The Reliability and Predictive Validity of Consensus-Based Risk Assessment

James Barber

Nico Trocmé

Deborah Goodman

Aron Shlonsky

Tara Black

Bruce Leslie

Funded by:
Social Sciences and Humanities Research Council of Canada (SSHRC) grant

CENTRES OF EXCELLENCE FOR CHILDREN’S WELL-BEING

Child Welfare

The Reliability and Predictive Validity of Consensus-Based Risk Assessment
was funded by Social Sciences and Humanities Research Council of Canada (SSHRC) grant.

ISBN: 978-0-7727-7894-9

Citation: James Barber, Nico Trocmé, Deborah Goodman, Aron Shlonsky, Tara Black,
Bruce Leslie. *The Reliability and Predictive Validity of Consensus-Based Risk Assessment*,
Toronto: Centre of Excellence for Child Welfare, 2007.

To order additional copies of this report, contact:
Centre of Excellence for Child Welfare
www.cecw-cepb.ca

Contents

Acknowledgements

Financial support for this research was provided by a 3-year Social Sciences and Humanities Research Council of Canada (SSHRC) grant. The research team also gratefully acknowledges the important assistance of our community partners in conducting this research: the many children's aid societies' staff who provided their perceptive thoughts in the gathering of the focus group data, and the significant contribution made by the Catholic Children's Aid Society of Toronto, in providing the quantitative case data with data extraction expertise from Kenneth Chan and Sinforiano Llano. We would also like to thank Joanne Daciuk for her dataset expertise, and a special thank you goes to Dolly Jakupi, a research assistant and case reader for the project. Without Ms. Jakupi's help, this project would not have been completed. Thank you to our additional case readers for the reliability study Jackie Dell and Jacquie Horsley and to our data entry clerks for the validity study: Heather Johnson, Liz Lambert, Garima Bhardwaj, and Chizuru Mitani. We would also like to thank our statistician, Aiala Barr, for her statistical expertise.

Introduction

Child welfare services across North America are struggling to target their limited resources in the face of growing demands for services (Waldfogel, 1998). In Ontario alone, the estimated number of substantiated child abuse and neglect cases doubled between 1993 and 1998 (Trocmé, Fallon, MacLaurin & Copp, 2002) and doubled again between 1998 and 2003 (Fallon, Trocmé, MacLaurin, Knoke, Black, Daciuk & Felstiner, 2005). In the United States, the number of maltreated children doubled from 1.4 million to more than 2.8 million between 1986 and 1993 (Sedlak and Broadhurst, 1996), and between 1990 and 2002 there was a 21.3% increase in the number of children who were the subject of CPS investigation or assessment (NCANDS, 2002). Since then, the inclusion of exposure to domestic violence as a reportable form of maltreatment in some North American jurisdictions has led to even more dramatic increases in reports to child protective services (Edleson, 2004; Trocmé et al., 2005).

In response to figures like these, jurisdictions around the world are increasingly turning to structured risk assessment to assist child protection authorities in rationing their services. In 1998, the Canadian Province of Ontario, with a population of 11 million, introduced a province-wide [ONTARIO] Risk Assessment Model (ORAM) to assist the 53 Children's Aid Societies (CAS) in this effort. The model is built on three assessment instruments that are now routinely completed by CAS fieldworkers: (1) the Eligibility Spectrum, which is grounded in the provincial legislation, where the eligibility tool is employed to determine whether the in-coming report has a *prima facie* claim to the services of CAS, (2) The Safety Assessment Tool (SAT), which assesses whether or not the child is in immediate danger, and (3) the Risk Assessment Tool. Ontario's RA Tool is based on an instrument that was developed in the early 1990s by the New York State Department of Social Services and it consists of five assessment categories related to the: (1) caregiver, (2) child, (3) family, (4) intervention (e.g. caregiver's receptivity to intervention), and (5) abuse/neglect history. Within each of these categories or "Influences" are related risk "elements," derived by a panel of experts from child welfare theory, research studies and field experience. In all, 22 risk elements are rated by the RA Tool on five-point scales of severity ranging from 0 to 4. After scoring each element, results are not combined arithmetically but workers are guided through a number of summary questions and prompts before using their clinical judgment to arrive at an overall rating, from 1 "No/Low Risk" to 5 "High Risk." In contrast to the crisis focus, "present-tense" lens of the first two instruments, the RA Tool is intended to assist workers by providing a "future" view that will predict the ongoing level of risk to the child until the next scheduled reassessment.

In general, risk assessment models can be divided into two kinds: consensus-based and actuarial, with some models combining elements of both. Actuarial models are based on the empirical study of child protection cases and their future maltreatment outcomes. The object of the exercise is to identify factors that are known to be statistically predictive of future maltreatment and to use this information in the construction of an instrument that can be scored in a purely mechanical fashion. Ontario's RA Tool, by contrast, is an instance of consensus-based risk assessment because workers rate selected characteristics that were originally identified by consensus among experts and these factors are then processed using professional judgment rather than according to a standard algorithm. Irrespective of whether the actuarial or consensus-based method is preferred, however, a frequently cited benefit of structured risk assessment is that it leads to greater consistency of evaluation and reliability of response among child protection workers. This movement towards structured risk assessment is representative of a trend in social work generally towards evidence-based practice, which emphasises methods that are built on a solid foundation of scientific research. Among the critics of this trend are those who object not so much to empirically validated practice, but to the loose way in which the research is sometimes applied. Some critics (see, for example, Rycus & Hughes, 2003; Wald & Woolverton, 1990) have objected that risk assessments are often used to inform decisions for which they have never been validated or for which no empirical evidence exists, such as whether or not to remove a child or how much intervention to provide. Furthermore, despite the increasingly sophisticated risk assessment methods in use today, large-scale validation studies have not kept pace with the growing use of the instruments. What data do exist suggest that most instruments currently in use have questionable reliability and/or validity or have not been subjected to empirical investigation at all (Camasso & Jagannathan, 2000; Lyons, Doueck & Wodarski, 1996; Rycus and Hughes, 2003).

In Ontario, in 1995 the Ministry of Community and Social Services [re named Ministry of Children and Youth Services in 2005] imported then modified the New York State instrument without re-examining its psychometric properties prior to implementation. Ironically, just as Ontario was beginning this process, research from around the world, including an internal evaluation by New York State of the instrument on which Ontario's RA Tool was based, was coming to the view that the actuarial approach performs better in the field than consensus models do. In the New York State study (Falco & Salovitz, 1997), an actuarial instrument was developed through a process of reading and coding case files for factors that might predict recurrence of maltreatment. As the consensus-based model had been in operation for some time by then, items from that instrument were also extracted. In a retrospective longitudinal analysis of case files, the resultant instrument was assessed for discrimination of cases, predictive validity, reliability, and generalizability to different jurisdictions. Notwithstanding the omission of some important methodological details, such as sample size and statistical analysis, for example, the subsequent report suggested that when weighted properly, a relatively small number of predictors could classify cases into four levels of risk that performed well against the evaluation criteria. Importantly, this (atheoretical) instrument outperformed the consensus-based one that was in operation at the time.

In the published literature, most studies addressing the reliability of risk assessment instruments have concentrated on inter-rater agreement and have used one of two methods. The first involves constructing case vignettes that contain sufficient information for blind raters to perform risk assessments, and the second involves blind readings of case files. The level of agreement between raters adjusted for chance is then calculated (cf. Bakeman & Gottman, 1986; Nasuti & Pecora, 1993). In an example of the first approach, Fluke et al. (1993) evaluated the inter-rater reliability of three common risk assessment models using case vignettes constructed to assess reliability for risk by type of maltreatment at different decision points during the case. The authors could find only moderate levels of agreement but because Fluke et al. (1993) did not control for level of risk in their case scenarios, it is impossible to judge whether their results were better or worse than would be expected in the field. Because it is always easier to differentiate high- from low-risk cases than it is to distinguish between cases in the middle range, an adequate test of reliability requires exposure to cases from across the full range of risk levels. Using the case reading technique, Baird, et al. (1999) compared the reliability of two consensus-based assessment tools and one actuarial model. In that study, Baird et al. selected 80 cases from four sites (20 from each) and trained

12 case readers (3 from each site) in one or other of the three risk assessment models. Copies of the 20 case files from each site were stripped of identifying information and sent to the case reading teams at the other three sites. Team members at the other sites then read each case and completed their respective risk assessment instruments, thereby producing four independent ratings of each of the 80 cases. This procedure produced a large amount of variance among raters in the risk levels assigned to cases in all of the systems, although reliability was significantly higher for the actuarial model than for either of the consensus-based approaches. In contrast, Camasso and Jagannathan's examination of the New Jersey Risk Assessment Matrix consensus based model found inter-rater reliability coefficients in the .85-.90 range (as cited in Jagannathan and Camasso, 1996), while Wood's (1997) assessment of inter-rater reliability of an actuarial instrument yielded a median *kappa* of only 0.66 for 63 randomly selected case files.

