

## University of Groningen

### Dilemmas in child protection

Bartelink, Cora

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*

2018

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Bartelink, C. (2018). *Dilemmas in child protection: Methods and decision-maker factors influencing decision-making in child maltreatment cases*. [Thesis fully internal (DIV), University of Groningen]. Rijksuniversiteit Groningen.

**Copyright**

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

**Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

## Chapter 4

# **Agreement on child maltreatment decisions: A nonrandomized study on the effects of structured decision-making**

Previously published as: 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 and Youth Care Forum*, 43, 639-654.

This study was financed by ZonMw (grant number 157004003).

## **Abstract**

**Background:** Practitioners investigating cases of suspected child maltreatment often disagree whether a child is subject to or at risk of abuse or neglect in the family and, if so, what to do about such abuse or neglect. Structured decision-making is considered to be a solution to the problem of subjective judgments and decisions.

**Objective:** This study investigates the effects of ORBA, a method for structured decision-making in Advice and Reporting Centres for Child Abuse and Neglect (ARCCAN), on interrater agreement of judgments and decisions.

**Methods:** Two groups of ARCCAN practitioners, one trained in using ORBA and one untrained, used a questionnaire to make judgments and decisions on the same case vignettes. Interrater agreement on the judgments was obtained by calculating the percentage of agreement, intra class correlation, and the Kappa coefficient.

**Results:** Both ORBA trained and untrained practitioners showed little agreement on judgments and decisions, except for the judgment on child maltreatment substantiation, for which trained practitioners showed fair agreement. Agreement among trained and untrained practitioners only differed for some judgments and decisions, and differences were not always in the same direction.

**Conclusions:** This result indicates no convincing evidence that structured decision-making leads to better agreement on decisions concerning child abuse and neglect. Recommendations for improvements in uniform decision-making and further research are given.

## 1. Introduction

In the Netherlands, practitioners of the Advice and Reporting Centres for Child Abuse and Neglect (ARCCAN) investigate cases of suspected child maltreatment. The practitioners assess if a child is raised in a threatening or unsafe situation and decide whether care or protection is needed. There are fifteen ARCCANs, one in each province and the three largest cities. ARCCAN social workers first judge the severity and urgency of reported cases. Then, the social workers decide between different courses of action – advising informants how to provide support to the child or family, telling informants to gather more information, starting their own investigation to determine if a child is indeed maltreated, or immediately referring a case either to a welfare organisation or to a child protection agency (Baeten, 2009; Ten Berge, & Vinke, 2006a). ARCCAN practitioners do not make individual decisions about cases, but discuss the cases in an interdisciplinary team at least twice. A physician, behavioural scientist or both are involved in the decisions about actions to be taken (Baeten, 2009). The quality of the initial judgments that are provided as input in this interdisciplinary process appears to be critical for the quality of the entire decision making process (Munro, 1999; 2008).

Judging whether child abuse or neglect actually occurs is difficult, because there are many uncertainties (Benbenishty, 1992; Kaplan, Pelcovitz, & Labruna, 1999). For example, information is often lacking or contradictory because parents and children do not want to talk about their problems (Forrester, Kershaw, Moss, & Hughes, 2008; Munro, 1999). Also, informants, such as a child's teacher or family doctor, may provide different information because they see the subtle signs of child maltreatment differently (Munro, 2008). Practitioners make decisions under time pressure, since help should be provided as soon as possible when a child is indeed being maltreated. Families are not always willing to cooperate with the practitioner because they may not be aware that their problems are serious (Forrester et al., 2008). They may also be wary of the power of the investigators, who have the power to remove the children from home (Dumbrill, 2005; Turnell & Edwards, 1999). Decisions can have far-reaching consequences for children and parents, and it is often unsure which decision will have the most desirable outcome. For example, an out-of-home placement is sometimes necessary because of serious threats to a child's health, wellbeing and development (Oosterman, Schuengel, Slot, Bullens, & Doreleijers, 2007; Van der Horst & Van der Veer, 2009; Van IJzendoorn, 2008), but at the same time it can be harmful, especially to young children who are still developing close attachment relations (Juffer, 2010; Perry, 2009). Because child maltreatment cases may result in long-term harm or even death of a child, it is necessary to make timely decisions and provide appropriate care (Cyr, Euser, Bakermans-Kranenburg, & Van IJzendoorn, 2010; Fearon, Bakermans-Kranenburg, Lapsley, & Roisman, 2010; Perry, 2002).

