Child Youth Care Forum (2017) 46:745–768
DOI 10.1007/s10566-017-9405-2

ORIGINAL PAPER

# Is It Safe? Reliability and Validity of Structured Versus Unstructured Child Safety Judgments

Cora Bartelink[1] · Leontien de Kwaadsteniet[2] · Ingrid J. ten Berge[1] · Cilia L. M. Witteman[2]

Published online: 19 May 2017

**Abstract**

*Background* The LIRIK, an instrument for the assessment of child safety and risk, is designed to improve assessments by guiding professionals through a structured evaluation of relevant signs, risk factors, and protective factors.
*Objective* We aimed to assess the interrater agreement and the predictive validity of professionals' judgments made with the LIRIK in comparison to unstructured judgments.
*Method* In study 1, professionals made safety and risk judgments for 12 vignettes with the LIRIK (group 1, $n = 36$) or without an instrument (group 2, $n = 43$). In study 2, we compared professionals' safety and risk judgments for 370 children made with the LIRIK (group 1, $n = 278$) or with no instrument (group 2, $n = 92$), with outcomes indicating actual unsafety in files 6 months later.
*Results* In study 1, agreement about safety and risks was poor to moderate in both groups. Differences between groups were small and inconsistent. In study 2, the predictive validity of judgments was weak to moderate in both groups. In neither group had unsafe outcomes increased consistently when unsafety or risks were assessed as higher.
*Conclusions* Judgments made with the LIRIK were not more reliable or valid than unstructured professional judgments. These findings raise important questions about the value of risk assessment instruments and about how professional safety and risk judgments can be improved.

**Keywords** Risk assessment · Child abuse and neglect · Agreement · Reliability · Validity

Cora Bartelink and Leontien de Kwaadsteniet contributed equally to the paper.

✉ Cora Bartelink
c.bartelink@nji.nl

[1] The Netherlands Youth Institute, Nederlands Jeugdinstituut, PO Box 19221, 3501 DE Utrecht, The Netherlands

[2] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

## Introduction

Child abuse and neglect seriously affect a child's healthy development (Edwards et al. 2005; Felitti et al. 1998; Nanni et al. 2012; Perry 2009). Therefore, it is crucial that youth care professionals correctly identify situations that are unsafe for children, and assess future risks. On the other hand, it is also important not to wrongly identify situations as unsafe or risky. Being accused of child abuse is very serious (Hacking 1992), and false accusations may be traumatic and harmful to families.

Evaluating children's safety and future risks is difficult, due to both the characteristics of the decision situation and professionals' limited cognitive capacities (Gambrill and Shlonsky 2000; Hardman 2009; Munro 1999). Information concerning a child's safety and risks is often incomplete and conflicting. Families may not be very willing to talk about their problems, out of shame or fear that a child will be removed from home (Dumbrill 2005; Munro 1999, 2008). Decisions are often made under time pressure, especially when a child's safety seems seriously threatened, and they can have far reaching consequences for children and families. At the same time, evidence about the effects of different interventions is scarce and mixed (e.g. Davidson-Arad 2005, 2010; Doyle 2007; Pinto and Maia 2013). Further, the media expose serious consequences of wrong decisions, making professionals' task also a public concern (Camasso and Jagannathan 2013).

Limited cognitive capacities further impede decision making. People select and combine information using heuristics that may lead to inconsistency and biases (Hardman 2009). In her analysis of errors in child protection work, Munro (1999) found that social workers are slow to revise their initial risk assessments, and that they rely on a limited amount of evidence, often evidence that is vivid and recent, and not necessarily the most relevant. Also, errors in communication of case information occur. Professionals need to rely on personal values, experiences and heuristics, since there are hardly any empirical guidelines (Arad-Davidson and Benbenishty 2008; Enosh and Bayer-Topilsky 2014). Unsurprisingly, professionals are found to reach different safety judgments and decisions for the same cases (e.g. Bartelink et al. 2014; Lindsey 1992; Schuerman et al. 1999).

### Risk Assessment Instruments

To improve judgments of children's safety and risks of abuse and neglect by professionals in child protection, risk assessment instruments have been developed that structure the judgment process. A distinction can be made between *consensus-based instruments* and *actuarial instruments* (Baird et al. 1999; D'Andrade et al. 2005; Gambrill and Shlonsky 2001). Consensus-based instruments present cues that experts have indicated as relevant based on empirical findings, theoretical literature, and practical knowledge, and ask users to evaluate these cues in a systematic, ordered way. Actuarial instruments present cues that haven been found, in experimental studies, to be predictive of outcomes. They ask users to systematically assess these cues, and they subsequently apply an algorithm that uses the empirically established weights of these cues to reach a conclusion.

There is not much evidence about the accuracy of safety judgments in child protection. There are only a few comparisons between the quality of safety and risk assessments made with different instruments, or between safety and risk assessments made with an instrument (structured judgments) or without instruments (unstructured judgments).

One of the few relevant studies compared the interrater reliability of judgments made with one actuarial instrument and two consensus-based risk assessment instruments (Baird et al.

1999). The research team found that agreement of risk judgments made with the actuarial instrument was moderate (average Cohen's kappa of .56), and poor for the consensus-based instruments (average Cohen's kappa of .18 for both instruments). Barber et al. (2007) similarly found that agreement in risk judgments made with a consensus-based instrument was rather poor (Cohen's kappa between .22 and .33). The actuarial instrument was also found to yield more valid risk judgments: children who had been assessed to be more at risk were later indeed reported for child maltreatment more often, their maltreatment was substantiated more often, and they were placed out of home more often (Baird and Wagner 2000). However, the predictive validity of the actuarial instrument was still limited (in the high risk group, new investigation rates were 46%, new substantiation rates 28%, and new placement rates 7%). Finally, Baumann et al. (2005) found that predictions made with actuarial models were not more valid than unstructured clinical predictions (however, for criticism on their methods see Johnson 2006). Otherwise, for most instruments there does not seem to be evidence about reliability or validity (D'Andrade et al. 2005; Gambrill and Shlonsky 2001), and there is as yet no agreement whether consensus-based or actuarial instruments are the most useful for assessing risks (e.g. White and Walsh 2006).

## The LIRIK

In this study, we evaluated the interrater reliability and predictive validity of the LIRIK, a Dutch instrument for child safety and risk assessment. This instrument aims to help reach conclusions on the actual safety of children and possible future risks of child maltreatment in the family situation (Ten Berge et al. 2014a, b). The Netherlands Youth Institute (Nederlands Jeugdinstituut) constructed the LIRIK in 2007, and revised it in 2014. While originally designed to be used in Youth Care Agencies (Bureaus Jeugdzorg) and Advice and Reporting Centers for Child Abuse and Neglect (ARCCANs), the LIRIK is increasingly used in other organizations, such as large organizations for both ambulant and residential youth care, organizations for (mentally) disabled children, and in general preventive youth health care. The LIRIK aims to assess risks in family situations for children aged 0–18 years, incurred from primary caretakers.

The LIRIK is based on a broad literature search for risk and protective factors and signs of child abuse and neglect. This search was combined with an analysis of the different steps in the process of judging about potential child maltreatment, and professionals' practical knowledge (Ten Berge and Vinke 2006a, b). The instrument is consensus-based, and systematically addresses relevant cues: factors in parent–child interaction, child signs, and risk factors and protective factors of parents, the child, the family and its social environment. The professional is asked to check all these cues and to come to three main conclusions: a conclusion about the child's current safety, and two risk assessments, one for the present situation and one taking into account foreseeable changes in the near future. Professionals are also asked to explain their safety judgment and both risk judgments. Thus, the LIRIK aims to structure and explicate professionals' risk assessments, while it leaves the weighting and combination of these cues to the individual professional. Training in the LIRIK is recommended and guidelines for implementation are provided in order to stimulate that professionals use the LIRIK correctly.

## This Study

Clinical judgments and decisions about child abuse and neglect are fallible and the public increasingly demands for public accountability of child protection work. In practice, more need is expressed for instruments that may lead to more objective judgments. Given the possibly high expectations about instruments' potential contributions to objective, efficient professional decision making, it is very important that the quality of instruments that are used to assess risks for child safety is investigated. To be really informative, it should be established whether the use of specific instruments leads to better judgments than the use of other instruments, or no instrument (unstructured judgements). This might prevent unwarranted confidence in (instruments to support) judgements that can affect people's lives tremendously (Camasso and Jagannathan 2013).

Evaluations have shown that youth care professionals believe that the LIRIK supports decision making about safety and risks, also for clients from a mentally challenged population (Ten Berge and Meuwissen 2013; Ten Berge and Van Rossum 2009). However, its psychometric qualities have not yet been established. Other Dutch risk assessment instruments also still lack good empirical evidence about their quality. Specifically, there is no evidence yet about their performance in comparison with other instruments or with unstructured judgments (Bartelink and Kooijman 2013). In two studies, we investigated the interrater reliability of professionals' judgments of child safety and risks made using the LIRIK (do different professionals agree about safety and risk for the same cases?) and their predictive validity (how well do safety and risk judgments predict future abuse and neglect?). To learn whether the use of the LIRIK has incremental value above unstructured judgments, we compared interrater reliability and predictive validity of judgments made with the LIRIK with that of judgments made without support of an instrument.