The few published studies into the validity of risk assessment that are available suggest that the predictive performance of most instruments is fairly poor (Baird and Wagner, 2000; Camasso and Jagannathan, 2000; Rittner, 2002). Generally speaking, the data suggest that less than one-third of the variance in maltreatment recurrence can be explained by the factors included in risk assessment instruments (Baird and Wagner, 2000; Camasso and Jagannathan, 1995; Fuller et al., 2001; Rittner, 2002). However, Baird and Wagner (2000) have objected that the viability of prediction is inherently problematic for low base rate phenomena such as child maltreatment, and for this reason a shift in focus is warranted, from predicting who will recidivate to assigning cases to risk categories based upon "observed rates of behaviour". Using this approach, several studies indicate that actuarial risk designations are indeed associated with different rates of subsequent maltreatment. In an evaluation of the Alaska Risk Assessment model, for example, 83% of very high risk cases of abuse (physical, emotional and sexual) that were not removed from the home were subsequently abused, as compared with only 3.3 % cases classified as very low risk; and that 70% of high risk neglect cases (physical, medical and emotional) not removed from home were subsequently reported for neglect compared with 7.5% of cases classified as low risk (Baird, 1988). Similarly, Wood (1997) found that 52% and 34% of abuse cases classified as high to very high risk were the subject of new allegations or new substantiations, respectively, compared with 12% and 5% of the low risk cases. The rate of new allegations or substantiated reports in high to very high risk cases of neglect was 45% and 19% over the same period of time, compared with 4% and 1% of the low risk cases. Finally, using the Vermont Family Risk Assessment Matrix, 61% of the families rated high risk were subsequently reported, compared with 36% and 24% of the moderate and low-risk groups, respectively (Weedon et al., 1988). Importantly, Baird and Wagner (2000) found a significant difference in the predictive validity of their consensus-based and actuarial instruments. Their study compared one actuarial instrument with two consensus-based risk assessment tools on ability to predict new investigations and new substantiations 18-months following assessment. One thousand and four hundred cases from four U.S. states were classified as low, medium or high risk based upon case reader assessments. The actuarial model produced substantially better risk classifications than either of the consensus-based approaches.

The present study forms part of a larger project assessing the reliability and predictive validity of Ontario's RA Tool, as well as its intended and unintended effects on social work practice. In the first study, we report on the inter-rater reliability of Ontario's RA Tool using a case reading approach. In the second study, the predictive validity of the RA Tool is assessed. The third study element explores workers' views on the ramifications of the tool on practice.

Methods

Study I: Reliability

The first study examined Ontario's RA Tool's reliability and drew a stratified random sample of 132 cases from one of Ontario's large children's aid societies. Initial risk scores for each of these cases were extracted from case files and compared with the scores assigned by three blind case readers, who read and rated each of the case files independently. The internal consistency and inter-rater reliability of risk judgments were then calculated. The second study examined the predictive validity of risk assessment scores for 1,118 cases selected according to Study 1 criteria. In a retrospective longitudinal design, all cases selected had received at least two RA Tool ratings: the first upon completion of the initial investigation (Time 1) and the second at the time of case closure (Time 2). These scores were then used to predict recurrence of maltreatment at any point up to 18 months post-closure (Time 3).

Sample

The case files used for reliability assessment were a sub-set of those included in Study II. Project cases were selected electronically from the administrative database of one of Ontario's largest children's aid societies (CAS), and selection parameters included that:

1. the case had closed at least 18 months prior to the case selection date;
2. the case had opened after the ORAM had been implemented and computerized;
3. the child had not been made a Crown Ward;
4. the youngest child did not turn 16 years of age during the period covered by the project;
5. the case had not been transferred to another CAS (so its electronic record was therefore complete);
6. the family had not moved to another jurisdiction and their whereabouts was known during the period covered by the project.

One-thousand-one-hundred-and-eighteen cases between December 2000 and March 2003 satisfied these criteria. From these files, a stratified random sample of 132 cases was extracted for case reading.[1] Cases were stratified by abuse type and severity level. Sexual abuse cases were excluded because of their low incidence and, for this same reason, overall risk ratings of 1 and 2 were collapsed into a single low

[1] The Eligibility codes for each case were determined (the Eligibility Spectrum is available upon request). The Spectrum has 10 Sections; Section 1-5 are the child protection sections and Sections 6-10 outline the non-protection or voluntary services; the five protection sections have scales that detail two to five different forms of child maltreatment respective of that section; each protection scale in each section is divided into four levels of severity (extremely, moderately, minimally, and not severe); and each level of severity has one or more descriptors. The child protection entry point for each scale is between the moderate severe and minimally severe levels. For the purposes of this study, all cases that were given a code of Section 1, Scale 3 (Sexual Abuse) were excluded from the reliability study due to the low frequency. All cases given a code of Section 1 fell under the category "Physical Abuse." All cases given Section 2 and Section 4, Scale 1 fell under the category of "Neglect." All cases given Section 3 and Section 4, Scale 2 fell under the category of Emotional Abuse. Finally, all cases given a code of Section 5 fell under the category "Caregiver with a Problem."

risk group, while ratings of 4 and 5 were collapsed into a high risk group. A breakdown of the resultant sample is presented in Table 1.

Table 1: Case reading sample broken down by maltreatment category and risk level

Maltreatment Type	Risk Category			Total
	Low	Medium	High	
Physical Abuse	11	11	11	33
Neglect	11	11	11	33
Emotional Abuse	11	11	11	33
Caregiver Problem	11	11	11	33
Total	44	44	44	132

Each case could have up to four caregivers and up to six children. The sample contains information for 252 caregivers, and 277 children (see Table 2). The sample of children was evenly split by gender and over one quarter (28 percent) were under the age of 3. Caregivers were predominantly mothers (51%), though over one third (34 percent) of the files included information on fathers.

Table 2: Characteristics of Reliability

	Number of Cases	Percentage
Child Age		
0–3 Years	75	27%
4–7 Years	78	28%
8–11 Years	66	24%
12–15 Years	45	16%
16+ Years	3	1%
Missing	10	4%
Total 0-16	**277**	**100%**
Gender of Children		
Male	138	50%
Female	134	48%
Missing	5	2%
Total Children Rated	**277**	**100%**
Caregiver Relationship to Child		
Mother	128	51%
Father	89	35%
Grandmother	16	6%
Grandfather	4	2%
Other	15	6%
Missing	0	0%
Total Caregivers Rated	**252**	**100%**

NOTE: *There can be multiple caregivers and multiple children per investigation.*

Procedure

The following five documents were extracted from electronic case files and provided to the blind case readers:

1. The initial "Referral Form," which is completed by the intake worker and provides information on the date and time of the report, a brief statement as to the nature of the allegation, agency codes relating to the child's eligibility (see above), the recommended response time, rationale for response time, and record of the protection investigation plan;
2. The "People Profile," which records demographic information on the child, caregiver and the person responsible for the alleged maltreatment, and internal and provincial record checks;
3. The "Safety Assessment," which is a 12-item tool, is intended to assess the child's immediate safety. Among the items included in the checklist are: whether the caregiver's current behaviour is "violent or out of control", the "child's whereabouts cannot be ascertained", or the caregiver "has previously harmed a child";
4. "Case Activity" information, which is a summary of prior Children's Aid Society contact, if applicable;
5. The "Investigation" module completed by the investigating worker, which contains narrative reports of the interviews conducted during the investigation. This module also records the final decision as to whether or not the allegation(s)/concern(s) was (were) verified, and whether the abuse was verified.

This information was printed out for each case at the participating agency and all identifying information was removed from the records on site.

All case readers were previous or current child protection workers in Ontario. All were trained to use the RA Tool by the Ontario Association for Children's Aid Societies (OACAS) in the standardized New Worker Training (NWT) program. The case readers were not additionally trained so results could reflect the child protection field.

Reliability analyses were conducted to assess the internal consistency of Ontario's RA Tool's subscales (called Influences), and the inter-rater reliability of each of the RA Tool's 22 risk elements and overall risk rating. In addition to establishing the psychometric properties of the RA Tool's, individual risk factors were sought that independently predicted subsequent child maltreatment for this sample. Similar to other analyses of risk assessment instruments in child welfare (Baird et al., 1999; Camasso and Jagannathan, 2000), reliability analyses compared the ratings of blind case readers who were exposed to the same case information. In addition, since the RA Tool had been fully implemented in the field prior to data abstraction for the current study, comparisons were also made with the investigative caseworker's risk assessment.

Study II: Validity

Sample

As noted previously, the Eligibility Spectrum is a screening tool used to ascertain whether maltreatment reports should be investigated. All cases given an eligibility screening code of Section 1, scales 1, 2 and 4 fell under the category "Physical Abuse." All cases given a code of Section 1, scale 3 fell under sexual

Table 3: Validity sample

Primary Maltreatment Type	Initial Risk Category			Total	
	Low	Medium	High	Number	%
Physical Abuse	32	181	85	**298**	**27**
Sexual Abuse	9	16	9	**34**	**3**
Neglect/Caregiver Incapacity	57	217	130	**404**	**36**
Emotional Abuse	4	10	7	**21**	**2**
Exposure to Domestic Violence	25	164	66	**255**	**23**
Other	13	64	24	**101**	**9**
Missing				**5**	**0**
Total	**140**	**652**	**321**	**1,118**	**100**

abuse. All cases given Section 2, Section 4 Scale 1, Section 5 Scale 3 and 4 fell under the category of "Neglect/Caregiver Incapacity." All cases given Section 3, scale 1 fell under the category of Emotional Abuse. All cases given Section 3, scale 2 fell under the category of Exposure to Domestic Violence. All other cases fell under the other category (see Table 3).

All 1,118 case files that were extracted according to the selection criteria described in Study I were used in the predictive validation study. For the most part, percentages were similar between the reliability and validity samples. The largest proportion of children was between eight and eleven years of age (26%).

