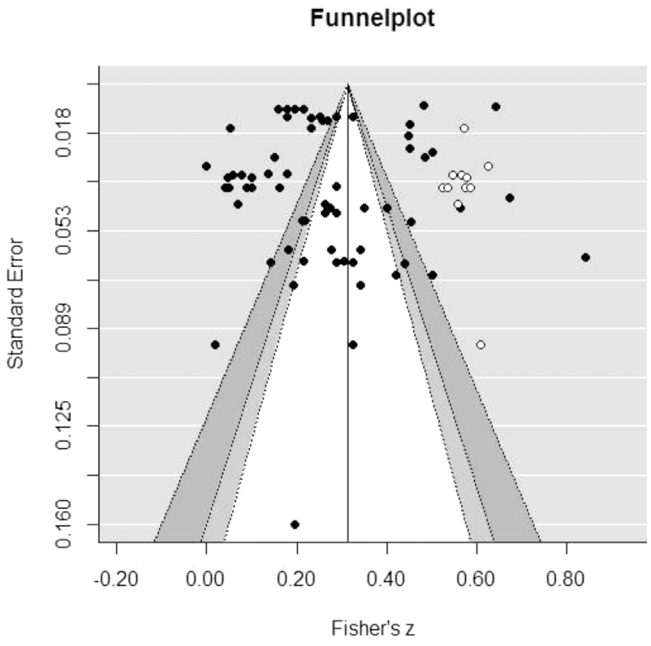

Fig. 3. Funnel Plot.
Note: Contour enhanced funnel plot with the standard error on the y-axis and Fisher's *z* on the x-axis. The black dots denote the observed effect sizes, and the white dots denote the filled effect sizes. The solid vertical line represents the overall mean effect. From inside to outside, the dashed lines limit the 90%, 95%, and 99% pseudo confidence interval regions.

3.4. Study design characteristics

A significant moderating effect was found for the sample used for validation. The mean effect size of instruments assessed in a construction sample (AUC = 0.745) was higher than the mean effect size of instruments assessed in a validation sample (AUC = 0.662). None of the study design characteristics (i.e., study design, follow-up length, type of follow-up, outcome measure, type of maltreatment, publication year, sample type, sample size, and percentage of cultural minority in samples) significantly moderated the overall effect.

3.5. Multiple moderator models

Lastly, we tested two multiple moderator models to determine the unique contribution of the variables that were tested in the bivariate models and significantly moderated the overall effect (see Table 4).

In Model 1 we examined whether the variable instrument type could significantly explain level 2 or level 3 variance when controlling for sample type. The results showed a significant negative moderating effect of both consensus-based instruments (versus actuarial instruments; $\beta = -0.100$, $p < 0.05$) and instruments for making a structured clinical judgment (versus actuarial instruments; $\beta = -0.127$, $p < 0.01$). A significant negative moderating effect was also found for validation sample (versus construction sample; $\beta = -0.111$, $p < 0.001$). Next, we built Model 2 by extending Model 1 with focus of risk assessment. In Model 2, there was also a significant negative moderating effect of both consensus-based instruments (versus actuarial instruments; $\beta = -0.095$, $p < 0.05$), instruments for making a structured clinical judgment (versus actuarial instrument; $\beta = -0.124$, $p < 0.01$), and validation sample (versus construction sample; $\beta = -0.113$, $p < 0.001$). A positive moderating effect was found for onset of maltreatment (versus recurrence of maltreatment; $\beta = 0.181$, $p < 0.01$). In both models, multicollinearity did not seem to be a problem, since all variance inflation factors were below 1.200 and all tolerance statistics were above 0.880.

4. Discussion

This meta-analysis investigated the predictive validity of risk assessment instruments for child maltreatment, and whether this is influenced by characteristics of instruments, studies, and samples. Overall, a significant medium effect was found (AUC = 0.681), indicating a moderate predictive accuracy of risk assessment instruments. This overall effect is comparable with effects sizes found in meta-analyses on risk assessment instruments in (juvenile) justice settings. For example, Schwalbe (2007) found a mean AUC of 0.640 for juvenile justice risk assessment instruments, and Fazel, Singh, and Grann (2012) found a mean AUC of 0.660 for juvenile and adult risk assessment instruments. The results of the trim-and-fill-analysis suggested that bias was present in the data set, and therefore a "corrected" overall effect was estimated, resulting in an AUC value of 0.704. Because there are several methodological shortcomings regarding the trim-and-fill method (see limitations section), this AUC value should not be interpreted as a true effect size, but only as

Table 3
Results for Categorical and Continuous Moderators Tested in Bivariate Models.