In "Study 1" section, we investigated the agreement of professionals' judgments for case vignettes. We also assessed agreement about the evaluation of the underlying cues. The LIRIK offers relevant factors to consider when judging child safety and risks, but it does not prescribe how to weigh them. Thus, although we did not expect interrater agreement about safety and risk conclusions to be very high, we did expect that the systematic evaluation of relevant cues would result in more reliable judgments than when no instrument was used. There is less research into agreement about individual risk items than about conclusions, so we had no clear expectations here (but see Barber et al. 2007; Orsi et al. 2014).

In "Study 2" section, we used a prospective design to investigate the predictive validity of safety and risk judgments, again comparing those made with the LIRK to unaided judgments. We related professionals' safety and risk judgments for clients they had seen in their own practice to outcomes of actual unsafety of these same clients 6 months later. A problem inherent to this design is that when an unsafe or risky situation is identified, this will lead to an intervention to improve that situation and less unsafety or risk will be detected after 6 months. Further, family factors that affect safety may change in time, preventing negative outcomes or worsening the situation. Predictive validity may thus be compromised by what happens in the intervening months.

To summarize, our main research questions for study 1 were to what extent youth care professionals agree about safety and risk judgments made with the LIRIK, and whether the interrater agreement of professionals' judgments about child safety and risks is higher when professionals use the LIRIK than when they do not use a risk assessment instrument (i.e. unstructured judgments). Our research questions for study 2 were to what extent safety and risk judgments made with the LIRIK predict later child maltreatment reports, child protection orders, safety interventions, and out-of-home placement, and whether the predictive validity of safety and risk judgments made with the LIRIK is better than that of unstructured safety and risk judgments.

## Study 1

### Method Study 1

#### *Participants*

Professionals were recruited from three youth care organizations, two large and one smaller organization. Managers of the two large organizations agreed to each ask 50 social workers and behavioural scientists whom they knew made safety and risk assessments in their practice, in different ambulant and residential departments, to participate during office hours. From the third, smaller organization a manager similarly recruited 8 professionals. We received 106 email addresses. Professionals were assigned to one of two groups in which they were asked to make safety and risk judgments for case vignettes, using the LIRIK (LIRIK group) or not using a risk assessment instrument (control group). Most professionals already worked with the LIRIK. Those who did not were assigned to the control group, and other participants were randomly assigned to the two groups. Of the 49 professionals assigned to the LIRIK group, 36 (73%) actually participated. Of the 57 professionals assigned to the control group, 43 (75%) participated. Reasons that professionals gave for not participating were lack of time, termination of contract, or illness.

Table 1 shows the gender, mean age, education, work experience, and experience and training with the LIRIK for the two groups. In the Netherlands, youth care professionals are often female (about 75%), and 49% is aged between 35 and 55 years, 38% is younger and 14% is older (Hollander et al. 2013). Dutch youth care professionals have mostly higher professional levels of education (63%), followed by academic levels (18%) (Hollander et al. 2013), so the education level in our sample is higher. This difference can be explained by the fact that we only included professionals who made risk assessments on a regular basis, and professionals with lower education levels do not usually make risk assessments. We had expected but did not find participants in the control group to have had less training and less experience with the LIRIK than participants in the LIRIK group. The participating institutions are from three different regions, and seem representative for Dutch youth care organizations.

#### *Procedure*

Participants received a link to an online questionnaire. In the LIRIK group, the LIRIK was part of the online questionnaire. The questionnaires contained six case descriptions and each case had questions about child safety and risks. First an instruction was given, explaining that the questions should be answered for one child, not other children in the family (if any); that participants could stop at any time to continue at a later moment; and that they should not discuss the case vignettes with colleagues to ensure independence of judgments. While answering the questions, participants always had access to the case vignettes in a separate window. After each case description we asked participants to rate the severity of the case in comparison with other cases in their daily practice.

Pilot testing showed that questionnaires with the LIRIK would on average take 2 h to complete (six cases). Questionnaires without the LIRIK took less time. Not all participants filled in the complete questionnaires. Two participants indicated that they had technical problems, and one participant indicated that he did not have enough information to answer the questions.

**Table 1** Participants' descriptives for the LIRIK and no-LIRIK conditions

| | LIRIK<br>($n = 36$) | | No LIRIK<br>($n = 43$) | | Test for differences |
|---|---|---|---|---|---|
| Female | 85.3% | (29/34) | 84.2% | (32/38) | $z = .13, p = .90$ |
| Age | $M = 40.7$ | $SD = 9.7$ | $M = 38.8$ | $SD = 10.0$ | $t(70) = .77, p = .44$ |
| Education | | | | | |
| Middle | 3.1% | 1 | 5.4% | 2 | |
| Higher | 46.9% | 15 | 48.6% | 18 | |
| Academic | 50.0% | 16 | 45.9% | 17 | $\chi^2 = .28, df = 2, p = .87$ |
| Experience | | | | | |
| In youth care | $M = 14.2$ | $SD = 8.7$ | $M = 12.2$ | $SD = 7.6$ | $t(70) = 1.04, p = .30$ |
| In function | $M = 6.9$ | $SD = 3.6$ | $M = 7.6$ | $SD = 4.4$ | $t(70) = -.79, p = .43$ |
| Training LIRIK | 59.3% | (16/27) | 70.8% | (17/24) | $z = -.86, p = .39$ |
| LIRIK used since | | | | | |
| ≥1 year | 48.1% | 13 | 45.8% | 11 | |
| 6–12 months | 22.2% | 6 | 29.2% | 7 | |
| 1–6 months | 7.4% | 2 | 12.5% | 3 | |
| ≤1 month | 22.2% | 6 | 12.5% | 3 | $U = 315.5, z = -.17, p = .86$ |

Frequencies do not add up to the total number of participants in each group, because not all participants had filled in (all) questions about personal descriptives

### *Materials*

Case descriptions were summaries of 521–732 words ($M = 607.5$) of real reports of 12 clients of two large youth care organizations. The cases were adapted to make them anonymous. We chose cases such that we had an equal number of boys and girls, of varying ages (5–16 years), different family situations (one/two parents, brothers/sisters/no siblings), different social and cultural backgrounds (e.g. poor/good SES, different ethnic backgrounds), and with varying type and severity of problems (suggesting physical, emotional or sexual abuse or neglect).

The twelve case descriptions were grouped into six combinations of six cases, such that each participant rated six different cases, each case would be rated equally often, and in different orders. The order of the cases in each combination was randomized. Each case was rated by 13–17 different raters in the LIRIK group and by 11–19 different raters in the control group.

In the LIRIK group, participants completed the LIRIK after reading a case description. The LIRIK contains items that ask for the presence of specific signs in parent–child interaction, child signs, risk factors, and protective factors of parents, the child, the family and its social environment (for all items, see the "Appendix"). Based on these cues, the professional is asked to give three main conclusions. First, a conclusion about the current safety of the child, with the options: the child seems currently safe, child maltreatment is a possibility, child maltreatment is substantiated, the child is in direct physical danger, or information is insufficient to reach a conclusion. This scale thus seems to imply both seriousness and certainty, what may explain why professionals are allowed to choose more than one option here (for a critical reflection on this scale, see the "General Discussion" section). After checking additional risk and protective factors, professionals are asked to provide two conclusions about the child's risk: in the current situation, and in the near future in the light of foreseen changes, both on a scale from low (1) to very high (5). In the control group, participants are only asked to give the three judgments of safety and risks for each case description (without the preceding structured assessment of cues).

### *Analyses*

First, we calculated the interrater agreement about the three main conclusions for child safety and current and future risks for both groups. Next, we looked at the interrater agreement in the LIRIK group about the cues. We performed additional analyses for the agreement about the main conclusions for participants who had followed training (see "Results" section).

The first judgment, of current child safety, posed two problems for the analysis of agreement. First, in line with actual use of the LIRIK in practice, participants were allowed to choose more than one option. Second, while the first four options formed an ordinal measurement scale indicating little to serious safety threats (but see our remarks in the "Discussion" section), the fifth category (insufficient information) was different. To deal with these problems, we transformed participants' safety judgments into a variable with only one value, or a missing value. First, we looked whether participants had chosen more than one option. If so, and if one of two chosen options was 5 (insufficient information), we transformed participants' answers into the other category chosen (disregarding the choice of the fifth category).[1] In those cases that two or three other categories than 5 were chosen, or only 5, we created a missing value. This strategy resulted in a loss of 10.7% of the judgments.[2]

We used Krippendorff's alpha ($\alpha$) as measure for interrater agreement about the three main judgments of safety and risks (Hayes and Krippendorff 2007). It expresses absolute agreement (whether judges reach the same ratings for the same case) on a scale of 0 (no more than chance agreement) to 1 (perfect agreement), treating coders as interchangeable. We also calculated bootstrap 95% confidence intervals for the alphas (1000 samples).