Table 4: Characteristics of Validity Samples

	Number of Cases	Percentage
Child Age		
0–3 Years	462	22%
4–7 Years	510	24%
8–11 Years	543	26%
12–15 Years	420	20%
16+ Years	84	4%
Missing	91	4%
Total 0-16	**2,110**	**100%**
Gender of Children		
Male	1,139	52%
Female	1,051	48%
Missing	15	1%
Total Children Rated	**2,205**	**100%**
Caregiver Relationship to Child		
Mother	1,084	50%
Father	824	38%
Grandmother	109	5%
Grandfather	31	1%
Other	129	6%
Missing	0	0%
Total Caregivers Rated	**2,177**	**100%**

NOTE: *There can be multiple caregivers and multiple children per investigation.*

Twenty-four percent were between four and seven, 22% were between zero and three years, and 20% were between 12 and 15 years. Fifty-two percent were boys and 48% were girls. Mothers represented 50% of the caregivers rated in the validity study. Thirty-eight per cent were fathers, 5% grandmothers, 1% grandfathers and 6% other caregiver (see Table 4).

Procedure

Demographic information together with initial (T1) risk assessments, case closure (T2) risk assessments, people profiles, and investigation module were extracted from case files. The date of any maltreatment verifications following case closure (T2) was also recorded from the Disposition B module of the ORAM, up to 18-months post closure. The date of the first verification (T3), if any, was then used in the assessment of the predictive validity of T1 and T2 risk assessment scores.

Study III: Focus Group

Intake workers, family service workers and supervisors from a wide range of Ontario's 53 Children's Aid Societies (CAS) were solicited in Fall 2005 through the OACAS website and mailing list to participate in a day-long focus group that both gathered field opinion on their experiences with the current ORAM Risk Assessment (RA) Tool, as well as had them review risk and clinical assessment tools that were being considered for use across the province as part of the province's child welfare Transformation Agenda. Qualitative data related to the respondents' views on the ORAM RA Tool were used to inform this study segment.

Sample

Since the 53 CASs vary in size and population, in order to maximize inclusion and variation, focus group participation was limited to two volunteers from each agency. In addition, focus groups were held in three locations across Ontario, Ottawa (East), Sudbury (North), and Toronto (South and West) in order to minimize travel for participants and garner a sample that was representative of the province. The clear distinction in job description and types of decision-making needs between intake and ongoing services workers necessitated conducting separate focus groups for each type of worker. Intake workers and supervisors were asked to participate in Component I of the focus groups and ongoing services workers and supervisors were asked to participate in Component II. In all, 92 workers and supervisors volunteered and ultimately participated in focus groups across the three settings.

Procedures

Prior to the focus groups, volunteers were asked to select and review the charts of three of their own cases that had been closed in the last six months. Component I (intake) participants were asked to review at least one case that was not opened and one case where the child was taken into protective custody. Component II (family service) participants were asked to review at least one case where the child was reunified and one case where the child was never placed in out of home care (i.e., received ongoing services without placement in foster care). For each of the three cases, Component I volunteers were asked to complete mock versions of two new proposed tools: a Safety Assessment and a Risk Assessment. Similarly, Component II respondents completed two mock tools on their three cases—a Risk Reassessment and a Family Reunification Assessment. For both Component situations, volunteers were to use only information that would have been available to them at each respective decision point and were requested to document their answers to a number of questions. Their opinion data related to the query: "What have workers found to be positive (strengths) and negative (limitations) with using Ontario's RA Tool? The RA Tool comments were extracted from the focus group data and analyzed for this study. Thematic analysis generated from the data was forwarded to an expert panel for their review and validation. The expert panel was made up of three senior child welfare staff from three different CASs.

Results

Study I: Reliability

Changes in Overall Risk Rating between Time 1 and Time 2

Among children who were subsequently maltreated, the largest proportion of cases indicated no change or worsening in risk score from Time 1 to Time 2 (23.4%) followed by those who showed an improvement of 2 levels (23%). Although a large number of cases were rated by caseworkers as improved between administrations, this did not appear to be related to whether maltreatment recurred (p=0.71) (see Table 5).

Table 5: Change from Time 1 to Time 2

Change from Time 1 to Time 2	Number and % of cases with Abuse at Time 3
No change or worsening	49 (20.9%)
Improvement of one level	93 (39.7%)
Improvement of 2 levels	70 (29.9%)
Improvement of 3 levels or more	22 (12.8%)

p-value = 0.7081

Internal consistency

The viability of each of the RA Tool's four categories (Caregiver, Abuse/Neglect, Child, and Family) was assessed using Cronbach's Alpha, which measures the degree to which each rater gives similar risk scores within a specific domain (Table 6). In order for a domain to be viable as an independent construct, raters should give similar scores for all items that comprise the domain. If ratings between subgroup elements are markedly different, they are not sufficiently related to be clustered as constructs. Alpha ratings range between 0 and 1, with 0.7 being considered minimally consistent.

For the case readers, the summary alphas of all three readers only met or exceeded the minimum requirement of 0.7 in the Caregiver Influence (α=0.73)[2]. The mean summary alpha ratings were below 0.7 for all other Influences. There was some variability among raters, with Reader 1 attaining consistency in three of the five Influences. However, Reader 2 was consistent in none of the Influences and Reader 3 attained consistency only in the Caregiver Influence (α=0.77). The percent agreements between specific raters can be found in Appendix A.

The original investigative worker most closely resembled Reader 1, rating consistently across three of the five Influences and rating highest among all raters in three of the five Influences. The family category was inconsistent across all readers. The most internally consistent rater was the original caseworker, which may be an indication that: 1) information that is not in the case file is being used to make ratings; 2) structural or organizational processes are in place that influence caseworker ratings, making them more consistent; 3) a combination of both of these.

[2] This alpha is an average of for the 3 summary alphas for the 3 case readers. The equation is: (.73+.68+.77)/3.

Table 6: Internal Consistency and Reliability of Scales

	Internal Consistency[1]				Inter-rater Reliability[3]			
Influence	Alpha Reader 1	Alpha Reader 2	Alpha Reader 3	Alpha Orig Worker	% Agree Case Readers	Average Kappas for Case Readers	% Agree Avg. Case Readers v. Orig. Worker	Average Kappas for Case Readers & Orig. Worker
Caregiver Influence								
Abuse/Neglect of Caregiver	0.79	0.74	0.85	0.79	*86.5	.41	§36.6	.10
Alcohol or Drug use	0.72	0.63	0.72	0.78	§66.3	.40	*63.8	.28
Caregiver's Expectations of child	0.65	0.59	0.68	0.72	45.4	.28	*43.1	n/a
Caregiver's Acceptance of Child	0.63	0.56	0.68	0.71	§57.4	.48	60	.30
Physical Capacity to care for Child	0.67	0.69	0.72	0.76	§63.2	.19	71.4	.26
Mntl/Emot/Int Capacity to care for Child	0.65	0.56	0.74	0.74	§49.2	.25	*50.4	.24
Summary Alpha	0.73	0.68	0.77	0.78				
Abuse/Neglect Influence								
Access to child by Perpetrator	0.82	0.49	0.50	0.68	*52.6	.15	*57.5	.19
Intent and Acknowledge Responsibility	0.73	0.18	0.50	0.78	51.8	.40	43.2	.19
Severity of Abuse/ Neglect	0.68	0.31	0.53	0.69	§42.9	.17	42.3	.17
Hx of Abuse/Neglect by Present Caregivers	0.67	0.49	0.71	0.73	35.9	.18	43.6	.16
Summary Alpha	0.78	0.45	0.64	0.77				
Intervention Influence[2]								
Caregiver's Motivation	na	na	na	na	§39.9	.23	*35.7	.15
Caregiver's Cooperation with Intervention	na	na	na	Na	*41.1	.26	*31.5	.20
Summary Alpha	na	na	na	Na				
Child Influence								
Child's Vulnerability	0.86	0.56	0.75	0.81	§78.3	.84	§79.1	.72
Child's Response to Caregiver	0.67	0.27	0.38	0.73	§69.3	.31	66.5	.24
Child's Behaviour	0.67	0.33	0.41	0.68	§71.2	.42	*69.5	.41
Child's Mental Health and Development	0.60	0.25	0.45	0.71	§70.6	.39	*67.2	.31
Child's Physical Health and Development	0.60	0.48	0.41	0.73	§84.7	.34	§82.4	.27
Summary Alpha	0.74	0.45	0.56	0.78				
Family Influence								
Family Violence	0.49	0.48	0.36	0.67	§46.7	.32	§45.1	.28
Ability to Cope with Stress	0.45	0.41	0.36	0.67	§31.2	.19	29.8	.14
Availability of Social Supports	0.63	0.62	0.52	0.70	42.6	.21	21.2	.08
Living Conditions	0.50	0.58	0.52	0.72	§75.5	n/a	69.5	.38
Family Identity and Interactions	0.39	0.37	0.32	0.73	34.0	.18	36.9	.19
Summary Alpha	0.56	0.56	0.48	0.66				
Overall Risk (Subjective rating 1-5)					47.3		42.3	

1. Cronbach's Alpha was calculated based on standardized variables
2. Alphas could not be calculated for this subscale since it only contains two elements
3. Cohen's Kappa was calculated for all combinations where distributions allowed.
** Significant (p<0.05) for all comparisons*
§ At least one comparison significant

Inter-rater Reliability

Agreement in scores between raters for each of the 22 risk items and the overall risk rating was assessed using Cohen's Kappa, which estimates the likelihood that agreement in rating was reached at levels greater than chance alone. For each item, the percent agreement (i.e., percent of time a rating was agreed upon by two raters) was calculated and, wherever possible, a Kappa statistic was generated (Table 6, last columns). The Kappa statistics are based on a pair of readers, and all pairs were compared. In general, significant Kappa scores of 0.3 to 0.49 indicate fair correspondence between raters, 0.5 to 0.69 indicates moderate correspondence, and 0.7 and above indicates high correspondence.