Moderator variables	# Studies	# ES	Intercept/mean *z* (95% CI)	β_1(95% CI)	Mean AUC	*F* (df1, df2)[a]	*p*[b]	Level 2 variance	Level 3 variance
Overall Effect	30	67	.328(.262, 0.393)***		0.681				
Instrument characteristics									
Type of risk assessment approach						14.579(2, 64)	< 0.001***	0.002***	0.032***
Actuarial (RC)	25	49	0.371(0.301, 0.441)***		0.704				
Consensus-based	3	9	0.259(0.151, 0.367)***	−0.112(−0.204, −0.021)*	0.644				
Structured clinical judgment	6	9	0.163(0.063, 0.264)**	−0.208(-0.293, −0.122)***	0.592				
Number of items	28	63	0.325(0.258, 0.391)***	0.000(−0.001, 0.000)	–	0.604(1, 61)	0.440	0.005***	0.028***
Type of assessor						0.054(3, 62)	0.983	0.005***	0.032***
Professional (RC)	15	38	0.325(0.227, 0.423)***		0.680				
Client (i.e., self-report)	4	6	0.309(0.107, 0.510)**	−0.017(−0.240, 0.207)	0.672				
Researcher	8	19	0.313(0.179, 0.447)***	−0.012(−0.178, 0.154)	0.674				
Computer system	2	3	0.368(0.102, 0.635)**	0.043(−0.240, 0.327)	0.703				
Focus of risk assessment						3.280(2, 64)	0.044*	0.005***	0.024***
Recurrence of maltreatment (RC)	16	39	0.285(0.202, 0.368)***		0.659				
Onset of maltreatment	9	12	0.448(0.334, 0.561)***	0.162(0.022, 0.303)*	0.744				
Both/not specified	5	16	0.249(0.097, 0.402)**	−0.036(−0.209, 0.138)	0.639				
Study design characteristics									
Study design						1.573(1, 65)	0.214	0.005***	0.031***
Prospective (RC)	20	50	0.306(0.228, 0.383)***		0.670				
Retrospective	10	17	0.379(0.274, 0.485)***	0.074(−0.044, 0.191)	0.709				
Sample used for validation						32.071(1, 65)	< 0.001***	0.002***	0.034***
Construction sample (RC)	10	14	0.450(0.369, 0.531)***		0.745				
Validation sample	24	53	0.291(0.220, 0.362)***	−0.159(−0.216, −0.103) ***	0.662				
Follow-up length (in months)	29	61	0.336(0.268, 0.403)***	0.000(−0.003, 0.003)	–	0.018 (1, 59)	0.893	0.006***	0.027***
Type of follow-up						0.279(2, 64)	0.758	0.005***	0.029***
Time after assessment (RC)	20	48	0.338(0.258, 0.417)***		0.687				
Time after case closure	9	13	0.328(0.210, 0.446)***	−0.010(−0.142, 0.123)	0.682				
Both/not specified	2	6	0.239(−0.013, 0.491)+	−0.099(−0.363, 0.166)	0.634				
Outcome measure						0.187(5, 61)	0.967	0.006***	0.028***
Substantiated maltreatment (RC)	18	32	0.325(0.242, 0.408)***		0.680				
Number of new reports	9	19	0.326(0.215, 0.438)***	0.001(−0.129, 0.132)	0.681				
Number of investigations	2	6	0.314(0.195, 0.432)***	−0.011(−0.101, 0.078)	0.674				
Recidivism/relapse	3	5	0.297(0.137, 0.457)***	−0.028(−0.186, 0.130)	0.665				
Supervision order	3	3	0.398(0.209, 0.587)***	0.073(−0.131, 0.277)	0.718				
Out of home placement	2	2	0.336(0.126, 0.546)***	0.011(−0.212, 0.234)	0.686				
Type of maltreatment						0.366(3, 63)	0.778	0.006***	0.027***
Multiple forms (RC)	23	46	0.326(0.255, 0.396)***		0.681				
Physical abuse	5	9	0.303(0.198, 0.408)***	−0.023(−0.119, 0.074)	0.668				
Neglect	5	7	0.300(0.195, 0.406)***	−0.025(−0.123, 0.072)	0.667				
Maltreatment not specified	3	5	0.398(0.209, 0.586)***	0.072(−0.124, 0.268)	0.718				
Publication year	29	66	0.319(0.252, 0.386)***	−0.004(−0.009, 0.002)	–	1.747(1, 64)	0.191	0.005***	0.028***
Type of sample						2.241(1, 65)	0.139	0.005***	0.026***

(*continued on next page*)

Table 3 (*continued*)

Moderator variables	# Studies	# ES	Intercept/mean *z* (95% CI)	β_1(95% CI)	Mean AUC	*F* (df1, df2)[a]	*p*[b]	Level 2 variance	Level 3 variance
Clinical sample (RC)	23	58	0.299(0.224, 0.373)***		0.666				
Non-clinical sample	7	9	0.424(0.292, 0.557)***	0.109(−0.036, 0.255)	0.732				
Sample size	30	67	0.327(0.261, 0.393)***	0.000(−0.000, 0.000)	–	1.165(1, 65)	0.284	0.005***	0.028***
Percentage cultural minority	17	40	0.322(0.228, 0.417)***	−0.002(−0.006, 0.002)	–	1.186(1, 38)	0.283	0.002***	0.030***

Note. # Studies = number of studies # ES = number of effect sizes; mean *z* = mean effect size (*z*) CI = confidence interval; β_1 = estimated regression coefficient; mean AUC = mean effect size expressed in an AUC value; AUC = area under the ROC curve; df = ° of freedom; Level 2 variance = variance between effect sizes extracted from the same study; Level 3 variance = variance between studies.