Under such time pressure and uncertainty, and due to limited cognitive resources, practitioners' information processing may be compromised. Practitioners may overlook relevant information, attach too much importance to irrelevant details or only search for information that confirms their previous judgments (Munro, 1999; 2008). There are no clear empirically based guidelines for child maltreatment decisions (Berben, 2000; Drury-Hudson, 1999; Munro, 1998; Ten Berge, 1998), and practitioners rely on personal experiences and beliefs (Arad-Davidzon & Benbenishty, 2008; Gambrill & Shlonsky, 2000; Garb, 2005; Osmo & Benbenishty, 2004; Schuerman, Rossi, & Budde, 1999). Given these circumstances, it is not surprising that research shows that practitioners disagree, sometimes strongly, on important judgments and decisions (Berben, 2000; Britner & Mossler, 2002; Gold, Benbenishty, & Osmo, 2001; Munro 2008; Schuerman, Rossi, & Budde, 1999; Ten Berge, 1998; Van Montfoort, 2004).

Although some mistakes due to a lack of information or other situations seem inevitable (Munro, 1996), other mistakes might be avoided when practitioners carefully consider available information and critically judge their own opinions and experiences (Gambrill, 2005).

A structured decision process may help. It may lead to a critical consideration of the information used and of alternative decision options and their potential consequences. Decision makers may consequently make more objective decisions (Hodgkinson, Bown, Maule, Glaister, & Pearman, 1999) and plan better interventions if they use a structured decision process (Bolton & Lennings, 2010; Léveille & Chamberland, 2010; Wagner, Johnson, & Caskey, 2001). A possible remedy to biased and subjective decision-making thus seems to be to explicate and systematise the decision-making process (see for example De Bruyn, Ruijsenaars, Pameijer, & Van Aarle, 2003; Munro, 2008; Pameijer & Van Beukering, 2004; Shlonsky & Wagner, 2005).

In 2005/2006 a systematic and empirically founded method for investigation, risk assessment and decision making on suspicion of child maltreatment was developed: ORBA. Grounded in evidence-based and practice-based knowledge, ORBA combines four theoretical and practical approaches: 1) principles of effective child protection (Munro, 2002), 2) the Risk/General Assessment and Decision-making Model (Dalglish, 1997), 3) Signs of Safety (Turnell & Edwards, 1999) and 4) deciding on interventions in families (Van Montfoort, 2004). These approaches emphasise the importance of making explicit judgments and decisions through a structured decision-making process to consider the extent of potential danger for the child and the possibilities to guarantee child safety and protection in the family (for further details of the development of ORBA, see Ten Berge & Vinke, 2006a; 2006b). Munro (2008) described the risk assessment process and advised how to prevent decision errors. Dalglish (1997) distinguished judgments and decisions. A judgment is an assessment of a situation given the current case information. A decision addresses whether or not to take a course of action. Turnell and Edwards (1999) emphasized that practitioners should be aware of both positive (i.e. strengths, resources, protective factors) and negative (i.e. problems, risk factors) factors in a family. All these approaches, combined in ORBA, emphasise the importance of making explicit judgments and decisions through a structured decision-making process in which the extent of potential danger for the child and the possibilities to guarantee child safety and protection in the family are considered (for further details of the development of ORBA, see Ten Berge & Vinke, 2006a; 2006b). ORBA offers guidelines, criteria, and checklists to assist in the process of collecting relevant information about cases, judging if there is a case of substantiated child maltreatment, and deciding whether care or protection is needed. ORBA distinguishes two key decisions: 'Do you accept the case for investigation?' and 'How do you close the case: Referring to child and youth care for help, passing the case to the Child Care and Protection Board, another action, or closing the case without further action?' (see also the section on Measures). For each key decision, ORBA describes which information practitioners should collect and organise, divided into several domains: signs of child maltreatment, child functioning, parenting, parental characteristics, family and environment, and history of services received. This information should include concerns, problems and risk factors as well as positive characteristics and protective factors. Practitioners analyse this information by answering questions about the nature and severity of the problems in these domains. Based on their analyses, practitioners reach final conclusions (judgments) about suspicion or substantiation of child maltreatment (for an example see Table 1; Ten Berge & Vinke, 2006b).