Results from this analysis are somewhat discouraging. Only nine of the 22 items met minimal Kappa levels of correspondence (0.3 or better) among case readers, and only five out of 22 items reached minimal Kappa levels when case readers were compared with the original case worker.

Moreover, only a small number of risk assessment items that reached minimal or better Kappa range also reached levels of statistical significance. Among case readers, these six items were: alcohol or drug use of caregiver (Kappa=0.40), caregiver's acceptance of the child (Kappa=0.48), caregiver intent and acknowledgement of responsibility (Kappa=0.40), child's response to caregiver (Kappa=0.31), child's behaviour (Kappa=0.42), and child's mental health and development (Kappa=0.39). When compared with the original case worker, no risk items met minimal levels of correspondence at statistically significant levels. Thus, case readers minimally agreed on fewer than half of the items on the tool but case readers (as a whole) never agreed with the original caseworker at accept levels (see Table 7).

In the caregiver category, there was a very high level of agreement for "Abuse/neglect of caregiver" as a child among case readers (average percent agreement=87%), but this level of agreement did not hold when comparisons were made between the three raters and the original worker on the case (average Kappa=0.10, average percent agreement=37%). In all likelihood, this is because the case file did not

Table 7: Reliable risk assessment items

Items of RA Tool		Rater 1 Vs. Rater 2	Rater 1 Vs. Rater 3	Rater 2 Vs. Rater 3	Average
Alcohol or Drug use	Kappa	0.42	0.40	0.38	0.40
	p-value	0.0048	0.0001	0.0009	-
	Percent agreement	65.2%	65.3%	68.3%	66.3%
Caregiver's Acceptance of Child	Kappa	NA	NA	0.48	0.48
	p-value	NA	NA	0.0004	-
	Percent agreement	51.6%	47.8%	72.8%	57.4%
Intent and Acknowledgement of Responsibility	Kappa	NA	NA	0.40	0.40
	p-value	NA	NA	0.0087	-
	Percent agreement	47.3%	40.8%	67.2%	51.8%
Child's Response to Caregiver	Kappa	0.31	NA	NA	0.31
	p-value	0.0235	NA	NA	-
	Percent agreement	67.9%	65.9%	74.1%	69.3%
Child's Behaviour	Kappa	0.48	0.37	0.41	0.42
	p-value	0.0414	0.0002	0.0062	-
	Percent agreement	74.4%	68.5%	70.7%	71.2%
Child's Mental Health and Development	Kappa	0.43	0.33	0.40	0.39
	p-value	0.0391	0.0001	0.0025	-
	Percent agreement	72.9%	67.8%	71.1%	70.6%

contain information about the childhood maltreatment history of the caregiver but the caseworkers may have been aware of such a history. There was a moderate correspondence between case readers and the original workers on the "alcohol or drug use" of the caregiver (average percent agreement=64%, average Kappa=0.28), a fair amount of agreement about the "caregiver's expectations" of the child (average percent agreement=43%, average Kappa not calculated), and moderate correspondence for "mental/emotional/intellectual capacity to care for the child" (average percent agreement=50%, average Kappa=0.24).

The reliability of the overall risk score (a number assigned by the case reader or caseworker that is based upon, but not systematically calculated by the ratings on the previous 22 items) was also assessed using Cohen's Kappa. These calculations indicate that, on average, the case readers did not agree with each other (Kappa=0.22) or with the original worker (Kappa=0.28) at minimally acceptable levels. Looking more closely at the paired comparisons, case readers (1 and 2) and (2 and 3) agreed at low but statistically significant levels (Kappa=0.33 and 0.30, respectively), indicating that some individual raters may have been more reliable than others. Nonetheless, Kappas for overall risk, while showing weak levels of agreement along a 5-point scale for some paired comparisons, have average agreement less than 50% of the time. The distribution of overall risk ratings revealed no strong, clear pattern of agreement among the three case readers or between case readers and the original caseworker. The choice of 5 risk levels, rather than some smaller number, may increase variation in response among raters. The total number of cases classified at the various risk levels and stratified by rater (Figure 1) reveals consistency at the upper and lower risk levels among some groupings of raters (e.g., raters 1 and 3), indicating that a decrease in the number of risk levels might improve reliability.

In the hopes of broadening our understanding of variability in the reliability of the overall risk scores and to identify individual elements that may be both reliable and valid predictors of risk of subsequent child maltreatment, an analysis of caseworker scoring of individual elements with respect to overall risk rating was undertaken (N=132). Spearman rank correlation was used to model bivariate correlations between the worker's overall risk judgment and the score for each element (Table 8). All but two correlations were significant (Abuse/Neglect of Caregiver and Child's Physical Health and Development). In such cases, it is most helpful to look at the size of the correlation for significant items. A correlation of 0.3 was considered strong enough to include in subsequent investigations of predictive validity. These eight more strongly correlated factors included (arranged from strongest correlation

Figure 1: Overall Risk Rating for Cases with Abuse at Time 3

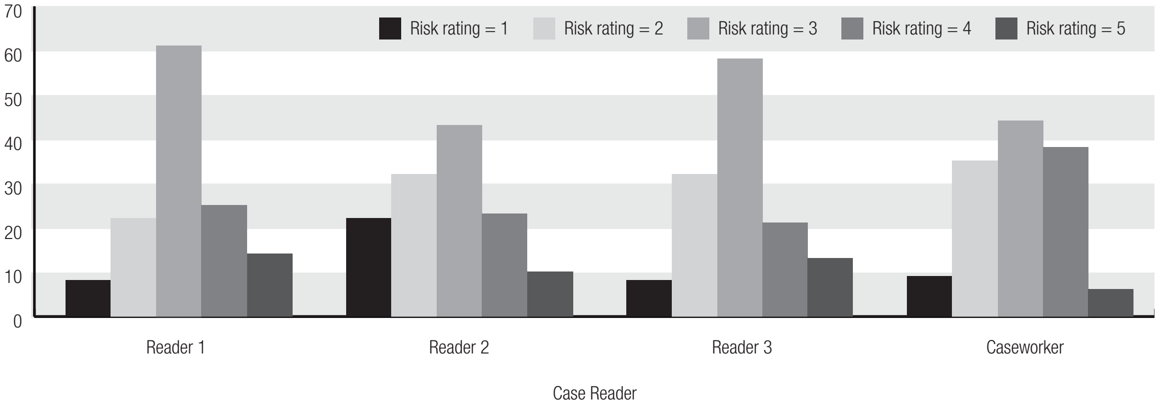


Table 8. Correlation of Individual Factors with Worker's Overall Risk Judgment (scale of 1-5)

Risk Item	Correlation (r)	p-value
Abuse/Neglect of Caregiver	0.04	0.51
Alcohol or Drug use	0.19	<0.01
Caregiver's Expectations of child	0.40*	<0.01
Caregiver's Acceptance of Child	0.35*	<0.01
Physical Capacity to care for Child	0.08	0.02
Mental/Emotional/Intellectual Capacity to care for Child	0.20	<0.01
Access to child by Perpetrator	0.30*	<0.01
Intent and Acknowledgement of Responsibility	0.37*	<0.01
Severity of Abuse/ Neglect	0.34*	<0.01
History of Abuse/Neglect Committed by Present Caregivers	0.29	<0.01
Caregiver's Motivation	0.44*	<0.01
Caregiver's Cooperation with Intervention	0.24	<0.01
Child's Vulnerability	0.06	0.04
Child's Response to Caregiver	0.22	<0.01
Child's Behaviour	0.07	0.03
Child's Mental Health and Development	0.12	0.01
Child's Physical Health and Development	0.03	0.41
Family Violence	0.23	<0.01
Ability to Cope with Stress	0.30*	<0.01
Availability of Social Supports	0.12	0.01
Living Conditions	0.19	<0.01
Family Identity and Interactions	0.37*	<0.01

1 *Spearman rank correlation was used to derive the correlations in this table*
* r≥0.3

to weakest): Caregiver's Motivation (r=0.44, p<0.01); Caregivers Expectations of the Child (r=0.40, p<0.01); Family Identity and Interactions (r=0.37, p<0.01); Caregiver's Acceptance of the Child (r=0.35, p<0.01); Severity of Abuse/Neglect (r=0.34, p<0.01); Access to Child by Perpetrator (r=0.30, p<0.01); and Ability to Cope With Stress (r=0.30, p<0.01). Across all multiple caseworkers, these eight factors were associated with overall risk rating, giving some indication that these factors were weighted more strongly in the caseworker's decision about the level of overall risk faced by children over the course of the original child maltreatment investigation.