[a] Omnibus test of all regression coefficients in the model.

[b] *p*-Value of the omnibus test.

+ $p < 0.1$.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

Table 4
Results for the Multiple Moderator Models.

Moderator variables	Model 1			Model 2		
	β (SE)	95% CI	*t* statistic	β (SE)	95% CI	*t* statistic
Intercept	0.442 (0.038)***	0.365, 0.519	11.475	0.388 (0.041)***	0.306, 0.469	9.490
Control variables						
Validation sample (vs. construction sample)	−0.111 (0.027)***	−0.165, −0.057	−4.116	−0.113 (0.027)***	−0.166, −0.060	−4.256
Structured Clinical Judgement (vs. actuarial instruments)	−0.127 (0.040)**	−0.208, −0.046	−3.148	−0.124 (0.040)**	−0.204, −0.045	−3.129
Consensus-based (vs. actuarial instruments)	−0.100 (0.038)*	−0.175, −0.024	−2.640	−0.095 (0.037)*	−0.169, −0.020	−2.533
Onset of maltreatment (vs. Recurrence of maltreatment)				0.181 (0.068)**	−0.045, 0.317	2.660
F (df1, df2)[a]	21.810 (3, 63)***			18.160 (4, 62)***		
Level 2 variance	0.001***			0.001***		
Level 3 variance	0.033***			0.027***		

Note: β = estimated regression coefficient; SE = standard error; CI = confidence interval; df = degress of freedom; Level 2 variance = variance between effect sizes extracted from the same study; Level 3 variance = variance between studies.

[a] Omnibus test of all regression coefficients in the model.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

an indicator of (possible) bias in the data.

Moderator analyses revealed a number of significant moderators. In line with our expectations, we found higher mean effect sizes in construction samples (AUC = 0.745) than in validation samples (AUC = 0.662). As mentioned in the introduction, predictive validity estimates reported for construction samples are commonly inflated, as "overfitting" data is a common problem for models built and tested in construction samples. An important limitation of the literature is that only very few instruments have been validated in multiple independent samples. With the exception of the CFRA and the CAPI, the reliability of predictive validity estimates is generally unknown. It is possible that high estimates of the predictive value of risk assessment instruments would decline when multiple estimates are averaged over independent samples. Also, a positive moderating effect was found for instruments predicting the onset of maltreatment (AUC = 0.744) compared to instruments predicting the recurrence of maltreatment (AUC = 0.659). A possible explanation for this finding is that predictive models benefit from greater variation in the prevalence of risk factors within general population samples compared to the generally higher risk clinical samples.

In line with our expectations we found that actuarial instruments (AUC = 0.704) performed better than consensus-based (AUC = 0.644) and SCJ instruments (AUC = 0.592). In other words, actuarial instruments have a better discriminative accuracy than clinical methods. Meta-analyses on the performance of risk assessment instruments in other disciplines, such as criminal justice, forensic mental health, and clinical psychology, also found that actuarial methods outperform clinical methods (Aegisdottir et al., 2006; Dawes et al., 1989; Grove & Meehl, 1996; Hanson and Morton-Bourgon, 2009; Hilton, Harris, & Rice, 2006). In contrast to our expectations, we found no significant difference in discriminative accuracy of instruments between consensus-based and SCJ instruments. A possible explanation is that the two types of instruments are similar in the sense that the weighting of risk factors and making the final decision is left to the professional. Rather than the degree to which both types of instruments are derived from empirical research, the aspect of subjectivity present in both instrument types, may be more determining − or perhaps even the decisive factor − in the predictive accuracy.

We did not find a moderating effect of the length of the instrument, as predictive validity did not vary with the number of items risk assessment instruments were comprised of. So, brief instruments were as predictive as longer instruments. This finding was not in line with what we expected, because Schwalbe (2007) found in his meta-analysis that brief instruments yielded smaller effect sizes than other types of instruments. It may be possible that brief instruments used in child welfare are more often actuarial in nature and derived from multivariate statistical techniques, assessing only risk factors that make a significant and unique contribution to the prediction of child maltreatment. From this perspective, it is not necessary (nor desirable) for clinical practitioners to use instruments with a large number of items for adequately capturing the risk for maltreatment. Further, predictive validity was not dependent on the type of assessor or study design, even though a prospective design is considered to be superior to a retrospective design. The predictive validity of instruments examined in prospective studies was lower (AUC = 0.670) than that of instruments examined in retrospective studies (AUC = 0.709), but this difference was not significant.