**Table 1.** Example of a key decision

<p>Key decision: Accept a case for investigation</p> <p>Possible outcomes:</p> <ul style="list-style-type: none"><li>• Do not accept the case</li><li>• Accept the case as a report for further investigation</li><li>• Give a single advice to the informant</li><li>• Give advice to the informant repeatedly</li></ul> <p>The ORBA-manual describes criteria what cases a practitioner should accept and not accept and when he should give advice to the informant.</p> <p>To make this decision a practitioner should assess:</p> <ul style="list-style-type: none"><li>• The nature and severity of the problems: facts and signs of child maltreatment</li><li>• Informant: his abilities and limitations, intentions and expectations</li><li>• Urgency: is there a crisis situation in the family?</li></ul> <p>Then the ORBA manual describes what information should be collected.</p> <p>Regarding the nature and severity of the problems information is needed on:</p> <ul style="list-style-type: none"><li>• The suspected child maltreatment: impressions and perceptions of the informant; facts; nature, duration, frequency; course (increase or decrease of the problems);</li><li>• The child: emotional, physical, behavioural signs of maltreatment or other problems;</li><li>• Parenting: basic care, safety, affection/attachment behaviour;</li><li>• Parents: parenting style, parenting knowledge and capacities; personal problems (psychological problems, addiction); relationship problems; physical or mental disabilities;</li><li>• Family and environment: family composition; social-economical situation; support network; social contacts; special situations (e.g., important events)</li></ul> <p>Regarding the information the practitioner should collect information on:</p> <ul style="list-style-type: none"><li>• The immediate cause: why reporting now? For how long does the informant worry about this child or family?</li><li>• Earlier report: same informant or same story? Earlier investigation?</li><li>• Reliability and credibility of the informant: reasons for reporting; personal interest in reporting; kind of relationship with the child/family; realistic view on children; his own perceptions or others.</li><li>• Abilities and limitations: Is the informant able to speak with the parents about his worries? Is the informant able to investigate the situation and/or offer help to the family?</li></ul> <p>Practitioners have a checklist that supports them to give attention to all these elements.</p>
---

International methods for child protection service assessment based on the same principles as ORBA include the Victorian Risk Framework (VRF; Armitage et al., 1999) from Australia and the Structured Decision Making (SDM) model (Children’s Research Center, 1999) from the United States. The Framework for the Assessment of Children in Need and their Families (FACNF) was developed in England (Department of Health, 2000). Although this framework was not specifically designed for decisions on child maltreatment cases, it also has a structured decision-making approach. Léveille and Chamberland (2010) studied the effects of the FACNF and concluded that practitioners using the framework make better assessments of complex situations in families, use a more child-centred point of view, and plan interventions better.

ORBA was implemented nationwide in 2007–2008. A recent analysis of case records in ARCCANs showed that ORBA leads to a more systematic and transparent decision-making process (De Kwaadsteniet, Bartelink, Witteman, Ten Berge, & Van Yperen, 2013). In the current paper we focus on the question whether ORBA leads to less subjective practitioner judgments, as expressed in higher practitioner agreement. Ideally, there is a gold standard or an agreed upon definition of what constitutes child maltreatment. However, there is neither consensus on a standard set of child maltreatment definitions among practitioners (Leeb, Paulozzi, Melanson, Simon, & Arias, 2008)<sup>4</sup>, nor are there reliable and valid child maltreatment assessment instruments for use as a reference tool in the Netherlands (Ten Berge & Bartelink, 2014). Therefore, interrater agreement was studied as the best available alternative.

The research question of this study is if the use of ORBA in the Netherlands leads to more uniformity of judgments and decisions than work without ORBA. The purpose of this study is to shed light on the following issues: 1) What is the interrater agreement of ARCCAN practitioners on key judgments and decisions? 2) Are there differences in the interrater agreement between practitioners trained in working with ORBA and untrained professionals? If ORBA reduces subjectivity, one expects that trained practitioners will agree more in their judgments and decisions than untrained novices.