For good agreement Krippendorff's alphas should be at least .80 (Krippendorff 2004).[3] Other authors similarly indicate that .80 is a minimum for good agreement, or .60 for sufficient agreement (see e.g. Cichetti 2001; Evers et al. 2010), although whether agreement is good or sufficient depends on what the data are used for.

[1] The fifth category, lack of information, can be considered to be an uninformative category with respect to participants' safety judgments. Uninformative categories in a coding system are best replaced by informative values, or otherwise excluded from analyses of agreement (Krippendorff 2011).

[2] To control for biases that might result from our recoding, we performed two additional transformations, that excluded no judgments. When two or three other categories than 5 were chosen, for one transformation we took the highest option, for the other transformation the lowest option. In both groups, calculation of agreement with the three different transformations did not result in significantly different alphas (see "Results" section, and Bartelink et al. 2015).

[3] We had 181 judgments in the LIRIK group and 183 judgments in the non-LIRIK group. These numbers are sufficient for reliably deciding ($p \leq .05$) whether Krippendorff's alpha would exceed a minimum of .80 [see Krippendorff (2011): Table 1, p 105]. For reliably deciding on lower minimum levels fewer judgments are needed.

**Table 2** Krippendorff's alphas for interrater agreement (with bootstrap 95%-confidence intervals) for participants that had and had not used the LIRIK in the study

| | LIRIK ($n$ = 36) | No LIRIK ($n$ = 43) |
|---|---|---|
| Conclusion safety | .48 (.40, .56) | .42 (.33, .50) |
| Risk assessment now | .39 (.33, .45) | .46 (.40, .51) |
| Risk assessment with changes foreseen | .19 (.10, .26) | .25 (.18, .32) |

**Table 3** Krippendorff's alphas for interrater agreement (with bootstrap 95%-confidence intervals) for participants that had followed a training, and that used and did not use the LIRIK in study 1

| | Training and LIRIK ($n$ = 16) | Training ($n$ = 33) |
|---|---|---|
| Conclusion safety | .40 (.32, .49) | .45 (.37, .52) |
| Risk assessment now | .37 (.30, .44) | .44 (.39, .51) |
| Risk assessment with changes foreseen | .29 (.21, .37) | .23 (.16, .31) |

## Results Study 1

Participants did not rate the cases as much more or less serious than cases in their daily practice. On a scale of 1 (much less serious) to 5 (much more serious) the mean ratings for the cases (averaged over participants) varied between 2.13 and 3.50 ($M = 2.95$, $SD = .46$) in the LIRIK group and between 1.74 and 3.47 ($M = 2.83$, $SD = .66$) in the control group (no statistically significant difference between groups, $t(22) = .59$, $p = .56$).

### *Interrater Agreement About Judgments of Child Safety and Risks*

Table 2 shows Krippendorff alphas (α) and bootstrap confidence intervals for the safety judgments and the judgments of current risks and risks in the near future. We found that agreement about children's current safety and risks never reached .50, and bootstrap confidence intervals also excluded values of .80 or .60; thus, agreement should be considered poor to moderate at best, in both the LIRIK and the control group. Assessment of future risks with foreseen changes taken into account was the least reliable. Different participants rated the same child's safety situation as safe or low risk to (seriously) dangerous.

Agreement about current safety was somewhat higher for participants who had used the LIRIK. Agreement about risks, with or without taking foreseen changes into account, was better for participants who had not used the LIRIK. Results are similar if only participants who had been trained to use the LIRIK are included in the analyses, see Table 3.

### *Interrater Agreement about the Presence of Relevant Cues*

The LIRIK is designed to help professionals assess the presence of cues that, according to the (empirical) literature, are related to child safety and risks for child abuse and neglect. It asks whether a cue is present, absent, or unknown. In practice it may be useful to see what is relevant but unknown, thus what information is still needed, but in our study we were interested in whether participants reliably identified these cues in the available information.

For most items agreement was rather poor. It turned out that for 22 (of 75) items alpha was lower than .20, for 30 items it was between .20 and .40, for 14 between .40 and .60, for five between .60 and .80, and only for four items alpha was .80 or higher (for the results for specific items, see the "Appendix").

To conclude, we found that agreement about safety and risks was insufficient, both when participants had used the LIRIK, and when participants had not used an instrument. In "Study 2" section, we investigated the predictive validity of safety and risk judgments, both of those made with the LIRIK and of unstructured judgments, by comparing them with outcomes indicating substantiated maltreatment after (approximately) 6 months.

## Study 2

### Method Study 2

#### *Participants*

Managers of two large youth care organizations (the same as in Study 1), and one child health care organization agreed to collect 320 safety and risk assessments for children made with the LIRIK, and 120 assessments made without an instrument (see "Procedure" section),[4] from case files completed at case opening or during the treatment process between September 2013 and November 2014. The collection stopped a bit earlier due to time constraints, when 428 assessments had been selected. Assessments for children in residential care at the moment of the risk assessment were excluded from the study, because the LIRIK is meant for judging safety and risks for children living with biological, foster, or adoptive parents. The study sample consisted of 370 children from three different agencies (two child welfare agencies and one child health care agency). For 278 children (168 boys) the LIRIK had been completed (LIRIK group), for 92 children (52 boys) an unstructured risk assessment was obtained (control group). The parents of included children had given permission for use of the case files for research purposes.

The proportion of boys and girls was not significantly different in the two groups ($\chi^2$ (1) = .26, $p$ = .62). The mean age in the LIRIK group was 9.6 years ($SD$ = 5.1 years). The mean age of children in the control group was significantly higher (11.7 years; $SD$ = 5.0 years; $t$ (367) = 3.35, $p$ = .001).

#### *Procedure*

Professionals had given the three main conclusions about safety and risks: current safety, current risk of child maltreatment, and risk of maltreatment with foreseen changes taken into account, by filling in the LIRIK (LIRIK group) or just a short form to complete the conclusions plus a question to describe the rationale for their conclusions (control group).

Six months after the initial safety and risk assessments, three raters (first author and two research assistants) independently coded the presence of the following outcomes in the case files of these children: child reported at an ARCCAN, child protection investigation by the Child Protection Board, child protection order, out-of-home placement, another specific safety intervention, or crisis intervention. All outcomes were coded as 0 (no) or 1

[4] The use of the LIRIK was a standard procedure within the agencies. To collect data for the control group, they adjusted their procedure temporarily.

(yes). We constructed an additional variable "unsafe outcome" to indicate that at least one of the outcomes mentioned above was present in a case (0 = no to all outcomes, 1 = yes to one or more outcomes). The case files were coded using a standardized checklist. After testing and small adaptations, Cohen's kappa ranged between .65 and 1.00, which can be considered sufficient to good (Landis and Koch 1977). The mean period between the risk assessment and the follow-up measurement (T2) was 225 days ($SD = 34$ days) in the LIRIK group and 201 days ($SD = 52$ days) in the control group, which was significantly shorter ($t$ (368) = −5.06, $p < .001$).

### *Analyses*

First, we calculated percentages for the safety and risk judgments (reported in Table 5), and for all T2 outcome measures (reported in Table 4), and performed Chi squared tests to check for differences between the LIRIK group and control group.[5] To analyse the relationship between the safety judgments and the outcomes, we excluded those cases in which professionals chose more than one option to indicate child safety (and one of these was not the option 'insufficient information'), or in which they chose the option insufficient information only, as in "Study 1" section. In the LIRIK group, 15 cases had to be excluded for this reason, and 9 cases in the control group. In the LIRIK group this resulted in the exclusion of all (3) cases in which professionals concluded that a life-threatening situation existed. For a fair comparison, the cases (4) in the control group in which professionals concluded that a life-threatening situation existed were also excluded from further analyses. Judgments of safety and risks were scored 1–4, a higher score meaning more unsafety or higher risks.[6] Spearman rank correlations and bootstrap confidence intervals were calculated to examine the relationship between judgments and the outcomes at T2. A correlation between .70 and 1.00 is interpreted as indicating a very strong relationship, between .50 and .69 strong, between .30 and .49 moderate, between .10 and .29 weak, and between .00 and .09 a negligible relationship (see also Cohen 1988; Hinkle et al.2003). Because correlations between a dichotomous variable and a variable with a four point scale are maximized, we also calculated maximum correlations (the highest possible values, given the categories).

To see in more detail to what extent unsafety outcomes occurred when professionals had judged cases as more unsafe or at higher risk, we additionally related the safety and risk judgments, ranked by level of threat, to the outcome measure "any unsafe outcome" at T2. The variable "any unsafe outcome" is a dichotomous variable that reports whether at least one outcome occurred in a case. Finally, we examined the relationship between separate LIRIK items and an unsafe outcome at T2 using Spearman rank correlations.