4.1. Clinical implications

Several implications for clinical practice can be derived from our results. Overall, a medium significant effect was found, indicating a moderate predictive accuracy. This result shows that it is important to use risk assessment tools, especially because *unstructured* clinical judgment is widely recognised to be flawed, due to lower transparency, reliability and predictive validity (see for example Dorsey, Mustillo, Farmer, & Elbogen, 2008; Munro, 1999; van der Put et al., 2017). Furthermore, our review showed that actuarial instruments are highly preferable to clinical instruments because actuarial instruments make a better distinction between

high-risk and low-risk cases. However, to be able to bring risk and needs assessment in child welfare to a higher level, it is important to improve actuarial instruments. Actuarial instruments in their current form are limited in their ability to guide case planning because they do not identify the full range of risk factors necessary for adequate intervention planning (Schwalbe, 2008; Shlonsky & Wagner, 2005). Most actuarial instruments that are currently used by child welfare professionals are brief instruments derived from multivariate statistical techniques consisting mainly of static risk factors. Therefore, these instruments are particularly suitable for the purpose of risk assessment (predicting future child maltreatment to determine intervention urgency and intensity) but not for the purpose of needs assessment (identifying targets of interventions in order to individualize case planning). Actuarial instruments for child maltreatment should therefore be further developed and strengthened by distinguishing between risk and needs assessment and by integrating risk assessment with case management.

Moderator analyses revealed that onset of maltreatment can be better predicted than recurrence of maltreatment, which stresses the importance of early detection and prevention of child maltreatment. Our review showed that the predictive validity of currently available screening instruments is sufficient to justify using these instruments in assessing risks for child maltreatment in the general population. For instance, different types of child and youth care professionals may screen for child maltreatment during regular health check-ups for children and juveniles. Currently, the most commonly employed prognostic process in child welfare services is assessing the risk of *recurrence* of child maltreatment. However, screening for potential child maltreatment before the maltreatment actually occurs contributes to the early detection of child maltreatment risks which is necessary to timely refer children and their families early intervention programs. Given the relatively good performance of screening tools, it is fruitful to invest time, money, and resources in developing and strengthening preventive strategies for child maltreatment.

4.2. Limitations

Several limitations need to be discussed. One limitation is related to the outcome measure assessed in primary studies, because in some studies it is assumed that a report, investigation, or a recurrence in child protection is indicative of abuse or neglect. However, the relationship between child protection system contact and maltreatment is not straightforward (Jenkins, Tilbury, Mazerolle, & Haues, 2017). First, studies showed that a large proportion of child maltreatment is not reported to child protection authorities (Cyr et al., 2013; Finkelhor, Ormrod, & Turner, 2005; Finkelhor, Ormrod, & Turner, 2009). Population based surveys showed that rates of maltreatment are more than ten times the rates of substantiated maltreatment in those same jurisdictions (Fergusson, Horwood, & Woodward, 2000; Finkelhor, 2008; MacMillan, Jamieson, & Walsh, 2003). Second, in the majority of cases reported to child protection authorities, maltreatment is not substantiated.

A second limitation is that the reliability of predictive validity estimates is generally unknown, because only very few instruments have been validated in multiple independent samples. For empirically derived actuarial instruments, ongoing replication studies are required to determine whether estimations of predictive validity are robust to random sampling variation. Even cross-validated instruments like the Y-ACNAT-NO (Assink,Van der Put, Hoeve et al., 2015) are vulnerable to random sampling error when construction and validation samples are randomly selected from the same sample.

A third limitation is that studies were included regardless of their methodological quality in order to analyze a representative sample of the literature. To address this limitation, possible sources of within- and between-study heterogeneity were examined, including features of methodological quality such as sample size, prospective or retrospective design, and length of follow-up. While multiple potential moderating variables were examined, it is possible that there are other study design, sample, and instrument characteristics that contribute to effect size variation, but which were not investigated. For example, clinical background of the professionals who administered the instruments was not included whereas previous research showed that this may be an important moderator of the predictive validity of risk assessment instruments (Aegisdóttir et al., 2006). Studies generally do not report on potentially important moderators such as clinical background of professionals.

Fourth, there are several methodological difficulties regarding the trim and fill method. First, Nakagawa and Santos (2012) mentioned that this method has originally been designed for meta-analyses in which independence of effect sizes can be assumed. Second, the performance of the trim and fill method is limited when effect sizes prove to be heterogeneous (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau, & Olkin, 2003). Third, the application of the trim and fill method could mean adding and adjusting for non-existent effect sizes in response to funnel plots that are asymmetrical, simply because of random variation (Egger, Davey-Smith, & Altman, 2001). Despite these shortcomings, there is no best method for detecting and handling missing data in meta-analysis, and therefore, the results from the trim and fill method should be interpreted with caution. In the present study, we only used the trim-and-fill method to calculate a corrected overall effect (R Core Team, 2015). .