## 2. Method

### 2.1 Participants

An experimental design with randomisation was not possible because ORBA has already been implemented nationally. Therefore, a study design with experienced and novice ARCCAN practitioners was chosen. A group of 40 ARCCAN practitioners who had received a one-day training in ORBA in 2007 or 2008 and who passed a test of proper ORBA application in 2009 or 2010 participated in this study. The group was selected from among the approximately 70 practitioners in four ARCCAN locations who had passed the test, via a multi-stage sampling. They constitute a random sample of eight to twelve participants from each of the four ARCCAN locations; numbers of participants depended on the size of the ARCCAN.

A second group consisted of 40 practitioners who had applied for the ORBA training but not yet followed it. They participated in the study prior to the training. These participants had been working in an ARCCAN for a maximum of four months, but most of them had at least 5 years of work experience in child and youth care or child protection.

Of the 80 participants, 69 were women. They worked at 12 out of the 15 ARCCAN locations. Most practitioners were social workers. Behavioural scientists and team leaders did not participate in the study and were only asked to assess the quality of the vignettes (see below). Age and work experience of both groups are reported in Table 2.

---

<sup>4</sup> A child maltreatment definition is established in the Law on Child and Youth Care Child, but offers no exact judgment and decision making rules: child maltreatment is any act or series of acts of physical, psychological or sexual commission or omission by a parent or other caregiver that results in harm, potential for harm, or threat of harm to a child.

**Table 2.** Group characteristics

	Mean (SD)		T-value
	Untrained	Trained	
<b>Gender</b>			
• Men	3	8	
• Women	37	32	
Age	37.0 (9.6)	43.6 (8.7)	-4.536**
Work experience at ARCCAN (in years)	0.8 (1.1)	6.6 (5.0)	-10.094**
Other relevant work experience (in years)	6.4 (6.6)	13.4 (9.6)	-5.391**

\*\*  $p < .001$

Thirty-three (82.5%) of the trained practitioners had received the official ORBA training; seven had been trained by a team leader or behavioural scientist. Nine of the untrained practitioners had received provisional instruction in ORBA from a team leader or behavioural scientist. In the analyses, we controlled for these unintended training effects by comparing not only the total groups ( $n = 40$  vs.  $n = 40$ ), but also the officially trained ( $n = 33$ ) with the completely untrained practitioners ( $n = 31$ ).

## 2.2 Measures

*Vignettes.* The design used vignettes. A vignette booklet was compiled with 16 case descriptions. Eight descriptions were of cases where the key decision was to accept or reject a case for investigation. The other 8 cases concerned the decision what to do after the ARCCAN investigation.

The vignettes had varied characteristics, including age and sex of the child or children, severity and complexity of the problems, and type of informant who notified the ARCCAN of the suspected child maltreatment (professional or private person). No cases with a self-evident decision were included. The vignettes were based on existing cases, which were rewritten for study purposes to be untraceable to actual persons. The vignettes were tested for usefulness and representativeness by 14 team leaders and behavioural scientists. Some adaptations to the vignettes were made based on their feedback.

*Questionnaire.* Each participant completed a questionnaire about four vignettes. The questionnaire surveyed ORBA key judgments and decisions and participants' justifications for these. Key judgments and decisions were critical questions that had to be answered with yes or no. In deciding whether to accept a case for a child maltreatment investigation, the key judgments and decisions were:

1. Is there a well-founded suspicion of child maltreatment? (judgment)
2. Do you accept the case for investigation? (decision)

For investigated cases, the key judgments and decisions were:

1. Is child maltreatment substantiated? (judgment)
2. Does this child or family need help? (judgment)
3. Can help be delivered on a voluntary basis? (judgment)
4. How do you close the case? (decision)



The last decision did not involve a dichotomous yes/no answer but was a choice between alternatives: referring the case to child and youth care for help on a voluntary basis; passing the case to the Child Care and Protection Board to investigate the necessity of a child protection order; another action, specified in free text; or close the case without further action.

Questionnaire items asked for the participant's background, including age, work experience and ORBA training.