[5] We found no difference between the LIRIK group and the control group, except for the safety judgments and number of out-of-home placements. The safety judgments of the control group were statistically significantly higher than those of the LIRIK group, meaning that professionals in the control group judged cases to be more unsafe ($\chi 2$ (3) = 10.52, $p = .02$). In the control group children were significantly more often placed out-of-home than in the LIRIK group ($\chi 2$ (1) = 14.84, $p = .00$).

[6] The LIRIK was revised just before the studies started, in response to youth care professionals' feedback. Differences between the earlier and the latest version consist of textual changes and changes in item ordering. The risk judgments changed from a four point to a five point scale. For the validity study, professionals' assessments of 6 months earlier were collected, therefore the risk judgments were on a four point scale here.

**Table 4** Spearman's rank correlations [with 95% percentile bootstrap confidence intervals] (and maximum correlations) for the relationship between judgments about safety and future risks, and outcomes on T2

| | Child maltreatment report | Child protection investigation | Child protection order | Out-of-home placement | Safety intervention | Crisis | Unsafe outcome[a] |
|---|---|---|---|---|---|---|---|
| LIRIK | | | | | | | |
| Current safety | **.10** [−.04, .24] (.55) | **.03** [−.11, .19] (.47) | **.24** [.07, .41] (.62) | **−.01** [−.13, .13] (.51) | **.07** [−.07, .22] (.69) | **.15** [−.02, 31] (.58) | **.31** [.17, .46] (.91) |
| Risk in foreseeable future | **.14** [−.13, .28] (.51) | **.04** [−.09, .16] (.41) | **.21** [.06, .35] (.63) | **.08** [−.04, .19] (.44) | **.20** [.06, .33] (.69) | **.12** [.00, .26] (.54) | **.37** [.24,.50] (.90) |
| Risk with changes foreseen | **.07** [−.10, .23] (.58) | **−.03** [−.17, .15] (.49) | **.14** [−.03, .31] (.70) | **.02** [−.13, .17] (.52) | **.04** [−.10, .21] (.76) | **.08** [−.08, .23] (.62) | **.17** [.01, .33] (.90) |
| Base rate[b] | 6.1 | 4.0 | 9.4 | 4.8 | 12.2 | 7.2 | 30.0 |
| Unstructured judgment | | | | | | | |
| Current safety | **.06** [−.12, .26] (.38) | **.06** [−.12, .27] (.48) | **.23** [.01, .47] (.69) | **.44** [.21, .64] (.71) | **.26** [.01, .46] (.60) | **.16** [−.08, .40] (.58) | **.43** [.20, .64] (.96) |
| Risk in foreseeable future | **.04** [−.12, .25] (.35) | **.11** [−.10, .30] (.45) | **.05** [−.17, .27] (.70) | **.26** [−.01, .50] (.72) | **.37** [.13, .56] (.61) | **.15** [−.07, .38] (.57) | **.39** [.18, .59] (.96) |
| Risk with changes foreseen | **.05** [−.11, .27] (.36) | **.04** [−.14, .32] (.47) | **.12** [−.11, .35] (.73) | **.35** [.09, .57] (.75) | **.21** [−.07, .45] (.63) | **.20** [−.08, .44] (.59) | **.28** [.04, .49] (.93) |
| Base rate | 3.3 | 5.4 | 15.2 | 17.6 | 9.8 | 8.7 | 33.3 |

[a] The variable "unsafe outcome" is a dichotomous variable that reports whether at least one outcome occurred in a case

[b] The base rate is the percentage of cases in which an outcome was found in a group (LIRIK or control group)

* $p < .05$ (two-sided); ** $p < .01$ (two-sided)

## Results

### *Relations Between Safety and Risk Judgments and Outcomes—Rank Correlations*

Table 4 shows the correlations and bootstrap confidence intervals for the judgments of current safety and risks and the outcomes at T2, and the maximum possible correlations. In the LIRIK group we found weak to moderate relationships, and in the control group moderate to strong relationships, between the level of judged safety and risk and the report of an order for child protection, a crisis, or any unsafe outcome. Correlations were higher in the control group than in the LIRIK group, except for child protection orders, although the wide bootstrap confidence intervals indicate large uncertainties.[7]

### *Relations Between Safety and Risk Judgments and Outcomes*

Table 5 shows a more detailed comparison between the safety and risk judgments, split by level of threat, and the outcome measure "unsafe outcome", for the LIRIK group and the control group. It also includes the base rates of the outcome in each group. For example, 23.9% of all children in the LIRIK group had an unsafe outcome; this is the base rate. The predictive validity of judgments is better if the percentage of unsafety outcomes increases with increasing levels of unsafety, that is: the more unsafe a child is judged to be, the (relatively) more often unsafety outcomes are expected to be present. In the LIRIK group, a consistent increase in the percentages of unsafe outcomes was found with judgments of increasing unsafety and risks. In the control group, no consistent increase was found. Further, the percentage of unsafety outcomes for children judged to be safe should be lower than the base rates of the outcomes, and the percentage of outcomes for children judged to be (possibly) unsafe should be higher than the base rates of the outcomes. For example, 15.5% of all children in the LIRIK group who had been judged to be safe at home (not maltreated) did not have any unsafe outcome. Percentages of outcomes for children assessed to be safe or not at risk were all below the base rate. Percentages of outcomes for children assessed to be unsafe or at risk were all above the base rate.

To summarize, judgments were often very poor predictors of specific outcomes, but they did predict weakly to moderately whether any unsafe outcome occurred at all. Differences between the LIRIK group and the control group were small, and while rank correlations between judgments and outcomes seemed higher for the control group, the LIRIK group seemed to have a slightly more consistent increase of unsafe outcomes when safety threats were judged to be higher.

### *Relationships Between Separate LIRIK Items and Outcomes*

We also looked at the correlations between the LIRIK items preceding the three main conclusions and an unsafe outcome at T2 (see "Appendix"). The majority of these correlations was low ($\rho < .09$), meaning that there was barely any relationship between LIRIK items and an unsafe outcome. Other items, about parent–child interaction, child signals and some risk factors of the parents, the child and the family and environment, correlated weakly ($.10 > \rho < .29$) to moderately ($.30 > \rho < .49$) with an unsafe outcome.

[7] The relations between judgments and outcomes were the same when the cases were included in which the professional concluded that there was a life-threatening situation.

**Table 5** Percentages of the outcomes per level of judgment about current safety and future risks

| Current unsafety | Case distribution[a] | Unsafe outcome[b] | Risk in foreseeable future | Case distribution[a] | Unsafe outcome[b] | Risk by changes foreseen | Case distribution[a] | Unsafe outcome[b] |
|---|---|---|---|---|---|---|---|---|
| LIRIK | | | | | | | | |
| No sign of child maltreatment | 47.8 | 15.5 | Low | 49.6 | 16.7 | Low | 32.1 | 24.4 |
| Child maltreatment is possible | 23.4 | 29.6 | Real | 20.5 | 38.6 | Real | 19.2 | 29.3 |
| Child maltreatment is substantiated | 9.4 | 78.6 | High | 9.0 | 72.0 | High | 7.0 | 50.0 |
| | | | Very high | 1.8 | 40.0 | Very high | 2.7 | 45.5 |
| Base rate[c] | | 23.9 | | | 28.9 | | | 30.0 |
| Unstructured judgment | | | | | | | | |
| No sign of child maltreatment | 63.0 | 15.5 | Low | 65.2 | 20.0 | Low | 64.1 | 23.7 |
| Child maltreatment is possible | 26.1 | 63.6 | Real | 22.8 | 57.1 | Real | 19.6 | 44.4 |
| Child maltreatment is substantiated | 5.4 | 33.3 | High | 9.8 | 55.6 | High | 9.8 | 66.7 |
| | | | Very high | 1.1 | 100 | Very high | 1.1 | 100 |
| Base rate | | 28.9 | | | 33.3 | | | 33.3 |

[a] The variable "case distribution" represents the percentage of judgments on current safety, risk in foreseeable future, and risk with changes foreseen. Note that the percentages do not sum up to 100%. In the LIRIK group, 19.6% of the judgments about current safety, 19.1 of the judgments about risk in the foreseeable future, and 39.0% of the judgments about risk with changes foreseen taken into account are missing. In the control group, 5.5% of the judgments about current safety, 1.1 of the judgments about risk in the foreseeable future, and 5.4% of the judgments about risk with changes foreseen taken into account are missing

[b] The variable "unsafe outcome" is a dichotomous variable that reports whether at least one outcome occurred in a case

[c] The base rate is the percentage in which an unsafe outcome was found in a group (LIRIK or control group)

## General Discussion

The LIRIK aims to support professionals in reaching judgments about a child's safety and future risks by asking them to systematically check cues that indicate (risks of) child abuse and neglect. Safety and risk judgments may have far-reaching consequences for the children and families involved, both in cases in which actual risks are not identified, and in cases in which supposed risks are not actually present. For this reason, but also in light of general guidelines for agreement (see e.g. Cichetti 2001; Evers et al. 2010; Krippendorff 2011), we conclude that the agreement about the safety and risk judgments we found in Study 1 was insufficient, both when the LIRIK was used and when it was not used. In "Study 2" section, we found that professionals' safety and risk judgments moderately predicted an unsafe outcome over a period of 6 months, both those made with and made without the LIRIK. So, we found no evidence that using the LIRIK leads to better, i.e. more reliable and more valid judgments, than when no instrument is being used.