Fifth, we included 40 years of research in our study (the oldest study we found was conducted in 1978). It is possible that in earlier years, risk assessment tools as well as research designs were less robust than in later years. However, moderator analyses showed that publication year was not a significant moderator of the predictive validity of risk assessment instruments.

A final limitation is that many moderator analyses were based on a small number of effect sizes, implying a low statistical power in testing potential moderators in the bivariate and multiple moderator models.

5. Conclusion

The present study is the first meta-analysis on the predictive validity of risk assessment instruments for child maltreatment, with the aim to learn more about the general effectiveness of these instruments and about the characteristics that influence the predictive validity. This study showed that the discriminative accuracy of actuarial instruments is better than the discriminative accuracy of

both consensus-based instruments and structured clinical judgment instruments, and therefore we conclude that actuarial instruments perform better than clinical instruments. Because actuarial risk assessment instruments used in child welfare are limited in their ability to guide case management, it is important that these instruments are further developed. One important improvement is to extend actuarial instruments with a broad array of dynamic risk factors that can be used for both formulating clinical hypotheses and identifying targets for interventions that are aimed at reducing the risk for (the recurrence of) child maltreatment.

References[*]

Aegisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341–382.

*Altemeier, W. A., O'Connor, S., Vietze, P., Sandler, H., & Sherrod, K. (1984). Prediction of child abuse: a prospective study of feasibility. *Child Abuse & Neglect, 8*(4), 393–400.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 3: Receiver operating characteristic plots. *British Medical Journal, 309*, 188.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct.* Routledge.

Arad-Davidson, B., & Benbenishty, R. (2008). The role of workers' attitudes and parent and child wishes in child protection workers' assessments and recommendation regarding removal and reunification. *Children and Youth Services Review, 30*, 107–121.

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology, 12*, 154–174.

*Assink, M., Van der Put, C. E., Oort, F. J., & Stams, G. J. J. M. (2015). The development and validation of the youth actuarial care needs assessment tool for non-Offenders (Y-ACNAT-NO). *BMC Psychiatry, 15*, 36.

Assink, M., Van der Put, C. E., Hoeve, M., de Vries, S. L., Stams, G. J. J. M., & Oort, F. J. (2015). Risk factors for persistent delinquent behavior among juveniles: A meta-analytic review. *Clinical Psychology Review, 42*, 47–61.

*Ayoub, C. C., & Milner, J. S. (1985). Failure to thrive: Parental indicators, types, and outcomes. *Child Abuse & Neglect, 9*(4), 491–499.

*Baird, C., & Wagner, D. (2000). The relative validity of actuarial-and consensus-based risk assessment systems. *Children and Youth Services Review, 22*(11), 839–871.

*Barber, J. G., Shlonsky, A., Black, T., Goodman, D., & Trocmé, N. (2008). Reliability and predictive validity of a consensus-based risk assessment tool. *Journal of Public Child Welfare, 2*(2), 173–195.

Barlow, J., Fisher, J. D., & Jones, D. (2012). *Systematic review of models of analyzing significant harm.* Oxford University. Retrieved online on November 29, 2016 at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/183949/DFE-RR199.pdf.

*Bartelink, C., De Kwaadsteniet, L., Ten Berge, I., & Witteman, C. (2015). *Betrouwbaarheid en validiteit van de LIRIK: Eindrapport LIRIK Valideringsonderzoek*. [Reliability and validity of the LIRIK: Final report LIRIK validation project] Utrecht, the Netherlands: Nederlands Jeugd Instituut. Retrieved from: http://www.nji.nl/nl10/Download-NJi/Publicatie-NJi/Betrouwbaarheid-en-validiteit-van-de-LIRIK.pdf.

Bartelink, C., Van Yperen, T. A., & Ten Berge, I. J. (2015). Deciding on child maltreatment: A literature review on methods that improve decision-making. *Child Abuse & Neglect, 49*, 142–153.

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

*Brayden, R. M., Altemeier, W. A., Dietrich, M. S., Tucker, D. D., Christensen, M. J., McLaughlin, F. J., & Sherrod, K. B. (1993). A prospective study of secondary prevention of child maltreatment. *The Journal of Pediatrics, 122*(4), 511–516.

Caldwell, R. A., Bogat, G. A., & Davidson, W. S., II (1988). The assessment of child abuse potential and the prevention of child abuse and neglect: A policy analysis. *American Journal of Community Psychology, 16*(5), 609–624.

*Camasso, M. J., & Jagannathan, R. (1995). Prediction accuracy of the Washington and Illinois risk assessment instruments: An application of receiver operating characteristic curve analysis. *Social Work Research, 19*(3), 174–183.