### 2.3 Procedure

Because of time restrictions, ARCCAN practitioners did not judge all sixteen vignettes. A blocked design was used to randomly combine vignettes. Each participant judged two vignettes for each key decision. Thus, each vignette was judged by 20 practitioners, 10 trained and 10 untrained practitioners, resulting in 160 judgments for each group. A power analysis was conducted. The power is higher than .99 with 160 judgements per group (40 practitioners x 4 vignettes) and a moderate effect size (Cohen's  $d = .50$ ;  $\delta = 4.47$ ;  $df = 1, 318$ ;  $\alpha = .05$ ).

The vignettes were sent by post or e-mail. The questionnaire was completed on the Internet.

### 2.4 Analysis

Almost all key judgments and decisions were yes-no questions that were analysed as dichotomous variables (see Van Yperen, 1990). Chance-controlled interrater agreement was calculated with the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979). The only exception was the decision about actions after the ARCCAN investigation. Cohen's kappa was calculated for this item. An ICC between  $-1.00$  and  $.30$  and kappa between  $-1.00$  and  $.20$  means no or slight agreement, an ICC between  $.31$  and  $.50$  and kappa between  $.21$  and  $.40$  means fair agreement, an ICC between  $.51$  and  $.70$  and kappa between  $.41$  and  $.60$  means moderate agreement, and an ICC between  $.71$  and  $1.00$  and kappa between  $.61$  and  $1.00$  means substantial to almost perfect agreement (see also Landis & Koch 1977; Van Yperen 1990). There is no consensus on the interpretation of differences in agreement or on a statistical way to investigate significance of differences between groups for ICC and kappa. Goldstein and Healy (1995) propose comparison of the (95%) confidence intervals for differences on interrater agreement between groups. A statistically significant difference ( $p$  value  $< .05$ ) roughly corresponds with an overlap in confidence intervals of one quarter of the length of the intervals, or half one arm of the interval, given sufficient sample size and no large differences in width (Belia, Fidler, Williams, & Cumming, 2005; Cumming & Finch, 2005).

Cohen's kappa and ICC can be unreliable measures if the distribution of answers is highly skewed. This situation is called the "base rate problem"; if one category is chosen most of the time, chance corrected measures for agreement cannot obtain high values (Spitzer, Endicott, & Robbins, 1978; Spitznagel & Helzer, 1985; Van Yperen, 1990). We set as limit for a useful interpretation of chance-controlled measures that skewness should not exceed 3.0 (Van Yperen, 1990).

## 3. Results

### 3.1 Overall interrater agreement

Table 3 shows the percentage of agreement and the chance corrected indexes (ICC and Cohen kappa) for key judgments and decisions. The ICC and kappa showed poor overall agreement, which was in fact often not substantially higher than could be expected by chance. The participants were more in agreement on the judgment on the substantiation of child maltreatment than on other

judgments and decisions, but even this agreement could only be qualified as fair. Table 4 shows that many item distributions were skewed. If we focus on the items that are not subject to this “base rate problem”, agreement appeared to be low.

Inspection of the percentages reveals that participants were more in agreement on judgments than decisions.

### **3.2 Differences between trained and untrained practitioners**

We expected trained practitioners to be in better agreement on key judgments and decisions than untrained practitioners. The results in the total group ( $n = 40$  vs.  $n = 40$ ) do not clearly support this expectation. The ICC and kappa seem to indicate that, compared with untrained participants, trained participants agreed more on the substantiation of child maltreatment and their decision on actions needed after the ARCCAN investigation. The 95% confidence intervals of the other judgments and decisions largely overlap.

A further comparison of the interrater agreement of those participants who had received a complete training ( $n = 33$ ) with those who had not had any introduction to ORBA yet ( $n = 31$ ) indicates no obvious differences between these two groups. Observed differences between the total groups ( $n = 40$  vs.  $n = 40$ ) seem to disappear or reverse direction in comparison with these results.