Our results are not unique. One study (Van der Put et al. 2016) found that the LIRIK poorly predicted re-occurrence of child rearing problems in families that already received help (AUC = .53). Other risk assessment instruments also show, if their psychometric qualities have been studied at all, limited reliability and validity (D'Andrade et al. 2005; Baird and Wagner 2000; Danktert and Johnson 2013; De Ruiter et al. 2012; Gambrill and Shlonsky 2001; Johnson 2011; Shlonsky and Wagner 2005; Van der Elst et al. 2012). Actuarial instruments are generally found to outperform consensus-based instruments, but they too have disappointing reliability and validity (D'Andrade et al. 2005). Only a few other researchers have compared an actuarial risk assessment instrument and unstructured clinical judgment and they showed mixed results (Baumann et al. 2005; Johnson 2011).

There are several explanations for the low reliability and predictive validity of risk assessment instruments. Disagreement among professionals may be due to differences in assumptions they make, hypotheses they generate, or additional information they need (Mandel et al. 1994). It may also be caused by different personal decision thresholds (Baumann et al. 2011; Dalgleish 2000; Schuerman et al. 1999), or professionals' different cultural contexts (Gold et al. 2001). We also found that agreement was poor for the individual cues. This is also in line with other findings (Barber et al. 2007: Orsi et al. 2014). If professionals disagree about the presence of relevant cues, they are also likely to disagree about conclusions that are based on these cues (cf. Baird et al. 1999; Shlonsky and Wagner 2005). If professionals disagree about safety and risks for a child, the child may receive different interventions, or perhaps no intervention, dependent on the specific judge.

Reliability and validity of judgments made with (actuarial and consensus based) risk assessment instruments will be compromised by unclear definitions of what exactly is and is not child abuse or neglect. Inconsistency of judgments, whether by differences in definitions, thresholds, or assumptions, interferes with the assessment of validity (Camasso and Jagannathan 2000, 2013; Gambrill and Shlonsky 2000). Also, risk assessment instruments may not be sensitive enough to assess changes in safety and risks, because they—in particular actuarial instruments—rely on static factors while risks have a dynamic nature. Also important is that although risk factors have been identified in empirical research, there is little knowledge about their actual impact in individual families (Gambrill and Shlonsky 2000; Munro 2014; Rycus and Hughes 2003). The presence of risk factors seems to be neither a necessary nor a sufficient condition for child maltreatment to occur. It is not even known what specific combination(s) of risk factors may lead to child maltreatment (Munro 2014). Some researchers found evidence that an accumulation of risk factors predicts child maltreatment better than the presence of specific single risk factors (Begle et al. 2010; MacKenzie et al. 2011). A further complication is that professionals may not be familiar with relevant literature, or do not learn from their experience, because they lack feedback about the accuracy of their judgments (Dawes et al. 1989; Finnila et al. 2012).

The LIRIK helps professionals to systematically assess relevant cues when judging child safety. It is thus assumed that professionals deliberate about risks. However, it seems cognitively demanding to deliberately consider so many cues. In practice, professionals might make intuitive assessments of risk, based on a holistic impression about the safety of a family situation (Munro 2005). It is uncertain whether the use of a risk assessment instrument can or cannot in fact improve unaided judgments (Baumann et al. 2005; cf. Kahneman and Klein 2009).

## Limitations

In "Study 1" section, we used case vignettes, and participants could not ask for additional information. In practice, professionals may have more information than is presented in vignettes, which may lead to more reliable judgments (see Barber et al. 2007). On the other hand, more information does not necessarily result in better predictions, as more information implies the necessity to integrate more cues, which makes the task more complex, or might distract attention from relevant information (Dana et al. 2013; Grove et al. 2000).

In both studies, professionals may not be fully representative of all professionals who make safety and risk assessments for children. Professionals were from three organizations, and in the control conditions, professionals were often already familiar with the LIRIK. Knowledge of the LIRIK may have affected their unstructured judgments and have led to smaller differences between the LIRIK-groups and the control groups, although it seems unlikely that professionals who gave judgments without the LIRIK did so in a similarly structured way as when actually filling in the LIRIK.

Several factors may weaken the relation between initial judgments of safety and risk, and later outcomes. First, as mentioned in the introduction, if situations are judged to be unsafe or high risk, interventions will be implemented to restore safety. Further, given the time span between the safety and risk assessments and the outcomes up to 6 months later, it is possible that outcomes occurred after more than 6 months. For example, the results of an investigation of a child maltreatment report, or the judgment of a magistrate of a juvenile court on the necessity of a child protection order may have come later.

Cases are regularly transferred from one professional to another and from one organization to another. In our study professionals who made the safety and risk judgments were not always the same professionals as those providing the help offered to families. This is very common in the Netherlands and it may also be the case in other countries. As a consequence, professionals may or may not adopt previous judgments of safety and risks made by another professional, and they can decide whether or not they will follow up the intervention recommendations made by the other. Research has repeatedly confirmed that intervention decisions do not only depend on case specific factors, but also for example on personal beliefs and attitudes of the professional, and knowledge about available interventions (Berben 2000; Lekkerkerker et al. 2011; Ten Berge 1998). Dalgleish and others have shown that most of the time the factors influencing risk assessments are case specific, while intervention decisions depend on factors related to the decision-maker, such as knowledge, skill, and experience (see Baumann et al. 2011; Dalgleish 1988, 2000, 2003).

Finally, it was a salient finding that case files often appeared to be incomplete. Notably, in the LIRIK group many conclusions on safety and risk were missing. This may also have affected our results. Possibly after filling in specific items, professionals felt that it was not necessary to further report their conclusions because they seemed self-evident from the items. Also, there were no reports of interventions, which one would expect in cases of serious threat. Missing information, which we often saw in our data when professionals did not give a conclusion or chose more options, may obscure the relation between judgments and outcomes. Specifically, if conclusions were missing because professionals felt too uncertain about them, our estimates of predictive validity may have been too optimistic. If, on the other hand, professionals did not report conclusions because they seemed so obvious to them (clear cases of maltreatment, or of safety), predictive validity could have been better had the conclusions been included.

## Implications

Implementation of risk assessment instruments such as the LIRIK is justified if they improve professionals' judgments, and thereby reduce risks and consequences of child abuse and neglect. Otherwise, they may just be an additional administrative burden, at the cost of time available for the children and families (Baumann et al. 2005; Munro 2005). The widespread use of the LIRIK in The Netherlands, and our disappointing results (within the studies' limitations), call for improvements.

Agreement about individual risk items in the LIRIK may be improved by providing clear decision criteria, especially in a digital version where these instructions may pop up with each question. An actuarial version of the LIRIK has been designed (Van der Put et al. 2016), but its ability to reliably and validly predict child maltreatment has not yet been determined. Also, the implementation of the LIRIK may be improved, given that so far implementation processes seem to have been lax (D'Andrade et al. 2005; Prins 2011). However, it seems impossible to meet participants' wishes for an objective risk assessment instrument, given the current state of knowledge about how families become abusive, and given poor definitions of abuse and of outcomes and the dynamic nature of risk (Gambrill and Shlonsky 2000; Rycus and Hughes 2003).

Another recommendation is that the use of three main conclusions should be reconsidered. The first scale, for current safety, seems to confuse certainty about whether a child is being maltreated and the seriousness of the (supposed) maltreatment, and this ambiguity may in itself induce disagreement and lack of predictive power. It might be reformulated such that a conclusion is being asked about the certainty of maltreatment, or the seriousness of the (supposed) maltreatment, or about both separately. Also, it can be doubted whether the third conclusion (future risks taking into account foreseeable changes) has additional value. This conclusion was meant to improve risk assessments if it could be foreseen that future events would change risks for a child. For example, a father coming back from prison may alter the safety situation. However, it seems unclear to what extent foreseeable changes might not already be taken into account in the conclusion about current risks, and interrater agreement and predictive validity was poor for this conclusion. Thus, it might be left out, perhaps with an explicit reminder to be alert to changes in the future that may affect safety.

Using risk assessment instruments such as the LIRIK may lead to overconfidence and less critical reflection on safety and risk judgments, specifically if professionals believe that they use an objective instrument (Regehr et al. 2010). Professionals are positive about the LIRIK's usefulness (Ten Berge and Meuwissen 2013; Ten Berge and Van Rossum 2009).

When professionals use an instrument, they should be aware of its limitations and realise the uncertainties that are inherent in the use of instruments, even if these are well validated (Gambrill and Shlonsky 2000; Hart et al. 2007; Regehr et al. 2010).