Cash, S. J. (2001). Risk assessment in child welfare: The art and science. *Children and Youth Services Review, 23*(11), 811–830.

*Chaffin, M., & Valle, L. A. (2003). Dynamic prediction characteristics of the child abuse potential inventory. *Child Abuse & Neglect, 27*(5), 463–481.

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*(2), 211–229.

*Coohey, C., Johnson, K., Renner, L. M., & Easton, S. D. (2013). Actuarial risk assessment in child protective services: Construction methodology and performance criteria. *Children and Youth Services Review, 35*(1), 151–161.

Cyr, K., Chamberland, C., Clément, M.-, Lessard, G., Wemmers, J.-A., Collin-Vézina, D., & Damant, D. (2013). Polyvictimization and victimization of children and youth: Results from a populational survey. *Child Abuse & Neglect, 37*(10), 814.

D'andrade, A., Austin, M. J., & Benton, A. (2008). Risk and safety assessment in child welfare: Instrument comparisons. *Journal of Evidence-Based Social Work, 5*(1–2), 31–56.

*Dankert, E. W., & Johnson, K. (2014). *risk assessment validation: A prospective study.* NCCD Children's Research Center.. Retrieved from: http://www.nccdglobal.org/sites/default/files/publication_pdf/risk-assessment-validation.pdf.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: Free Press.

*De Ruiter, C., Hildebrand, M., & Van der Hoorn, S. (2012). Gestructureerde risicotaxatie bij kindermishandeling: De Child Abuse Risk Evaluation-Nederlandse versie [Structured risk assessment for child maltreatment: The Dutch version of the Child Abuse Risk Evaluation (CARE-NL)]. *Psychologie, 3*, 10–17.

De Vogel, V., De Ruiter, C., Hildebrand, M., Bos, B., & Van de Ven, P. (2004). Type of discharge and risk of recidivism measured by the HCR-20: A retrospective study in a Dutch sample of treated forensic psychiatric patients. *International Journal of Forensic Mental Health, 3*(2), 149–165.

DePanfilis, D., & Girvin, H. (2005). Investigating child maltreatment in out-of-home care: Barriers to good decision-making. *Children & Youth Services Review, 27*, 353–374.

Dorsey, S., Mustillo, S. A., Farmer, E. M. Z., & Elbogen, E. (2008). Caseworker assessments of risk for recurrent maltreatment: Association with case-specific risk factors and re-reports. *Child Abuse & Neglect, 32*, 377–391.

Doueck, H. J., English, D. J., DePanfilis, D., & Moote, G. T. (1992). Decision-making in child protective services: A comparison of selected risk-assessment systems. *Child Welfare, 72*(5), 441–452.

Duval, S., & Tweedie, R. (2000a). A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–99.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Systematic Reviews in Health Care: Meta-analysis in Context. In M. Egger, G. Davey Smith, & D. Altman (Eds.). (2nd ed.). London: BMJ Books.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behavior in 73 samples involving 24827 people: Systematic review and meta-analysis. *British Medical Journal,* 345 [e4692].

Fergusson, D. M., Horwood, L. J., & Woodward, L. J. (2000). The stability of child abuse reports: a longitudinal study of the reporting behaviour of young adults. *Psychological Medicine, 30*, 529–544.