**Table 3.** Interrater agreement of untrained and trained workers on ORBA key judgments and decisions

Variable	Untrained				Trained				Total			
	%	ICC	95% CI	%	ICC	95% CI	%	ICC	95% CI	%	ICC	95% CI
<b>Reports</b>												
<i>Suspicion of child maltreatment?</i>												
• Total (n = 40 vs. n = 40)	67.5	.10* (-)	[-.01, .44]	77.5	.03 (-)	[-.16, .89]	75.1	.09* (-)	[.04, .37]			
• Selection (n =31 vs. n = 33)	67.9	.06* (-)	[-.09, .45]	73.3	.07 (-)	[-.08, .47]	70.0	.12*	[.02, .44]			
<i>Decision: accept for investigation?</i>												
• Total	65	.22** (-)	[.05, .61]	52.5	.18* (-)	[.03, .55]	58.8	.20** (-)	[.07, .55]			
• Selection	60.8	.30** (-)	[.06, .71]	56.7	.11* (-)	[-.06, .52]	66.0	.21** (-)	[.06, .57]			
<b>ARCCAN investigations</b>												
<i>Child maltreatment substantiated?</i>												
• Total	77.5	.21* (-)	[.04, .59]	90	.64** (+)	[.40, .89]	83.8	.39** (-/+)	[.19, .74]			
• Selection	86.2	.27* (-)	[.04, .68]	90	.65** (+)	[.38, .90]	91.5	.39** (-/+)	[.18, .74]			
<i>Need help?</i>												
• Total	87.2	.04 (-)	[-.05, .37]	92.3	.05 (-)	[-.04, .40]	92.3	.08* (-)	[.00, .43]			
• Selection	86.2	.09 (-)	[-.07, .50]	89.7	.20* (-)	[-.02, .66]	87.6	.12* (-)	[.01, .49]			
<i>Voluntary or involuntary help?</i>												
• Total	69.2	.25** (-)	[.06, .67]	61.5	.27** (-)	[.08, .69]	69.2	.21** (-)	[.06, .63]			
• Selection	68.9	.21* (-/+)	[.00, .64]	55.1	.09 (-)	[-.08, .54]	63.8	.21**(-)	[.05, .61]			
<i>Decision after investigation</i>												
• Total	37.5	Kappa = .06 (-)	n.a.	50	Kappa = .24* (-/+)	n.a.	40.1	Kappa = .10 (-)	n.a.			
• Selection	51.7	Kappa .26* (-/+)	n.a.	40	Kappa .09 (-)	n.a.	39.0	Kappa .07 (-)	n.a.			

Note 1. % = percentage of agreement; ICC = intraclass correlation coefficient, unless otherwise reported; no or slight agreement = -; fair agreement = -/+; moderate agreement = +; substantial to almost perfect agreement = ++; \* p < .05; \*\* p < .001. Note 2. For explanation of the selection of participants see section "Participants".

**Table 4.** Distribution of responses

Variable	Untrained		Trained		Total	
	Distribution of answers	Skewness	Distribution of answers	Skewness	Distribution of answers	Skewness
<b>Reports</b>						
Suspicion of child maltreatment?	78.8%	-1.51	88.8%	-2.50	83.8%	-1.90
Decision: accept for investigation?	55.0%	-0.23	48.8%	0.05	51.9%	-0.09
<b>ARCCAN investigations</b>						
Child maltreatment substantiated?	86.3%	-2.60	88.8%	-2.80	87.5%	-2.67
Need help?	92.5%	-3.66	95.0%	-4.93	93.8%	-4.1
Voluntary or involuntary help?	67.5%	-0.81	60.0%	-0.51	63.8%	-0.64
Decision after investigation						
• Child and youth care	38.8%	0.45	22.5%	1.24	30.6%	0.80
• Child Care and Protection Board	23.8%	1.24	35.0%	0.68	29.4%	0.93
• Other	37.5%	0.62	42.5%	0.34	40.0%	0.47

Note. In the column 'Distribution of answers' the percentage of participants answering 'yes' has been reported.

Additional analyses were performed to control for homogeneity of agreement among participants and among vignettes. All vignettes showed similar levels of agreement; practitioners did not agree much more or less on some vignettes. There were no practitioners who clearly agreed less on judgments and decisions than their colleagues, indicating that the above-mentioned results are not influenced by deviant practitioner decisions or deviant vignettes.

We controlled for the influence of age and amount of work experience on interrater agreement, but did not find any effect for these variables. There is no reason to assume that gender influences interrater agreement.

We also considered characteristics of the vignettes (child gender, child age, informant, severity of the case as judged by the respondents) as moderators of the interrater agreement. The interrater agreement improved when the respondents judged the vignettes as more severe and the report was of a professional informant and not a lay person. The child's gender and age had no moderating effect on the interrater agreement (more information on these analyses is available in Ten Berge, Bartelink, & De Kwaadsteniet, 2011).