In sum, the overall level of agreement among case readers and between case readers and the original caseworker is fairly low. Although a few of the 22 individual risk items reach low to moderate agreement when adjusted for chance, the majority of items show substantial variation between raters. The overall risk rating appears to be inconsistent across raters. However, some of the individual elements predicted overall risk rating for caseworkers, indicating that these factors were generally considered most important when assessing overall risk in the field.

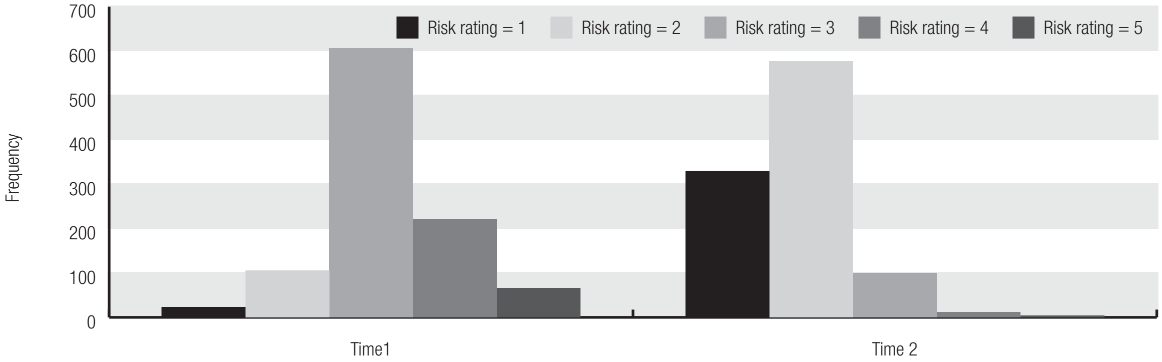
Study II: Validity

Poor to fair reliability does not bode well for predictive validity. Nevertheless, given minimal indications of reliability (particularly for the original investigative worker), it was important to ascertain whether all or part of Ontario's RA Tool had some degree of predictive validity. Predictive capacity of the RA Tool was assessed using the entire sample of cases closed between December 2000 and March of 2003 and otherwise meeting the inclusion criteria (N=1,118). Substantiated maltreatment recurrence within 18 months of case closure was set as the benchmark for ascertaining the extent to which the RA Tool or its items predicted subsequent child maltreatment. For this sample, the RA Tool was administered at least twice. Time 1 administration occurred at the end of the CPS investigation (usually within 30 days) and Time 2 administration occurred just prior to case closure. Thus, the sample being used in this study is an exit cohort with an average length of service of 398.4 days (SD=193). That is, children and families were only included if cases were open for services (ranging from in-home to foster care) and their cases were closed within the study period. The RA Tool bases its assessment of risk on the family as a whole. Multiple caregivers and multiple children could be rated per risk factor; therefore, the highest risk rating in each category was used to determine the predictive validity of that item and a recurrence of maltreatment was based on whether any child in the household experienced subsequent maltreatment.

Changes in Overall Risk Rating between Time 1 and Time 2

The distribution of scores at Time 1 and Time 2 generally showed a decrease in risk rating between administrations (Figure 2). The vast majority of cases at Time 1 were rated as 3 or higher while the vast majority of cases at Time 2 were rated as 2 or lower. Among the cases included in the study, 83.3% (852/1023) indicated improvement in overall risk rating between Time 1 to Time 2, 16.1% (165/1023) showed no change, and almost no cases (6/1023=0.5%) had a higher level of risk at Time 2.

Figure 2: Overall Risk Ratings at Time 1 and Time 2 (N=1,023)



Among children who were subsequently maltreated, the majority of cases indicated an improvement of one level in risk score from Time 1 to Time 2 (41.1%) followed by those who showed an improvement of 2 levels (31.6%). Although a large number of cases were rated by caseworkers as improved between administrations, this did not appear to be related to whether maltreatment recurred (χ^2=0.97, p=0.81) (see Table 9).

Table 9: Change from Time 1 to Time 2

	Number and % of cases with Abuse at Time 3		
Change from Time 1 to Time 2	number	%	% within abuse pts
No change or worsening	47	(25.5%)	[16.7]
Improvement of one level	116	(25.1%)	[41.1]
Improvement of 2 levels	89	(26.7%)	[31.6]
Improvement of 3 levels or more	30	(22.4%)	[10.6]

Prediction of Maltreatment Recurrence

Approximately 25% of the sample experienced abuse at Time 3 (283 cases out of 1,118 cases). The time between case closure and follow-up varied considerably for this sample with a mean length of follow-up of 823 days (SD=411.4), requiring an analysis that accounted for the amount of time children lived with their caregiver. That is, given the same overall likelihood of experiencing maltreatment, abuse/neglect will be observed more often for children who are followed for longer periods of time. Survival analysis (which includes Cox Proportional Hazards models) is a family of statistical techniques that controls for variable length of exposure to risk and has been employed widely in child welfare studies (Wulczyn and George, 1994; Courtney, 1996; Guo and Wells, 2003; Shlonsky, Festinger, & Brookhart, 2006; Simpson et al., 2000). As a form of data reduction, simple (bivariate) Cox Proportional Hazards regression was conducted using maltreatment recurrence as the binary outcome measure and overall risk score at Time 1 and Time 2 as the predictor. Similar to regression, this technique generates a hazard ratio that indicates whether one group (in this case, risk ratings of either "2 or 3," or "3 or 4') has a different likelihood of experiencing the outcome (maltreatment recurrence) than the baseline or reference group (in this case, children who had an overall risk rating=1). Hazard Ratios with values greater than 1.0 indicate that the presence of this characteristic increases the likelihood of maltreatment recurrence while values less than 1.0 indicate a decreased likelihood of subsequent maltreatment.

Families who had an overall risk rating of "2 or 3" at Time 1 (end of the original maltreatment investigation) were no more likely (Hazard Ratio=0.97, p=0.43) than families who had an initial overall risk rating of "1" to experience a recurrence of maltreatment up to 18 months after case closure (Table 10). Even cases rated initially at the highest overall risk level (4,5) were no more likely than cases rated as lowest risk (1) to have a recurrence (Hazard Ratio=1.42, p=0.50). Similarly, Time 2 overall risk rating

Table 10: Simple Cox Proportional Hazards Models for Overall Risk Rating (Continuous and Categorical) at Time 1 and Time 2

	p-value	Hazard Ratio (95% Confidence Interval)
Overall risk at Time 1		
Overall Risk Rating=1 (reference group)		
Overall Risk Rating=2, 3	0.1379	2.12 (0.79, 5.76)
Overall Risk Rating=4, 5	0.1123	2.25 (0.83, 6.14)
Overall risk at Time 2		
Overall Risk Rating=1 (reference group)		
Overall Risk Rating=2, 3	0.1077	1.24 (0.96, 1.60)
Overall Risk Rating=4, 5	0.4388	1.39 (0.61, 3.18)

was unrelated to the likelihood of maltreatment recurrence 18 months after case closure. Overall risk rating appears to be an invalid predictor of maltreatment recurrence for this sample of families.

Predictive Capacity of Individual RA Tool Items

Using the same method for individual RA Tool items, each was separately entered into a simple Cox Proportional Hazards regression in order to determine which elements predicted recurrence of maltreatment with at least 18 months of follow-up. Items were entered as ordinal variables with a range of 0 "no risk" to 4 "highest risk," and items were considered worthy of further exploration if they obtained a significance level of $p<0.10$.

Beginning with the Time 1 assessment of risk (Table 11), the following variables were identified as possible predictors of abuse: Child behaviour (Hazard Ratio=1.13, p=0.02); History of Abuse/Neglect Committed by Present Caregivers (Hazard Ratio=1.13, p=0.04); and Access to Child by Perpetrator (0.91, p=0.06). That is, for every one unit increase in child behaviour problems as indicated by the caseworker on the initial RA Tool, a child had a 13% greater hazard of experiencing subsequent maltreatment. Access to the child by the perpetrator at Time 1 was associated with an almost 10 percent decrease in the hazard that maltreatment would recur.[3]

Table 11: Time to Abuse by RA Tool Items at Time 1 Using Simple Cox Proportional Hazards Regression

Item of RA Tool at Time 1	p-value	Hazard Ratio (95% Confidence Interval)
Abuse/Neglect of Caregiver	0.0498**	1.11 (1.00, 1.22)
Alcohol or Drug use	0.1556	1.07 (0.97, 1.19)
Caregiver expectations of child	0.4714	1.04 (0.94, 1.14)
Caregiver's Acceptance of child	0.7185	1.02 (0.91, 1.15)
Physical capacity to care for child	0.4354	1.06 (0.92, 1.23)
Mental/Emotional/Intellectual capacity to care for child	0.0988	0.91 (0.82, 1.02)
Child's Influence/Child vulnerability	0.1595	1.08 (0.97, 1.21)
Child's Influence/Child response to caregiver	0.4753	0.96 (0.86, 1.07)
Child's Influence/Child behaviour	0.0198**	1.12 (1.02, 1.23)
Child's Influence/Child Mental health development	0.8177	0.99 (0.88, 1.11)
Child's Influence/Child Physical health development	0.1578	0.89 (0.75, 1.05)
Intervention Caregiver's Motivation	0.5182	1.04 (0.93, 1.16)
Intervention Caregiver's cooperation with Intervention	0.4928	1.03 (0.94, 1.13)
Abuse/Neglect Influence Access to child by perpetrator	0.0553*	0.92 (0.85, 1.00)
Abuse/Neglect Influence Intent and Acknowledgement of Responsibility	0.4651	0.96 (0.85, 1.08)
Abuse/Neglect Influence Severity of Abuse/Neglect	0.2369	0.94 (0.84, 1.04)
Abuse/Neglect Influence History of Abuse/Neglect committed by present caregivers	0.0132**	1.14 (1.03, 1.26)
Family Influence – Family Violence	0.2115	1.05 (0.97, 1.15)
Family Influence – Ability to cope with stress	0.9855	1.00 (0.88, 1.13)
Family Influence – Availability of Social Supports	0.7204	1.02 (0.92, 1.14)
Family Influence – Living conditions	0.5029	1.06 (0.90, 1.24)
Family Influence – Identity and Interactions	0.8943	0.99 (0.90, 1.10)

** p-value<0.10; ** p-value<0.05*

[3] This finding should be regarded with some suspicion and was dropped from subsequent analysis despite a finding of significant effect. For instance, if the child is returned to the maltreating parent by Time 2 (the vast majority of cases), then access to the perpetrator is guaranteed. In other words, one can never return a child to a perpetrator without that perpetrator having full access to the child. Unfortunately, the structure of the data did not allow us to control for the problematic nature of this question.