[*] References marked with an asterisk were included in the meta-analysis

Finkelhor, D., Ormrod, R. K., & Turner, H. A. (2005). The victimization of children and youth: A comprehensive national survey. *Child Maltreatment, 10*(1), 5–25.
Finkelhor, D. (2008). *Childhood Victimization. Violence, Crime and Abuse in the Lives of Young People.* Oxford: Oxford University Press.
Finkelhor, D., Ormrod, R. K., & Turner, H. A. (2009). Lifetime assessment of poly-victimization in a national sample of children and youth. *Child Abuse & Neglect, 33*(7), 403–411.
*Flaherty, C. W. (2001). *An artificial neural network model for the prediction of child physical abuse recurrences. Doctoral Dissertation.* The University of Tennessee [ISBN: 0-493-33897-7].
Gambrill, E., & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth Services Review, 22*, 813–837.
Gershater-Molko, R. M., Lutzker, J. R., & Sherman, J. A. (2003). Assessing child neglect. *Aggression and Violent Behavior, 8*(6), 563–585.
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy and Law, 2*, 293–323.
*Hamilton, C. E., & Browne, K. D. (1999). Recurrent maltreatment during childhood: A survey of referrals to police child protection units in England. *Child Maltreatment, 4*, 275–286.
Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*, 1–21.
Hilton, N. Z., Harris, G. T., & Rice, M. E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist, 34*(3), 400–409.
*Horikawa, H., Suguimoto, S. P., Musumari, P. M., Techasrivichien, T., Ono-Kihara, M., & Kihara, M. (2016). Development of a prediction model for child maltreatment recurrence in Japan: A historical cohort study using data from a Child Guidance Center. *Child Abuse & Neglect, 59*, 55–65.
Houben, M., Van den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin, 141*, 901–930.
Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates.
*Hunter, R. S., Kilstrom, N., Kraybill, E. N., & Loda, F. (1978). Antecedents of child abuse and neglect in premature infants: A prospective study in a newborn intensive care unit. *Pediatrics, 61*(4), 629–635.
Jenkins, B. Q., Tilbury, C., Mazerolle, P., & Hayes, H. (2017). The complexity of child protection recurrence: The case for a systems approach. *Child Abuse & Neglect, 63*, 162–171.
Johnson, M. A., Stone, S., Lou, C., Vu, C. M., Ling, J., Mizrahi, P., & Austin, M. J. (2008). Family assessment in child welfare services: Instrument comparisons. *Journal of Evidence-based Social Work, 5*(1-2), 57–90.
*Johnson, W., Clancy, T., & Bastian, P. (2015). Child abuse/neglect risk assessment under field practice conditions: Tests of external and temporal validity and comparison with heart disease prediction. *Children and Youth Services Review, 56*, 76–85.
Johnson, W. (2006a). The risk assessment wars: A commentary response to Evaluating the effectiveness of actuarial risk assessment models by Donald Baumann J. Randolph Law, Janess Sheets, Grant Reid, and J. Christopher Graham. *Children and Youth Services Review, 27*, 465–490.
Johnson, W. (2006b). Post-battle skirmish in the risk assessment wars: Rebuttal to the response of Baumann and colleagues to criticism of their paper: Evaluating the effectiveness of actuarial risk assessment models. *Children and Youth Services Review, 28*, 1124–1132.
*Johnson, W. L. (2011). The validity and utility of the California Family Risk Assessment under practice conditions in the field: A prospective study. *Child Aabuse & Neglect, 35*(1), 18–28.
Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*(17), 2693–2710.
Knoke, D., & Trocmé, N. (2005). Reviewing the evidence on assessing risk for child abuse and neglect. *Brief Treatment and Crisis Intervention, 5*(3), 310–327.
Kuppens, S., Laurent, L., Heyvaert, M., & Onghena, P. (2013). Associations between parental psychological control and relational aggression in children and adolescents: A multilevel and sequential meta-analysis. *Developmental Psychology, 49*(9), 1697–1712.
*Lealman, G., Phillips, J., Haigh, D., Stone, J., & Ord-Smith, C. (1983). Prediction and prevention of child abuse—an empty hope? *The Lancet, 321*(8339), 1423–1424.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousands Oaks: Sage.
Loeber, R., Slot, N. W., & Stouthamer-Loeber, M. (2008). A cumulative developmental model of risk and promotive factors. In R. Loeber, N. W. Slot, P. H. Van der Laan, & M. Hoeve (Eds.). *Tomorrow's criminals. The development of child delinquency and effective interventions* (pp. 133–161). Farnham: Ashgate.
*Loman, L. A., & Siegel, G. L. (2004). *An evaluation of the Minnesota SDM family risk assessment: Final report*St. Louis, MO: Institute of Applied Research. Retrieved from: http://www.iarstl.org/papers/FinalFRAReport.pdf.
Lyons, P., Doueck, H. J., & Wodarski, J. S. (1996). Risk assessment for child protective services: A review of the empirical literature on instrument performance. *Social Work Research, 20*(3), 143–155.
MacMillan, H. L., Jamieson, E., & Walsh, C. A. (2003). Reported contact with child protection services among those reporting child physical and sexual abuse: results from a community survey. *Child Abuse & Neglect, 27*(12), 1397–1408.
*Milner, J. S., Gold, R. G., Ayoub, C., & Jacewitz, M. M. (1984). Predictive validity of the child abuse potential inventory. *Journal of Consulting and Clinical Psychology, 52*(5), 879.
Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269.
Morton, T. (2003). *The risk wars. Commentary.* Atlanta.GA: Child Welfare Institute.
Mullen, B. (1989). *Advanced BASIC meta-analysis.* Hillsdale: Erlbaum.
Munro, E. (1999). Common errors of reasoning in child protection work. *Child Abuse & Neglect, 23*, 745–758.
Munro, E. (2004). A simpler way to understand the results of risk assessment instruments. *Children and Youth Services Review, 26*(9), 873–883.
*Murphy, S., Orkow, B., & Nicola, R. M. (1985). Prenatal prediction of child abuse and neglect: A prospective study. *Child Abuse & Neglect, 9*(2), 225–235.
Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology, 26*(5), 1253–1274.
*Ondersma, S. J., Chaffin, M. J., Mullins, S. M., & LeBreton, J. M. (2005). A brief form of the child abuse potential inventory: Development and validation. *Journal of Clinical Child and Adolescent Psychology, 34*(2), 301–311.
Pecora, P. J. (1991). Investigating allegations of child maltreatment: The strengths and limitations of current risk assessment systems. *Child & Youth Services, 15*(2), 73–92.
Peters, R., & Barlow, J. (2003). Systematic review of instruments designed to predict child maltreatment during the antenatal and postnatal periods. *Child Abuse Review, 12*(6), 416–439.
Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine, 26*(25), 4544–4562.
Pfister, H., & Böhm, G. (2008). The multiplicity of emotions: A framework of emotional functions in decision making. *Judgment and Decision Making, 3*, 5–17.
R Core Team (2015). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: Foundation for Statistical Computing. Retrieved from: http://www.R-project.org/.
Rapp, R. C., Van den Noortgate, W., Broekaert, E., & Vanderplasschen, W. (2014). The efficacy of case management with persons who have substance abuse problems: A three-level meta-analysis of outcomes. *Journal of Consulting and Clinical Psychology, 82*(4), 605–618.
Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615.
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.
Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper, & L. V. Hedges (Eds.). *The handbook of research synthesis*New York, NY: Sage pp. 39.
Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*(1), 19–30.
Schwalbe, C. S. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior, 31*(5), 449–462.
Schwalbe, C. S. (2008). Strengthening the integration of actuarial risk assessment with clinical judgment in an evidence based practice framework. *Children and Youth Services Review, 30*(12), 1458–1464.
Shlonsky, A., & Wagner, D. (2005). The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case