#### **4. Discussion**

The primary question in this study was whether structured decision-making using ORBA leads to more uniformity in judgments and decisions in cases of suspected child maltreatment than in situations where ORBA is not used. Interrater agreement about key judgments and decisions of ORBA trained ARCCAN practitioners was compared to interrater agreement of untrained ARCCAN practitioners. In both groups, interrater agreement appeared to be low, with the exception of the judgment on child maltreatment substantiation, which was fair for trained participants. There were a few differences in agreement between trained and untrained participants but not always in the same, expected direction. So, our findings did not show that ORBA leads to more uniform judgments and decisions.

We may have found no effect of ORBA training, because also untrained participants had already some ORBA experience. Untrained participants often worked more than a few months in the ARCCANs, and had sometimes been instructed by a team leader or behavioural scientist in the use of ORBA. However, interrater agreement for trained users who had received a complete training was not substantially better than for the completely untrained users. A possible explanation for this lack of differences is the fact that ORBA had been implemented in the ARCCAN client registration system. Consequently, all practitioners used ORBA to some extent already. Still, given that we found such poor interrater agreement in both groups, ORBA trained and untrained, - we clearly cannot conclude that structured decision making using ORBA leads to better or good uniformity of judgments and decisions.

The results of this study are disappointing but consistent with previous studies. Agreement on child maltreatment cases and decisions about necessary help is generally low (Benbenishty, Osmo, & Gold, 2003; Berben, 2000; Britner & Mossler, 2002; Schuerman, Rossi, & Budde, 1999). A more recent study agreed with our results and found that practitioners disagree on decisions even when they use systematic and validated instruments (Regehr, Bogo, Shlonsky, & LeBlanc, 2010). Further, practitioners who agree on the 'diagnosis' can still disagree on what help should be delivered (Arad-Davidzon & Benbenishty, 2008; Lindsey, 1992). We similarly found that agreement about the presence of child maltreatment or the necessity of help was higher than agreement about case acceptance for investigation or steps to be taken after investigation.

There are several possible explanations for this lack of agreement, especially regarding decisions. An important explanation is that structured decision making still leaves much room for a practitioner's individual considerations. ORBA provides guidelines about data collection and analysis that can be applied to cases of child abuse or neglect in general (Ten Berge & Vinke, 2006b). It is not an actuarial instrument, and in each case, the practitioner still has to weigh the different signs to conclude whether this situation is a case of child maltreatment, and what is the best course of action. The absence of clear child maltreatment definitions encourages application of personal values in the judgment and decision making. Application of personal values that are not shared by colleagues or society is undesirable, because these values may not lead to decisions based on equal rights of the parents. But given the state of knowledge, current general guidelines cannot provide formulas to weigh the signs and clearly define decision-making cut-off points (Baartman, 2009).

Further, decisions are influenced by factors other than case characteristics. Baumann, Dalgleish, Fluke and Kern (2011) present their Risk/General Assessment and Decision-making Model, which states that a decision depends on two components: the assessment of the problems and the decision maker's personal threshold to make a certain decision, such as deciding to accept a case for investigation. The assessment of the problems depends on case-specific characteristics, such as child, parental, and familial characteristics. The decision of what to do depends on the decision maker's personal threshold. The threshold refers to the point at which the assessment of the case information (e.g., amount and weight of evidence) is serious enough for a person to decide to take action. This threshold is influenced by personal and professional experiences and subjective values, in addition to other factors. ORBA's criteria may influence the assessment, but not practitioners' personal thresholds. The question arises whether or not clear child maltreatment definitions can be formulated that decrease personal influences subsequently.

A final possible explanation is that practitioners did not sufficiently know or use the ORBA guidelines and criteria. This study did not investigate this possibility. Although experienced practitioners were trained and had passed a test, this does not guarantee that the practitioners continue to use ORBA. ORBA use depends on the quality of organisational implementation (Stals, 2012). Prins (2011) showed some implementation concerns, including ORBA's integration into the ARCCANs' client registration system and sustained training on the job.