Because professionals were found to reach different safety judgments with the same information, we recommend that they should always be asked to explain their judgments. This recommendation seems even more important in the light of our finding that relevant safety information (conclusions, interventions) was so often missing, and specifically if professionals had filled in the LIRIK. Professionals may wrongly believe that their conclusions and decisions follow logically from the information that they have assessed (and reported) (cf. De Kwaadsteniet et al. 2013). Involving colleagues and supervisors might improve judgments, although here too there seems to be little evidence for improved judgements (but see Smithgall et al. 2015).

Assessing safety and risks in collaboration with the family, which some professionals indicated to prefer, might lead to more reliable and valid signalling of relevant factors. Families may feel more supported when professionals involve them in the assessment and may be more willing to participate in the assessment process. That may result in more openness to share about their situation, which may lead to better safety and risk assessments. Approaches such as Signs of Safety (Turnell and Edwards 1999), which emphasize collaboration between parents and professionals, are quite popular with professionals, but little is known about effects on the quality of the assessment. There is some evidence, that including clients in decision making might lead to better outcomes (cf. Golnik et al. 2012; Vis et al. 2011).

Despite the limitations of our study, which are to some extent inherent to research in practice settings, we conclude that one should be cautious to expect that structuring the process will have substantial effects on the reliability and validity of safety and risk judgments. Pending future improvements of the LIRIK, it should not be used in practice without explicit warning that it cannot be expected to result in objective judgments. We agree with others' recommendations (e.g. Munro 2005; Rycus and Hughes 2003; Shlonsky and Wagner 2005) that risk assessment needs to be viewed in the larger context of child protection, and that not only instruments but processes too should be evidence-based, and families should be involved in judgments and decisions that concern them. It is crucial that investments in improving professional assessment and decision-making result in better outcomes for children, i.e. less child maltreatment, safe homes, and effective interventions in child maltreatment cases.

**Acknowledgements** Stichting Kinderpostzegels Nederland sponsored this study. We thank all participants for their cooperation. We specifically thank Carolien Konijn (Spirit), Mariëtte Blik and Mariëtte Brandenburg (Juzt), Mariëlle Hornstra (GGD-Zuid-Gelderland) for their collaboration and for their insightful contributions. We thank Floor Minkhorst for designing the online questionnaires and analyzing participants' comments about the LIRIK, Raoul Hakkenberg for coding the case files, Pieter Groenestijn for support with Perseus and Pierre Souren for support with the statistical analyses, and all members of the research committee for their valuable feedback during the whole research project.

**Funding** This study was funded by Stichting Kinderpostzegels Nederland (Grant No. 8329641).

### Compliance with Ethical Standards

**Conflict of interest** Cora Bartelink was involved in the development of the LIRIK. Leontien de Kwaadsteniet declares that she has no conflict of interest. Ingrid ten Berge was involved in the development of the LIRIK. Cilia Witteman declares that she has no conflict of interest.

**Ethical Approval** All procedures were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

# Appendix

See Table 6.

**Table 6** Proportions of how often participants in the LIRIK group believed a specific factor was present, measured over cases and raters, and Krippendorff's alphas for interrater agreement (with bootstrap 95%-confidence intervals) for these items in Study 1, and Spearman's rank correlations between the items and any unsafe outcome in Study 2

| | Study 1 | | | Study 2 |
|---|---|---|---|---|
| | Proportion 'yes' | Alpha | Bootstrap 95% confidence interval | $R_s$ unsafe outcome |
| 1. Current safety | | | | |
| 1A. Direct safety | | | | |
| Serious threat by parent(s)/primary caretaker(s) | .30 | .40 | (.19, .57) | .07 |
| Serious threat by child himself/herself | .12 | .16 | (−.23, .48) | .01 |
| Serious threat by other family member/other person | .03 | .00 | (−.88, .81) | .13 |
| *Suspicions of/Signs for life threatening situation/physical danger?* | *.35* | *.36* | *(.17, .54)* | *.08* |
| 1B. Interaction parent(s)–child | | | | |
| Are there facts that indicate recent | | | | |
| Physical violence | .30 | .40 | (.20, .60) | .14 |
| Psychological violence | .49 | .28 | (.09, .48) | |
| Physical neglect | .03 | .08 | (−.73, .68) | .23 |
| Emotional neglect | .58 | .26 | (.06, .46) | |
| Sexual abuse | .02 | .09 | (−.87, .77) | |
| Witnessing domestic violence | .40 | .41 | (.21, .59) | .21 |
| Parenting | | | | |
| Protection and safety | .72 | .35 | (.13, .58) | .28 |
| Basic care | .15 | .23 | (−.08, .50) | .24 |
| Emotional warmth (support) | .71 | .37 | (.15, .59) | .16 |
| Rules and boundaries | .70 | .31 | (.08, .51) | .21 |
| Stimulation | .64 | .20 | (.01, .39) | .19 |
| Stability | .55 | .29 | (.11, .48) | .31 |
| *Signs for threats or neglect by parent(s)?* | * | *.29* | *(.17, .41)* | *.16* |
| 1C. Child | | | | |
| Psycho-social functioning | .91 | .06 | (−.42, .47) | .22 |
| Physical health | .15 | .15 | (−.15, .44) | .22 |
| Skills and cognitive development | .66 | .27 | (.07, .47) | .09 |
| *Child signals for child maltreatment?* | * | *.29* | *(.16, .43)* | *.25* |
| 1D. Risk and protective factors | | | | |

**Table 6** continued

| | Study 1 | | | Study 2 |
|---|---|---|---|---|
| | Proportion 'yes' | Alpha | Bootstrap 95% confidence interval | R$_s$ unsafe outcome |
| *Risk factors of the parent(s)* | | | | |
| Functioning as parent | | | | |
| Former abuse or neglect of a child | .18 | .18 | (−.12, .46) | .26 |
| Insufficient parenting knowledge and/or skills | .74 | .09 | (−.16, .33) | .20 |
| Problems in the parent–child interaction | | | | .21 |
| Playing down/Denial of substantiated child maltreatment | .27 | .10 | (−.16, .32) | .12 |
| Negative attitude towards the child | .40 | .44 | (.25, .63) | .15 |
| Personal functioning | | | | |
| Psychiatric problems | .32 | .47 | (.29, .65) | .09 |
| Addiction problems | .26 | .94 | (.85, 1.00) | .08 |
| Intellectual disability | .16 | .79 | (.63, .93) | .16 |
| Availability for the child | | | | |
| Physical availability | .39 | .32 | (.14, .52) | −.02 |
| Emotional availability | .61 | .24 | (.04, .44) | .05 |
| History | | | | |
| Became parent at young age (<18 jaar) | .04 | .19 | (−.47, .73) | .05 |
| Victim of child maltreatment | .22 | .80 | (.65, .91) | .05 |
| History of violence against persons | .22 | .45 | (.22, .65) | .12 |
| Problematic partner relationship | .58 | .51 | (.35, .67) | .21 |
| *Risk factors of the child* | | | | |
| Young child (<5 jaar) | .08 | .37 | (−.06, .72) | .17 |
| Burdened prehistory (e.g. premature) | .21 | .39 | (.18, .61) | .28 |
| (Serious) disease, handicap, or disability | .16 | .65 | (.44, .85) | .11 |
| Behavioural and/or developmental problems | .91 | .10 | (−.38, .50) | .08 |
| Difficult temper | .48 | .04 | (−.15, .22) | .04 |
| Unwanted child | .09 | .63 | (.31, .87) | .20 |
| *Risk factors of the family and/or environment* | | | | |
| Low educational level | .26 | .62 | (.43, .79) | |
| One parent family, step family, big family | .43 | .85 | (.76, .94) | .26 |
| Many conflicts | .63 | .47 | (.28, .64) | .27 |
| Domestic violence | .42 | .55 | (.35, .69) | .19 |
| Instable, disordered life | .28 | .21 | (−.01, .44) | .25 |
| Material/financial problems (unemployment, housing) | .43 | .83 | (.70, .94) | .30 |
| Important life events | .76 | .31 | (.08, .54) | .28 |
| Social isolation/social conflict | .46 | .31 | (.12, .50) | .08 |
| *Risk factors for child maltreatment?* | * | *.33* | *(.20, .45)* | *.22* |
| Protective factors | | | | |
| *Protective factors of the parents* | | | | |
| Feeling of competence, capacity | .30 | .22 | (−.01, .44) | −.03 |
| Positive self-image | .27 | .59 | (.40, .77) | −.04 |