management. *Children and Youth Services Review, 27*(4), 409–427.
Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law, 31*(1), 8–22.
*Sledjeski, E. M., Dierker, L. C., Brigham, R., & Breslin, E. (2008). The use of risk assessment to predict recurrent maltreatment: A classification and regression tree analysis (CART). *Prevention Science, 9*(1), 28–37.
*Staal, I. I., Hermanns, J. M., Schrijvers, A. J., & van Stel, H. F. (2013). Risk assessment of parents' concerns at 18 months in preventive child health care predicted child abuse and neglect. *Child Abuse & Neglect, 37*(7), 475–484.
Stoltenborgh, M., Bakermans-Kranenburg, M. J., Alink, L. R., & IJzendoorn, M. H. (2015). The prevalence of child maltreatment across the globe: Review of a series of meta-analyses. *Child Abuse Review, 24*(1), 37–50.
Stowman, S. A., & Donohue, B. (2005). Assessing child neglect: A review of standardized measures. *Aggression and Violent Behavior, 10*(4), 491–512.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1–26. http://dx.doi.org/10.1111/1529-1006.001.
Tabachnik, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Allyn and Bacon.
Tatara (1996). A survey of states on CPS risk assessment practice: preliminary findings. *Paper presented at the 10th annual national roundtable on CPS risk assessment*.
Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*, 2113–2126.
*Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine, 45*(3), 354–359.
Van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*(5), 765–790.
Van den Noortgate, W., López-López, J. A., Marin-Martinez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594.
Van den Noortgate, W., López-López, J. A., Marin-Martinez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*(4), 1274–1294.
Van der Put, C. E., Hermanns, J., & Van Rijn-van Gelderen, L. (2016). Detection of unsafety in families with parental and/or child developmental problems at the start of family support. *BMC Psychiatry, 16*(1).
Van der Put, C. E., Assink, M., & Stams, G. J. J. M. (2016a). The effectiveness of risk assessment methods: commentary on Deciding on child maltreatment: A literature review on methods that improve decision-making. *Child Abuse and Neglect, 59*, 128–129.
*Van der Put, C. E., Assink, M., & Stams, G. J. J. M. (2016b). Predicting relapse of problematic child-Rearing situations. *Children and Youth Services Review, 61*, 288–295.
*van der Put, C. E., Bouwmeester-Landweer, M. B., Landsmeer-Beker, E. A., Wit, J. M., Dekker, F. W., Kousemaker, N. P. J., & Baartman, H. E. (2017). Screening for potential child maltreatment in parents of a newborn baby: The predictive validity of an Instrument for early identification of Parents At Risk for child Abuse and Neglect (IPARAN). *Child Abuse & Neglect, 70*, 160–168.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.
Walker, C. A., & Davies, J. (2010). A critical review of the psychometric evidence base of the Child Abuse Potential Inventory. *Journal of Family Violence, 25*(2), 215–227.
Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care. *JAMA Psychiatry, 70*(7), 750–761.
White, A., & Wash, P. (2006). *Risk assessment in child welfare: An issues paper*. Sydney: NWS Department of Community Services. Retrieved online on November 29, 2016 at http://www.communicty.nws.gov.au.
*Wood, J. M. (1997). Risk predictors for re-abuse or re-neglect in a predominantly Hispanic population. *Child Abuse & Neglect, 21*(4), 379–389.