Time 2 assessment of risk had a greater yield of potentially predictive items (10). Significant predictors of maltreatment recurrence at the p<0.05 level included: abuse/neglect of caregiver (Hazard Ratio=1.12, p=0.05); caregiver alcohol or drug use (Hazard Ratio=1.16, p=0.036); child behaviour (Hazard Ratio=1.26, p<0.01); history of Abuse/Neglect committed by present caregivers (Hazard Ratio=1.19, p<0.01); family violence (Hazard Ratio=1.15, p=0.02); caregiver ability to cope with stress (Hazard Ratio=1.22, p=0.01); and family living conditions (Hazard Ratio=1.28, p=0.04). Significant predictors at the 0.10 level included: child mental health development (Hazard Ratio=1.13, p=0.07); access to child by perpetrator (Hazard Ratio=0.9, p=0.06); family identity and interactions (Hazard Ratio=1.13, p=0.07).

Table 12: Time to Abuse By RA Tool Elements at Time 2 Using Simple Cox Proportional Hazards Regression

Item of RA Tool at Time 2	p-value	Hazard Ratio (95% Confidence Interval)
Family Influence – Identity and Interactions	0.8943	0.99 (0.90, 1.10)
Abuse/Neglect of Caregiver	0.0145**	1.13 (1.02, 1.24)
Alcohol or Drug use	0.0147**	1.17 (1.03, 1.32)
Caregiver expectations of child	0.1682	1.09 (0.96, 1.23)
Caregiver's Acceptance of child	0.5106	1.05 (0.90, 1.23)
Physical capacity to care for child	0.1911	1.11 (0.95, 1.31)
Mental/Emotional/Intellectual capacity to care for child	0.8975	1.01 (0.89, 1.14)
Child's Influence/Child vulnerability	0.1513	1.08 (0.97, 1.21)
Child's Influence/Child response to caregiver	0.6673	1.03 (0.89, 1.20)
Child's Influence/Child behaviour	0.0010**	1.21 (1.08, 1.35)
Child's Influence/Child Mental health development	0.1285	1.10 (0.97, 1.24)
Child's Influence/Child Physical health development	0.6200	0.95 (0.79, 1.15)
Intervention Caregiver's Motivation	0.4047	1.06 (0.93, 1.20)
Intervention Caregiver's cooperation with Intervention	0.9830	1.00 (0.92, 1.10)
Abuse/Neglect Influence Access to child by perpetrator	0.0751*	0.92 (0.83, 1.01)
Abuse/Neglect Influence Intent and Acknowledgement of Responsibility	0.3580	0.94 (0.82, 1.07)
Abuse/Neglect Influence Severity of Abuse/Neglect	0.9691	1.00 (0.89, 1.14)
Abuse/Neglect Influence History of Abuse/Neglect committed by present caregivers	0.0007**	1.21 (1.08, 1.35)
Family Influence – Family Violence	0.0504*	1.11 (1.00, 1.23)
Family Influence – Ability to cope with stress	0.0439**	1.16 (1.00, 1.32)
Family Influence – Availability of Social Supports	0.2669	1.08 (0.94, 1.24)
Family Influence – Living conditions	0.0143**	1.28 (1.06, 1.56)
Family Influence – Identity and Interactions	0.1157	1.10 (0.98, 1.25)

**p-value<0.10; ** p-value<0.05*

After data reduction, multivariate Cox Proportional Hazards Regression was conducted separately for Time 1 and Time 2 using all covariates that reached a p<0.10 level of significance. For Time 1, the presence of child behavior problems (HR=1.12, p=0.018) and prior caregiver history of maltreating behavior (HR=1.13, p=0.018) were associated with a higher likelihood of maltreatment recurrence after

Table 13: Multivariate analysis using Time 1 predictors

Time 1		
Risk Item	p-value	Hazard Ratio
Child Influence/Child behaviour	0.0183	1.12 (1.02, 1.24)
Abuse/Neglect Influence History of Abuse/Neglect committed by present caregivers	0.0179	1.13 (1.02, 1.25)

Time 2		
Risk Item	p-value	Hazard Ratio
Abuse/Neglect of Caregiver	0.0231	1.12 (1.02, 1.23)
Alcohol or Drug use	0.0265	1.16 (1.02, 1.33)
Child Influence/Child behaviour	0.0011	1.22 (1.08, 1.37)
Abuse/Neglect Influence History of Abuse/Neglect committed by present caregivers	0.0015	1.21 (1.08, 1.37)

a minimum of 18 months of follow-up. For Time 2, the presence of child behavior problems (HR=1.22, p=0.001) and prior caregiver history of maltreating behavior (HR=1.21, p=0.002) were also associated with a higher likelihood of subsequent maltreatment, though the magnitude of the association was slightly higher. In addition, Caregiver substance misuse issues (HR=1.16, p=0.023) and a history of abuse/neglect experienced by the caregiver (HR=1.12, p=0.023) were associated with a higher likelihood of maltreatment recurrence.

The individual items identified above were then compared with individual items that showed some degree of reliability between case readers and the original case worker. Table 14 lists the two items of the risk assessment tool which were both reliable and valid. First, child's behaviour was reliable and valid at Time 1 and Time 2. Second, caregiver's alcohol or drug use was reliable but did not meet minimal levels of correspondence when compared to the original case reader (kappa=.28), and was only valid at Time 2.

Table 14: Reliable and Valid Predictors of Maltreatment Recurrence Among Closed CPS Cases – 18 Month Follow-up (N=1,023).

RA Tool Items	% Agree Avg. Case Readers (Kappa)	% Agree Avg. Case Readers v. Orig. Worker (Kappa)	CPH Significant at Time 1	CPH Significant at Time 2
Alcohol or Drug use	* (.40)	* (.28)		*
Child's Influence/Child behaviour	* (.42)	* (.41)	*	*

** p-value<0.05*

Study III: Focus Group

Overall, the volunteers' negative perceptions far outweighed any positive comments about the tool. Two focus groups/expert panel strength-based themes were derived from the data. The dominant theme centered on "RA Tool Familiarity Breeds Worker Comfort" and the other theme was "RA Tool Structures Workers' Information Gathering." Essentially, after eight years of field use, the RA Tool was a familiar tool, therefore workers were comfortable in completing the tool because they completed it with great frequency. The layout of the tool was viewed positively in its ability to structure information and focus workers' data gathering.

The two dominant negative themes were: "RA Tool is a Compliance Tool not a Valuable or Valid Tool". This theme underscored workers experience with the tool as a compliance recording task. Many volunteers commented that the RA Tool was not a tool that informed their assessments, it was not clinically valuable, and they were dubious about its predictive validity. In short—they used it because they had to but did not trust it. The tool's time consuming demands informed the second theme: "RA Tool: A High Workload + Low Relevance Tool". A significant concern of the volunteers was that the RA Tool had a high workload requirement but there with little benefit to the worker (better assessments and interventions), the child (more accurate identification of children at-risk of future maltreatment) or the family (accurate prediction of high-risk families results in more services and interventions to adequately mitigate risk).

In summary, the qualitative, focus group data explored and expanded on the "why" behind the field's adverse views and negative experiences with the tool. The quantitative analysis substantiated workers' concerns with the validity and reliability of the RA Tool.

Discussion

The results of these studies add to a mounting body of evidence that clinical judgment using a structured risk assessment tool is an unreliable foundation on which to base critical child protection decisions. While summary risk assessment ratings proved to be marginally internally consistent across raters in relation to Caregiver Influences, summary Abuse/Neglect Influences achieved acceptable levels of internal consistency in the case of the original intake worker and one of the three cases readers only. Internal consistency fared no better at the subscale level, where only a minority of subscales achieved acceptable levels across raters. Mitigating this finding somewhat is the fact that the original intake workers produced much more internally consistent results than case readers did. In all but two subscales ("access to child by perpetrator" and "severity of abuse/neglect'), intake workers produced acceptable levels of internal consistency across summary scales and subscales alike. This difference between original worker and blind case reader results is presumably attributable to the original worker having access to more information about the family under investigation than was contained in the case notes. Not surprisingly given the results of internal consistency analysis, inter-rater reliability was also found to be marginal, especially when the original worker was included in the analysis. Six of the 22 risk items had modest reliability among case readers and no items attained even limited reliability when the original worker was included. The overall risk scores were generally unreliable, with or without the inclusion of the original caseworker. Again, this disparity may be a function of the difference between deriving risk ratings from a case file versus calculating them in the field. Nevertheless, such findings call into question both the content and structure of the risk assessment items.