There are some limitations to this study. First, although vignettes are typically being used to study interrater agreement (e.g., Mezzich, Mezzich, & Coffman, 1985), a vignette study does not fully correspond to reality because practitioners have limited written information and lack access to oral or non-verbal information. Some may object to the use of vignettes, claiming that they are too distant from the actual clinical situation. However, the use of vignettes provided several benefits. We were better able to control the information provided to the therapist, for example, by ensuring that similar categories of information were contained in each vignette. Vignettes also permitted us to systematically vary type of maltreatment and severity, which made it possible to analyse the impact of these factors on the decisions (Eels, Lombart, Kendjelic, Turner, & Lucas, 2005). A second limitation specific to this study concerns the two groups of research participants. It would have been better to study the effects of structured decision-making in a randomized controlled trial, in which work experience in an ARCCAN would not be related to prior knowledge and use of ORBA. Because ORBA was used in all ARCCANs at the time of this study, a design with ORBA trained and untrained practitioners was the only, although suboptimal, option. This limits our ability to draw conclusions about the effects of ORBA on interrater agreement. Still, additional

analyses did not reveal effects of age or amount of work experience. Also additional analyses with more clearly defined groups (only officially trained and fully untrained practitioners) did not improve interrater agreement. But most important, although, we cannot exclude the effects of other variables like experience, given that we found such poor agreement about judgments and decisions in both groups we can safely conclude that our study does not show that structured decision-making results in reliable decisions about cases of suspected child abuse or neglect.

## 5. Recommendations

An earlier study showed that ORBA lead to greater uniformity in the collection of information (De Kwaadsteniet et al., 2013). Despite this enhanced uniformity, our study reveals that ORBA has not lead to clearly improved or good interrater agreement. Use of the same information and criteria for the assessment is a necessary condition for practitioner uniformity, but this use insufficiently increases interrater agreement. Poor agreement is a serious flaw in child protection decision making. In similar cases, the decisions made should not vary according to the assigned practitioner.

Structured decision making methods like ORBA may not lead to improved agreement about child maltreatment decisions. Actuarial instruments, providing algorithms for decision thresholds, may lead to more reliable decisions. Actuarial instruments have been found to lead to more reliable child safety taxations, but current empirical knowledge about the best actions in cases of (suspected or substantiated) child maltreatment is insufficient yet to actuarially derive the best decisions (e.g., Baird & Wagner, 2000). We suggest that to address the problem of the lack of uniformity in judgments and decisions, first practitioners should be made aware of subjective influences on their decisions. De Kwaadsteniet et al. (2013) concluded that practitioners hardly provide the rationale for their decisions. Practitioners should explicitly identify and discuss the rationale for their conclusions and decisions. Disagreement on decisions could be caused by different explanations for the problems (De Kwaadsteniet, Hagmayer, Krol, & Witteman, 2010) and should be subject to discussion. Team decision-making can contribute to this process if the team conversation is explicit, structured and well-founded to avoid biases (for example, see Pijnenburg, 1996).

Differences in thresholds for decisions as well as personal or societal incidents that may affect a practitioner's threshold should also be regularly paid attention to in supervisory meetings. Discussions on thresholds, decisions made and the suspected effects of decisions can possibly improve agreement (Baumann et al., 2011; Munro, 2008).

Also, it may be studied whether interrater agreement improves with a decision-making team consisting of practitioners (executive practitioner, behavioural scientist, and team leader), the family and one or more persons from the family's social network. These persons have different stakes, and more objective thresholds can be potentially formulated based on discussions of these stakes.

This study raised the question of the extent to which practitioners in one team or the same region agree on judgments and decisions. Regional differences in arrangements and the availability of care may affect agreement, with agreement within regions or teams being better than between regions or teams. This possibility also needs further study. Further research may also determine the case characteristics and decision-maker characteristics, particularly attitudes, knowledge and skills, and traumatic experiences (see Baumann et al., 2011; Regehr et al., 2010), that influence agreement.

Finally, longitudinal research is needed to study the long-term impact of specific decisions on children and families. Longitudinal studies can examine and determine patterns of the

development and maintenance of child maltreatment (e.g., Folger & Wright, 2013). Thus, more objective thresholds for child maltreatment judgments and decisions might be established. Also, observing the impacts that judgments and decisions have on children and families by practitioners themselves may provide them with currently lacking feedback about positive or negative effects of their decisions.