**Table 6** continued

| | Study 1 | | | Study 2 |
|---|---|---|---|---|
| | Proportion 'yes' | Alpha | Bootstrap 95% confidence interval | $R_s$ unsafe outcome |
| Supporting partner | .15 | .29 | (−.03, .57) | .01 |
| Can deal with own youth experiences | .08 | .14 | (−.32, .54) | .03 |
| Positive youth experiences | .16 | .55 | (.29, .78) | .02 |
| Can ask for/profit from support | .32 | .16 | (−.07, .36) | .09 |
| Emotional availability | .15 | .33 | (.01, .60) | −.03 |
| Flexibility | .04 | .17 | (−.43, .76) | .04 |
| Willingness and ability to change | .32 | .49 | (.29, .66) | .05 |
| *Protective factors of the child* | | | | |
| Socially able | .17 | .40 | (.14, .62) | −.03 |
| Positive self-image | .03 | .01 | (−.87, .81) | −.03 |
| Above average intelligence | .10 | .47 | (.16, .78) | −.08 |
| Attractive physical appearance | .30 | .62 | (.46, .79) | .00 |
| Good relation with important adult(s) | .38 | .15 | (−.07, .35) | −.04 |
| Ego resilience (stress resistance) | .08 | .00 | (−.52, .41) | −.05 |
| Willingness and ability to change | .13 | .27 | (−.04, .59) | −.05 |
| *Protective factors of family and environment* | | | | |
| Support informal network | .43 | .21 | (.03, .41) | .16 |
| Support formal network | .17 | .09 | (−.21, .38) | .16 |
| *Protective factors?* | * | *.23* | *(.09, .37)* | −.31 |
| ***Conclusion current safety*** | | ***.48*** | ***(.42, .56)*** | ***.31*** |
| 2. Risk assessment | | | | |
| 2A. Additional risk factors in case of possible/substantiated child maltreatment | | | | |
| (Suspected) perpetrator has direct access to the child | .50 | .31 | (.12, .48) | .13 |
| No supervision of others on child | .17 | .04 | (−.27, .37) | .13 |
| 2B. What can happen? | | | | |
| Possible risks for the child | | | | |
| Life threatening situation/direct physical danger | * | .30 | (.18, .42) | .23 |
| Prolonged/repeated child abuse | * | .34 | (.22, .47) | .33 |
| Onset of child abuse | * | .15 | (−.02, .29) | .23 |
| Expected consequences for the child | * | .20 | (.06, .34) | .34 |
| 2C. Protective factors | | | | |
| Protective factors that can decrease risks? | * | *.14* | *(−.05, .31)* | −.31 |
| ***Conclusion current risks*** | | ***.39*** | ***(.33, .45)*** | ***.37*** |
| ***Conclusion risks with changes foreseen*** | | ***.19*** | ***(.10, .26)*** | ***.17*** |

Items in italics are intended to summarize answers on preceding items. Items in bold and italics are intended to draw conclusions. The "proportion 'yes'" indicates how often participants chose the option 'yes' for the items about the presence of specific factors, for all cases together. These are not proportions of agreement, but they indicate whether a specific cue seems relatively rare (proportions close to 0) or common (close to 1). Wide confidence intervals mostly occur with those cues that seem to be relatively rare or common, and so do low alphas.

* These questions did not have the options yes/no/unknown, but many/some/none/unknown

## References

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.

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

Baird, C., Wagner, D., Healy, T., & Johnson, K. (1999). Risk assessment in child protection services: Consensus and actuarial model reliability. *Child Welfare, 78,* 723–748.

Barber, J., Trocmé, N., Goodman, D., Shlonsky, A., Black, T., & Leslie, B. (2007). *The reliability and predictive validity of consensus-based risk assessment*. Toronto: Centre of Excellence for Child Welfare.

Bartelink, C., De Kwaadsteniet, L., ten Berge, I., Witteman, C., & Van Gastel, W. (2015). *Betrouwbaarheid en validiteit van de LIRIK: Eindrapport LIRIK valideringsonderzoek. [Reliability and validity of the LIRIK: Final report LIRIK validation study.]*. Utrecht: Nederlands Jeugdinstituut.

Bartelink, C., & Kooijman, K. (2013). Inschatten van veiligheid en kans op kindermishandeling: Noodzaak, instrumenten en ontwikkelingen. [Estimating the safety and risk of child maltreatment: Necessity, instruments and developments.]. *Tijdschrift voor Sociale Geneeskunde, 91,* 391–393.

Bartelink, C., Van Yperen, T. A., ten Berge, I. J., De Kwaadsteniet, L., & Witteman, C. L. M. (2014). Agreement on child maltreatment decisions: A nonrandomized study on the effects of structured decision-making. *Child & Youth Care Forum, 43,* 639–654.

Baumann, D. J., Dalgleish, L., Fluke, J., & Kern, H. (2011). *The decision-making ecology*. Washington: American Humane Association.

Baumann, D. J., Law, J. R., Sheets, J., Reid, G., & Graham, J. C. (2005). Evaluating the effectiveness of actuarial risk assessment models. *Children and Youth Services Review, 27,* 465–490.

Begle, A. M., Dumas, J. E., & Hanson, R. F. (2010). Predicting child abuse potential: An empirical investigation of two theoretical frameworks. *Journal of Clinical Child & Adolescent Psychology, 39,* 208–219.

Berben, E. G. M. J. (2000). *Als iedereen hetzelfde was… indicatiestelling in de jeugdzorg. [If everybody would be the same… assessment of youth care.]*. Maastricht: Shaker Publishing B.V.

Camasso, M. J., & Jagannathan, R. (2000). Modeling the reliability and predictive validity of risk assessment in child protective services. *Children and Youth Services Review, 22,* 873–896.

Camasso, M. J., & Jagannathan, R. (2013). Decision making in child protective services: A risky business? *Risk Analysis, 33,* 1636–1649.

Cichetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23,* 695–700.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

D'Andrade, A., Benton, A., & Austin, M. J. (2005). *Risk and safety assessment in child welfare: Instrument comparisons*. Berkeley: Bay Area Social Services Consortium.

Dalgleish, L. I. (1988). Decision-making in child abuse cases: Applications of social judgment theory and signal detection theory. In B. Brehmer & C. R. B. Joyce (Eds.), *Human Judgment: The SJT view* (pp. 317–360). North Holland: Elsevier.

Dalgleish, L. I. (2000). Assessing the Situation and Deciding to do Something: Risk, Needs and Consequences. *Paper presented at the 13th International congress on child abuse and neglect*, Durban, September 2000.

Dalgleish, L. I. (2003). Risk, needs and consequences. In M. C. Calder (Ed.), *Assessments in child care: A comprehensive guide to frameworks and their use* (pp. 86–99). Dorset: Russell House Publishing.

Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8,* 512–520.

Danktert, E. W., & Johnson, K. (2013). *Risk assessment validation: A prospective study*. Los Angeles: California Department of Social Services, Children and Family Services Division.

Davidson-Arad, B. (2005). Fifteen month follow-up of children at risk: Comparison of the quality of life of children removed from home and children remaining at home. *Child and Youth Services Review, 27,* 1–20.

Davidson-Arad, B. (2010). Four perspectives on the quality of life of children at risk kept at home and removed from home in Israel. *British Journal of Social Work, 40,* 1719–1735.

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

De Kwaadsteniet, L., Bartelink, C., Witteman, C. L. M., ten Berge, I. J., & Van Yperen, T. A. (2013). Improved decision making about suspected child maltreatment: Results of structuring the decision process. *Children and Youth Services Review, 35,* 347–352.

De Ruiter, C., Hildebrand, M., & Van der Hoorn, S. (2012). Risicotaxatie bij kindermishandeling: De Child Abuse Risk Evaluation–Nederlandse versie (CARE-NL). [Risk assessment in child maltreatment cases: The Child Abuse Risk Evaluation–Dutch version (CARE = NL.). In H. P. B. Lodewijks & L. Van Domburg (Eds.), *Instrumenten voor risicotaxatie: Kinderen en jeugdigen [Instruments for risk assessment: Children and youth]* (pp. 169–190). Amsterdam: Pearson.

Doyle, J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *The American Economic Review, 97,* 1583–1608.

Dumbrill, G. C. (2005). *Child welfare in Ontario: Developing a collaborative intervention model*. Toronto: Ontario Association of Children's Aid Societies.

Edwards, V. J., Anda, R. F., Dube, S. R., Dong, M., Chapman, D. F., & Felitti, V. J. (2005). The wide-ranging health consequences of adverse childhood experiences. In K. Kendall-Tackett & S. Giacomoni (Eds.), *Victimization of children and youth: Patterns of abuse, response strategies*. Kingston, NJ: Civic Research Institute.

Enosh, G., & Bayer-Topilsky, T. (2014). Reasoning and bias: Heuristics in safety assessment and placement decisions for children at risk. *British Journal of Social Work, 45*, 1–17.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de Kwaliteit van Tests [COTAN review system for the quality of tests]*. Amsterdam: NIP, COTAN.

Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., et al. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The adverse childhood experiences (ACE) study. *American Journal of Preventive Medicine, 14,* 245–258.

Finnila, K., Santtila, P., Mattila, J., & Niemi, P. (2012). The effects of experience, outcome feedback, and cognitive feedback on decision-making in child sexual abuse cases: A simulation study. *Nordic Psychology, 64,* 242–257.