When analysis was restricted to intake workers in the predictive validity study, results were again generally unsupportive of Ontario's RA Tool. In the first place, change in risk rating between T1 and T2 was unrelated to likelihood of recurrence by T3; furthermore, overall risk ratings at T1 and T2 were also unrelated to recurrence. There is more than one possible explanation for these results. The most straightforward explanation, of course, is that the factors incorporated in Ontario's RA Tool are genuinely unrelated to risk of future harm. However, it could also be that the link between risk assessment and case closure actually works in reverse; that is, workers do not conduct an impartial risk assessment leading to their case decision, but rather they make their decision to close the case and then record ratings aligned with that decision. Such a phenomenon has been observed before (Lyle and Graham, 2000) and would be expected under conditions where workers are placed under external organizational and/or legal pressure to close cases prematurely or according to some externally imposed timeline. Further research is needed on this issue incorporating a sample of cases that have not yet been closed. Whatever the explanation, the absence of any association between overall risk ratings at T1 or T2 and the likelihood of re-abuse indicates that Ontario's RA Tool, as currently operating, produces inaccurate and therefore unsafe results. Notwithstanding this conclusion, certain individual items drawn from the RA Tool did yield results that would justify including them in future revisions of the instrument. The variables that contained predictive power at T1 included: "child behaviour" and "history of abuse committed by caregiver." At T2, these same predictors re-emerged, along with "caregiver alcohol or drug abuse," "family violence," "caregiver ability to cope with stress," and "family living conditions." However, as structured in the Ontario Risk Assessment Tool, only two of these constructs have acceptable levels of reliability when measured among readers of case files: "child behavior" and "caregiver alcohol or drug abuse."

The individual items contained in the RA Tool are based on common predictors of maltreatment recurrence. The results of this single study do not necessarily discredit any of the 22 risk assessment items. Rather, the study indicates that the way in which these items are presented in the RA Tool and, more importantly, the manner in which the risk assessment tool has been implemented in the field, is neither reliable nor valid. The one risk assessment item that, perhaps, should be discarded is the caseworkers' perception of overall risk. This item had very little reliability and no predictive validity, and it is easy to see why this might be the case. Instead of adding up scores and classifying people into ascending risk categories, this item asks the caseworker to make a clinical judgment based on the other 22 items in the tool. Even if some items are reliable and valid, caseworkers have to weigh the strength of such factors and, inevitably, some degree of reliability and validity are lost.

Limitations

Before concluding, it is important to note a number of important limitations in the present study that limit the scope of its findings. Among the most significant of these is the sampling frame for the study. Rather than obtaining a sample of cases that were investigated for maltreatment and following them over time, we were constricted to using an exit cohort of families who had received services. For this reason, there are two important sub cohorts missing. First, some children received limited services of short duration while others were placed in foster care and returned home. Thus, children had differential exposure to risk that was uncontrolled in the analysis (i.e., some children were exposed to potentially maltreating parents for greater periods of time than others). Second, cases of children who remained in care and stayed open for services during the data collection window were missed. Somewhat related, the structure and content of the data made it impossible to tell whether children were returned to the offending parent, another parent, or a relative. Finally, case record reviews and administrative data have their drawbacks including missing data, insufficient information, and conflicting details. Though these were largely sorted out through double checks of the case files and additional queries of the administrative database, some small portion of the information is likely to be inaccurate.

Despite these limitations, this study is one of the few systematic attempts to assess the measurement properties of a child maltreatment risk assessment instrument and its findings generally corroborate evidence suggesting that consensus-based instruments are ill-suited to the task.

References

Baird, C. (1988). Development of risk assessment indices for the Alaska Department of Health and Social Services. In T. Tatara (Ed.), *Validation Research In CPS Risk Assessment: Three Recent Studies*, pp. 85–139, Washington, D.C.: American Public Welfare Association.

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

Bakeman, R. &. Gottman, J.M. (1986). *Observing Interaction: An Introduction to Sequential Analysis.* New York: Cambridge University Press.

Barber, J.G. (2002). *Evaluation of the Service Improvement Project.* Brisbane, Australia: Queensland Department of Families.

Barber, J.G., Cooper, L. & Delfabbro, P.H. (1999). *First Progress Report on Alternative Care for Children in South Australia.* Adelaide, Australia: South Australian Department of Human Services.

Barber, J.G. & Delfabbro, P.H. (2003). *Children in Foster Care*. London: Taylor and Francis.

Baumann, D.J., Esterline, J.A., Zuniga, G., Smith, S., Whiteside, D., Fluke, J., Goertz, B. & Cohen, M. (1997). The implementation of risk assessment. In Texas Department of Protective and Regulatory Services, *The WISDOM Project*. Austin, TX: Author.

Camasso, M.J. & Jagannathan, R. (1995). Prediction accuracy of the Washington and Illinois risk assessment instruments: An application of receiver operating characteristic curve analysis. *Social Work Research, 19*, 174–184.

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

Dawes, R.M. (1994). *House of Cards: Psychology and Psychotherapy Built on Myth*. New York: Free Press.

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

Doueck, H.J., Bronson, D.E. & Levine, M. (1992). Evaluating risk assessment implementation in child protection: Issues for consideration. *Child Abuse and Neglect, 16*, 637–646.

Edleson, J.L. (2004). Should childhood exposure to domestic violence be defined as child maltreatment under the law. In P.G. Jaffe, L.L. Baker and A. Cunningham (Eds.), *Protecting Children from Domestic Violence: Strategies for Community Intervention*. New York, NY: Guildford Press.

English, D.J. & Graham, J.C. (2000). An examination of relationships between children's protective services social worker assessment of risk and independent LONGSCAN measures of risk constructs. *Children and Youth Services Review, 22,* 897–933.

English, D.J. & Marshall, D. (1996). Correspondence of caseworkers rating of risk and independent caregiver self report measures. *Proceedings of the 10th National Roundtable on CPS Risk Assessment.* San Francisco.

English, D.J., Marshall, D., Brummel, S. & Coghlan, L. (1998). *Decision-Making in Child Protective Services: A Study of Effectiveness. Phase I: Quantitative Analysis*. Washington, DC: National Center on Child Abuse and Neglect.

Falco, G. & Salovitz, B. (1997). *Actuarial Risk Assessment in Child Protective Services: Results From Recent Research in New York*. New York State Department of Social Services.

Fallon, B., Trocmé, N., MacLaurin, B., Knoke, D., Black, T., Daciuk, J., Felstiner, C. (2005). *Ontario Incidence Study of Reported Child Abuse and Neglect – 2003 (OIS-2003): Major Findings.* Toronto, ON: Centre of Excellence for Child Welfare.

Fuller, T.L., Wells, S.J., and Cotton, E.E. (2001). Predictors of maltreatment recurrence at two milestones in the life of a case. *Children and Youth Services Review*, 23 (1), 49–78.

Gambrill, E., & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth Services Review, 22(11/12),* 813–837.

Grove, W.M. & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and form (mechanical, algorithmic) prediction procedures. *Psychology, Public Policy, and Law, 2*, 293–323.

Hindley, N., Ramchandani, P. G. & Jones, D. P. H. (2006). Risk factors for recurrence of maltreatment: A systematic review. *Archives of Disease in Childhood.* Retrieved July 27, 2006, from adc.bmjjournals.com on.

Jagannathan, R., & Camasso, M.J. (1996). Risk assessment in child protective services: A canonical analysis of the case management function. *Child Abuse and Neglect*, 20 (7), 599–612.

Kolko, D.J. (1998). CPS operations and risk assessment in child abuse cases receiving services: Initial findings from the Pittsburgh service delivery study. *Child Maltreatment, 3,* 262–275.

Knoke, D. & Trocmé, N. (2004). *Risk assessment in child welfare*. Toronto, ON: Centre of Excellence for Child Welfare (CECW).

Leschied, A., Chiodo, D., Whitehead, P., Hurlkey, D., & Marshall, L. (2003). The empirical basis of risk assessment in child welfare: The accuracy of risk assessment in clinical judgement. *Child Welfare, 82(5),* 527–540.

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

Lyle, G., & Graham, E. (2000). Looks can be deceiving: Using a risk assessment instrument to evaluate the outcomes of child protective services. *Children and Youth Services Review, 22,* 935–949.

Lyons, P, Doueck, H.J. & Wodarski, J.S. (1996). Risk assessment for child protective services: A review of the empirical literature on instrument performance. *Social Work Research, 20,* 143–154.

Ministry of Community and Social Services (1997). *Risk Assessment Model for Child Protection in Ontario.* Toronto: Queen's Printer for Ontario.

Ministry of Community and Social Services (2000). *Risk Assessment Model for Child Protection in Ontario (Revised 2000).* Toronto: Queen's Printer for Ontario.

Myers, H. (2003). Hospital fall risk assessment tools: A critique of the literature.

International Journal of Nursing Practice, 9, 223–235.

Nasuti, J.P. & Pecora, P.J. (1993). Risk assessment scales in child protection: A test of the internal consistency and interrater reliability of one statewide system. *Social Work Research and Abstracts, 29,* 28–33.

National Child Abuse and Neglect Data System (NCANDS; 2002). *Child Maltreatment.* Washington, DC: U.S. Department of Health and Human Services.

Ontario Association of Children's Aid Societies (OACAS; 2000). *Ontario child welfare: Eligibility Spectrum.* Revised 2000. Toronto, ON: OACAS.

Pecora, P.J. (1991). Investigating allegations of child maltreatment: The strengths and limitation of current risk assessment systems. In M. Robin (Ed.) *Assessing Child Maltreatment Reports: The Problem of False Allegations.* New York: Haworth Press.

Sedlak A.J., & Broadhurst, D.D. (1996). *Third National Incidence Study of Child Abuse and Neglect.* Washington, DC: U.S. Department of Health and Human Services.

Rittner, B. (2002). The use of risk assessment instruments in child protective services case planning and closure. *Children and Youth Services Review, 24,* 189–207.

Rodwell, M.K. & Chambers, D.E. (1989). Promises, promises: Child abuse prevention in the 1980s. *Policy Studies, 8,* 749–773.

Rossi, P., Schuerman, J. & Budde, S. (1996). *Understanding child maltreatment decisions and those who make them.* Chicago: Chapin Hall Center for Children.

Ruscio, J. (1998). Information integration in child welfare cases: An introduction to statistical decision making. *Child Maltreatment, 3*, 143–156.

Rycus, J.S. & Hughes, R.C. (2003). *Issues in Risk Assessment in Child Protective Services: A Policy White Paper.* Columbus, Ohio: North American Resource Center for Child Welfare.

Trocmé, N., Fallon, B., MacLaurin, B., & Copp, B. (2002). *The changing face of child welfare investigations in Ontario: Ontario Incidence Studies of Reported Child Abuse and Neglect (OIS 1993/1998).* Toronto, ON: Centre of Excellence for Child Welfare, Faculty of Social Work, University of Toronto.

Trocmé, N., Fallon, B., MacLaurin, B., Daciuk, J., Felstiner, C., Black, et al. (2005). *Canadian Incidence Study of Reported Child Abuse and Neglect – 2003 (CIS-2003): Major Findings.* Minister of Public Works and Government Services Canada.

Trocmé, N., Fallon, B., MacLaurin, B., Tonmyr, L., & de Marco, R. (2004, September). *Preliminary findings from the 2003 Canadian Incidence Study of Reported Child Maltreatment: Interpreting changes between the 1998 and 2003 cycles*. 15th International Congress on Child Abuse and Neglect. Brisbane, Australia.

Trocmé, N., Mertins-Kirkwood, B., MacFadden, R., Allagia, R., & Goodman, D. (1999). *Ontario risk assessment model Phase 1: Implementation and training.* Toronto, Ontario: University of Toronto, Faculty of Social Work, Bell Canada Child Welfare Research Unit, 69 pages.

Tatara, T. (1987). An overview of current practices in CPS risk assessment and family systems assessment in public child welfare. In *National Rountable on CPS Risk Assessment and Family Systems Assessment: Summary of Highlights*, Washington, DC: American Public Welfare Association.

Wald, M.S. & Woolverton, M. (1990). Risk assessment: The emperor's new clothes? *Child Welfare, 69*, 483–511.

Waldfogel, J. (1998). *The future of child protection: How to break the cycle of abuse and neglect.* Cambridge, Massachusetts: Harvard University Press.

Weedon, J., Torti, T.W., & Zunder, P. (1988). Vermont Division of Social Services Family Risk Assessment Matrix research and evaluation. In T. Tatara (Ed.), *Validation Research In CPS Risk Assessment: Three Recent Studies,* pp. 3–43, Washington, D.C.: American Public Welfare Association.

Wood, J.M. (1997). Risk predictors for re-abuse or re-neglect in a predominantly Hispanic or population. *Child Abuse and Neglect, 21,* 379–389.

Appendix A

Percent Agreement between specific raters

Caregiver Influence – Inter raters' reliability between raters 1,2,3

Variable	Statistics	Rater 1 vs. Rater 2	Rater 1 vs. Rater 3	Rater 2 vs. Rater 3	Average
Abuse/Neglect of Caregiver	Kappa	0.41	NA	NA	0.41
	p-value	0.2011	NA	NA	-
	Percent agreement	84.8%	86.9%	87.8%	86.5%
Alcohol or Drug use	Kappa	0.42	0.40	0.38	0.40
	p-value	0.0048	0.0001	0.0009	-
	Percent agreement	65.2%	65.3%	68.3%	66.3%
Caregiver's Expectations of child	Kappa	0.17	0.22	0.45	0.28
	p-value	<0.0001	<0.0001	0.0283	-
	Percent agreement	34.8%	36.7%	64.6%	45.4%
Caregiver's Acceptance of Child	Kappa	NA	NA	0.48	0.48
	p-value	NA	NA	0.0004	-
	Percent agreement	51.6%	47.8%	72.8%	57.4%
Physical Capacity to care for Child	Kappa	NA	NA	0.19	0.19
	p-value	NA	NA	<0.0001	-
	Percent agreement	71.7%	53.4%	64.6%	63.2%
Mental/Emotional/Intellectual Capacity to care for Child	Kappa	0.27	0.22	0.27	0.25
	p-value	0.0002	<0.0001	<0.0001	-
	Percent agreement	53.7%	43.4%	50.4%	49.2%

Intervention Influence – Inter raters' reliability between raters 1,2,3

Risk Item	Statistics	Rater 1 vs. Rater 2	Rater 1 vs. Rater 3	Rater 2 vs. Rater 3	Average
Caregiver's Motivation	Kappa	0.13	0.27	0.30	0.23
	p-value	<0.0001	<0.0001	<0.0001	-
	Percent agreement	30.3%	42.0%	47.3%	39.9%
Caregiver's Cooperation with Intervention	Kappa	0.22	0.24	0.31	0.87
	p-value	<0.0001	<0.0001	<0.0001	-
	Percent agreement	38.1%	37.3%	48.0%	41.1%

Abuse/Neglect Influence – Inter raters' reliability between raters 1,2,3.

Risk Item	Statistics	Rater 1 vs. Rater 2	Rater 1 vs. Rater 3	Rater 2 vs. Rater 3	Average
Access to child by Perpetrator	Kappa	0.14	0.05	0.25	0.15
	p-value	<0.0001	<0.0001	0.0002	-
	Percent agreement	47.1%	37.6%	73.1%	52.6%
Intent and Acknowledgement of Responsibility	Kappa	NA	NA	0.40	0.40
	p-value	NA	NA	0.0087	-
	Percent agreement	47.3%	40.8%	67.2%	51.8%
Severity of Abuse/ Neglect	Kappa	0.17	NA	NA	0.17
	p-value	<0.0001	NA	NA	-
	Percent agreement	36.5%	35.7%	56.5%	42.9%
History of Abuse/Neglect Committed by Present Caregivers	Kappa	NA	0.18	NA	0.18
	p-value	NA	0.0002	NA	-
	Percent agreement	30.2%	39.3%	38.1%	35.9%

Child Influence – Inter raters' reliability between raters 1,2,3

Variable	Statistics	Rater 1 vs. Rater 2	Rater 1 vs. Rater 3	Rater 2 vs. Rater 3	Average
Child's Vulnerability	Kappa	NA	NA	0.84	0.84
	p-value	NA	NA	0.1581	-
	Percent agreement	72.5%	73.3%	89.1%	78.3%
Child's Response to Caregiver	Kappa	0.31	NA	NA	0.31
	p-value	0.0235	NA	NA	-
	Percent agreement	67.9%	65.9%	74.1%	69.3%
Child's Behaviour	Kappa	0.48	0.37	0.41	0.42
	p-value	0.0414	0.0002	0.0062	-
	Percent agreement	74.4%	68.5%	70.7%	71.2%
Child's Mental Health and Development	Kappa	0.43	0.33	0.40	0.39
	p-value	0.0391	0.0001	0.0025	-
	Percent agreement	72.9%	67.8%	71.1%	70.6%
Child's Physical Health and Development	Kappa	NA	0.34	NA	0.34
	p-value	NA	0.1756	NA	-
	Percent agreement	84.3%	82.9%	86.8%	84.7%

Family Influence – Inter raters' reliability between raters 1,2,3

Variable	Statistics	Rater 1 vs. Rater 2	Rater 1 vs. Rater 3	Rater 2 vs. Rater 3	Average
Family Violence	Kappa	0.33	0.30	0.32	0.32
	p-value	0.0080	0.0003	0.2439	-
	Percent agreement	47.3%	43.6%	49.2%	46.7%
Ability to Cope with Stress	Kappa	0.19	NA	NA	0.19
	p-value	0.0107	NA	NA	-
	Percent agreement	36.4%	17.3%	40.0%	31.2%
Availability of Social Supports	Kappa	0.21	NA	NA	0.21
	p-value	<0.0001	NA	NA	-
	Percent agreement	36.4%	36.1%	55.4%	42.6%
Living Conditions	Kappa	NA	NA	NA	
	p-value	NA	NA	NA	-
	Percent agreement	72.9%	74.4%	79.2%	75.5%
Family Identity and Interactions	Kappa	NA	0.18	NA	0.18
	p-value	NA	0.0003	NA	-
	Percent agreement	41.1%	31.6%	29.2%	34.0%