Gambrill, E., & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth Services Review, 22,* 813–837.

Gambrill, E., & Shlonsky, A. (2001). The need for comprehensive risk management programs in child protective services. *Children and Youth Services Review, 23,* 79–107.

Gold, N., Benbenishty, R., & Osmo, R. (2001). A comparative study of risk assessment and recommended interventions in Canada and Israel. *Child Abuse and Neglect, 25,* 607–622.

Golnik, A., Maccabee-Ryaboy, N., Scal, P., Wey, A., & Gaillard, P. (2012). Shared decision making: Improving care for children with autism. *Intellectual and Developmental Disabilities, 50,* 322–331.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Hacking, I. (1992). World-making by kind-making: Child abuse for example. In M. Douglas & D. Hull (Eds.), *How classification works* (pp. 180–238). Edinburgh: Edinburgh University Press.

Hardman, D. (2009). *Judgment and decision making: Psychological perspectives*. West Sussex: BPS Blackwell.

Hart, S. D., Mitchie, C., & Cooke, (2007). Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *British Journal of Psychiatry, 190*(49), s60–s65.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77–79.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Hollander, M., Van Klaveren, S., Faun, H., & Spijkerman, M. (2013). *Arbeidsmarkteffectrapportage transitie jeugdzorg (labour market outcomes report transition youth care)*. Zoetermeer: Panteia.

Johnson, W. (2006). 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. *Children and Youth Services Review,* 28, 704–714.

Johnson, W. L. (2011). The validity and utility of the California Family Risk Assessment under practice conditions in the field: A prospective study. *Child Abuse and Neglect, 35,* 18–28.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515–526.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendation. *Human Communication Research, 30,* 411–433.

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures, 5,* 93–112.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lekkerkerker, L., Bartelink, C., & Eijgenraam, K. (2011). *De indicatiestelling bij de Brabantse Bureaus Jeugdzorg nader bekeken. Een onderzoek naar de kwaliteit van het indicatieproces en de betrouwbaarheid van het indicatiebesluit. [A closer look at the assessment of Youth Care Agency Brabant. A study on the quality of the assessment process and reliability of the care decision.].* Utrecht: Nederlands Jeugdinstituu.

Lindsey, D. (1992). Reliability of the foster care placement decision: A review. *Research on Social Work Practice, 2,* 65–80.

MacKenzie, M. J., Kotch, J. B., & Lee, L. (2011). Toward a cumulative ecological risk model for the etiology of child maltreatment. *Children and Youth Services Review, 33,* 1638–1647.

Mandel, D. R., Lehman, D. R., & Yuille, J. C. (1994). Should this child be removed from home? Hypothesis generation and information seeking as predictors of case decisions. *Child Abuse and Neglect, 18,* 1051–1062.

Munro, E. (1999). Common errors of reasoning in child protection work. *Child Abuse and Neglect, 23,* 745–758.

Munro, E. (2005). Improving practice: Child protection as a systems problem. *Children and Youth Services Review, 27,* 375–391.

Munro, E. (2008). *Effective child protection.* London: Sage.

Munro, E. (2014). Understanding the causal pathways to child maltreatment: Implications for health and social care policy and practice. *Child Abuse Review, 23,* 61–74.

Nanni, V., Uher, R., & Danese, A. (2012). Childhood maltreatment predicts unfavorable course of illness and treatment outcome in depression: A meta-analysis. *The American Journal of Psychiatry, 169,* 141–151.

Orsi, R., Drury, I. J., & Mackert, M. J. (2014). Reliable and valid: A procedure for establishing inter-item level interrater reliability for child maltreatment risk and safety assessments. *Children and Youth Services Review, 43,* 58–66.

Perry, B. D. (2009). Examining child maltreatment through a neurodevelopmental lens: Clinical applications of the neurosequential model of therapeutics. *Journal of Loss and Trauma, 14,* 240–255.

Pinto, R. J., & Maia, A. C. (2013). Psychopathology, physical complaints and health risk behaviours among youths who were victims of childhood maltreatment: A comparison between home and institutional interventions. *Children and Youth Services Review, 35,* 603–610.

Prins, D. (2011). *Een onderzoek naar de ORBA-werkwijze: Onderzoek, Risicotaxatie en Besluitvorming van de Advies- en Meldpunten Kindermishandeling. [A study on the ORBA method: Investigation, risk assessment and decision-making in the advice and reporting centres on child abuse and neglect.].* Utrecht: University of Utrecht (masterthesis).

Regehr, C., Bogo, M., Shlonsky, A., & LeBlanc, V. (2010). Confidence and professional judgment in assessing children's risk of abuse. *Research on Social Work Practice, 20,* 621–628.

Rycus, J. S., & Hughes, R. C. (2003). *Issues in risk assessment: Policy white paper.* Columbus, Ohio: North American Resource Center for Child Welfare.

Schuerman, J., Rossi, P. H., & Budde, S. (1999). Decisions on placement and family preservation: Agreement and targeting. *Evaluation Review, 23,* 599–618.

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,* 409–427.

Smithgall, C., Jarpe-Ratner, E., Gnedko-Berry, N., & Mason, S. (2015). Developing and testing a framework for evaluating the quality of comprehensive family assessment in child welfare. *Child Abuse and Neglect, 44,* 194–206.

Ten Berge, I. J. (1998). *Besluitvorming in de kinderbescherming. De ontwikkeling en evaluatie van een checklist voor de beoordeling van meldingen bij de raad voor de kinderbescherming. [Decision-making in Child Protective Services. The development and evaluation of a checklist for decision-making at Child Protective Services intake.].* Dissertation, Eburon, Delft.

Ten Berge, I. J., Eijgenraam, K., & Bartelink, C. (2014a). *Licht instrument risicotaxatie kindveiligheid: Herziene versie juni 2014 [Light instrument risk assessment child safety: Revised version June 2014].* Utrecht: Nederlands Jeugdinstituut.

Ten Berge, I. J., Eijgenraam, K., & Bartelink, C. (2014b). *Licht instrument risicotaxatie kindveiligheid: Toelichting en instructie [Light instrument risk assessment child safety: Explanation and instruction].* Utrecht: Nederlands Jeugdinstituut.

Ten Berge, I., & Meuwissen, I. (2013). *Bruikbaarheid en mogelijke aanpassingen van de LIRIK voor de toepassing in de (L)VB-sector: Bevindingen van de pilot augustus 2012–oktober 2013 [Utility and possible adaptations tot he LIRIK for the use in the (mild) mental disabilities field: Findings from the pilot August 2012–October 2013]*. Utrecht: Nederlands Jeugdinstituut.

Ten Berge, I., & Van Rossum, J. (2009). *Evaluatie en bijstelling GCT en LIRIK. Samenvatting resultaten en aanpassingen [Evaluation and adaptation GCT and LIRIK. Summary results and adaptations]*. Utrecht: Nederlands Jeugdinstituut.

Ten Berge, I., & Vinke, A. (2006a). *Beslissen over vermoedens van kindermishandeling: Eindrapport project Onderzoek, Risicotaxatie en Besluitvorming Advies- en Meldpunten Kindermishandeling (ORBA) [Deciding on suspicions of child maltreatment: Final report on ORBA project]*. Utrecht, Woerden: Nederlands Jeugdinstituut, Adviesbureau Van Montfoort.

Ten Berge, I., & Vinke, A. (2006b). *Beslissen over vermoedens van kindermishandeling: Handreiking en hulpmiddelen voor het Advies- en Meldpunt Kindermishandeling [Deciding on suspicions of child maltreatment: Practice manual and tools for the Advice and Reporting Centres of Child Abuse and Neglect]*. Utrecht, Woerden: Nederlands Jeugdinstituut Adviesbureau Van Montfoort.

Turnell, A., & Edwards, S. (1999). *Signs of safety: A solution and safety oriented approach to child protection casework*. New York, London: Norton.

Van der Elst, M., Sondeijker, F., Vogel, I., Jansen, W., & Hermanns, J. (2012). *Veiligheidsrisicotaxatie bij Opvoedhulp en Opgroeihulp aan Gezinnen met Kinderen van 0-12 jaar: Validering van de California Family Risk Assessment. [Safety risk assessment in child and youth care to families with children aged 0–12 year: Validation of the California Family Risk Assessment]*. Woerden: GGD Rotterdam-Rijnmond Van Montfoort Collegio.

Van der Put, C. E., Assink, M., & Stams, G. J. J. M. (2016). Predicting relapse of problematic child-rearing situations. *Children and Youth Services Review, 61,* 288–295.

Vis, S. A., Strandbu, A., Holtan, A., & Thomas, N. (2011). Participation and health: A research review of child participation in planning and decision making. *Child and Family Social Work, 16,* 325–335.

White, A., & Walsh, P. (2006). *Risk assessment in child welfare: An issues paper*. Ashfield: Centre for Parenting & Research.