Cover Page

Title
The recurrence of child maltreatment: Predictive validity of risk assessment instruments.

Reviewers

Aron Shlonsky, Associate Professor
University of Toronto
Faculty of Social Work
246 Bloor St. W.
Toronto, Ontario M5S 1A1, Canada
Phone: (416) 978-5718
Fax: (416) 978-7072
email: aron.shlonsky@utoronto.ca

Michael Saini
Systematic Review Specialist
University of Toronto
Faculty of Social Work
246 Bloor St. W.
Toronto, Ontario M5S 1A1, Canada
Phone: (416) 978-5718
Fax: (416) 978-7072

Statistician
Meng-Jia Wu
Assistant Professor,
Research Methodology
Loyola University Chicago School of Education
820 N. Michigan Ave., Lewis Towers 11th Floor
Chicago, IL 60611
Phone: (312) 915-7086
email: mwu2@luc.edu

Sources of support
The Jessie Ball DuPont Foundation
Bell Canada Child Welfare Research Center
Nordic Campbell Centre

Submitted
Nov. 5, 2007

Contents

Protocol for a Systematic Review of
The Recurrence of Child Maltreatment:
Predictive Validity of Risk Assessment Instruments

2. Background for the Review

Accurate and timely identification of abused and neglected children who are at risk of further maltreatment is essential for effective targeting of child protection services. Risk assessment in child protection generally involves estimating the likelihood of future maltreatment for parents alleged to have maltreated their children and deciding whether protective measures are needed to reduce the risk of future harm (Rycus & Hughes, 2003). Incorrect decisions can lead to any number of detrimental outcomes including leaving children in potentially dangerous environments or providing families with over-intrusive child protection services when children are not at risk of future harm (Gambrill & Shlonsky, 2000).

Clinical Judgement

During the process of investigating allegations of child maltreatment, social services workers have historically relied upon their own professional experience, intuition and individual heuristics to evaluate the potential for recurring child maltreatment (Gambrill & Shlonsky, 2000), but these may be inadequate predictors of future harm. Clinical estimates of risk are compromised by varying contextual and individual circumstances that quickly become too complicated to evaluate objectively (Dawes, 1993; 1994). Circumstances contributing to uncertainty can include differences in parenting styles, cultural variations of parenting, environmental characteristics, the values and policies of agencies and the broader community towards child maltreatment, worker training related to decision making and the organizational context (Gambrill & Shlonsky, 2000; 2001). Mistakes at any decision point have serious implications for children and families as inaccurate identification of risk can lead to subsequent maltreatment and/or unwarranted separation between the children and their families (Shlonsky & Wagner, 2005).

Standardized Risk Assessment

In an effort to address the shortfalls of clinical judgment, standardized risk assessment measures have been developed in child welfare to formalize the prediction of maltreatment recurrence and to improve the accuracy of estimating future risk (Rycus & Hughes, 2003; Shlonsky & Wagner, 2005). Standardized risk assessment instruments have existed for many years in other practice fields, including medicine, health, and criminology with modest to positive results (Gottfredson & Moriarty, 2006; Meehl, 1954).

Child maltreatment risk assessment instrument instruments seek to rank or segregate families based on the positive and negative likelihood of maltreating children in the future. Standardized risk assessment tools differ regarding their scope, content and construction, but they share a common goal: to help social service workers more accurately assess the likelihood of child maltreatment based on the presence of certain family characteristics and environmental conditions considered to be highly associated with maltreatment (Rycus & Hughes, 2003; Shlonsky & Wagner, 2005).

Standardized tools generally involve four separate functions: 1) screening for potential maltreatment from the general population; 2) screening for the presence of maltreatment in cases under investigation by child protection services; 3) assessing the risk of recurrence of maltreatment in populations already investigated by child protection services; and 4) assessing the risk of maltreatment among children who have been returned to their parents after residing in foster care. The difference between screening (assessing potential for maltreatment before it occurs) and assessing risk of recurrence in cases where there has already been an allegation of child maltreatment is crucial since the populations assessed and their risk of maltreatment differ, as do and relevant predictive factors associated with each (Cash, 2001). This review will focus solely on risk of maltreatment recurrence for children being investigated by child protection services. While the other types of assessment do occur in practice, assessing for recurrence at the point of investigation is arguably the most important and commonly employed prognostic process in child welfare services involving maltreatment.

There are two major approaches to standardized risk assessment of subsequent child maltreatment: consensus-based and statistically-based models. Although both models generate a list of characteristics (e.g., caregiver and child characteristics or attributes, abuse circumstances, or environmental circumstances) that should predict an outcome of interest (e.g., the recurrence of abuse), they are developed in different ways.

C*onsensus-based models* are usually developed by expert opinion of risk, best available theories of child maltreatment, and best available research on the recurrence of maltreatment. Risk factors derived from these sources are operationalized as items on the instrument, usually with four or more categories or risk levels per item. Consensus-based instruments tend to use a single instrument to predict all forms of child maltreatment. These instruments can be assessed numerically by categorizing families by their total score or they can be assessed by coding items as high, moderate or low based on the worker's judgement (Austin, D'Andrade, Lemon, Benton, Chow, & Reyes, 2005; English, 1999). Austin et al., (2005) highlight that consensus-based models have been criticized because of 1) poor conceptualization; 2) inconsistencies in the type of number of variables included; 3) use of the same variables to predict physical abuse, neglect, and sexual abuse; and 4) over reliance on maltreatment with little attention to the recurrence of maltreatment.

Statistically-based models incorporate client characteristics shown to be statistically predictive of future maltreatment (Rycus & Hughes, 2003). Examples of the latter include actuarial models, configural models, and neural network models. Statistical models are generally developed by taking a sample of children and families involved in the child welfare system, analyzing their paths in the system, relating those paths to a set of family characteristics or events, and identifying those characteristics and events which are highly associated with an outcome of interest – usually, the recurrence of maltreatment (Gambrill & Shlonsky, 2000).

Both types of **i**nstruments also differ in the number and type of risk criteria included and in approaches to presenting and aggregating risk information, and there remains no consensus about which elements are required to accurately assess risk. For instance, McDonald and Marks (1991) found that a large proportion of the risk items investigated in their seminal review had

not been empirically validated. As well, many of the commonly used items composing the risk assessment portion of the National Survey of Child and Adolescent Well-Being are not predictive of subsequent child maltreatment (Shlonsky, 2007). Statistically-driven approaches can rule out non-predictive factors, but they still require decisions about which factors to include prior to modeling.

Predicting the Occurrence or Recurrence of Maltreatment

Lyons, Doueck and Wodarski (1996) suggest that there is no reason to assume that the same variables predict both the occurrence and recurrence of child maltreatment. A comparison between initial and recurring maltreatment suggests that factors of initial maltreatment are more likely associated with maternal and paternal depression; substance abuse; unemployment; social isolation; unrealistic expectations of the child; parent's history of being abused; and increased stress (Cash, 2001). Other potential predictors of initial maltreatment, such as cultural factors (Lyons, Doueck & Wodarski, 1996), have also been proposed but have not been rigorously tested.

In contrast, factors of recurring maltreatment are more likely related to the parents' previous involvement with child protection services; parents' unrealistic expectations of the child; the child's level of fear towards the perpetrator; and the child's contact with the perpetrator, neglect, parental conflict; and parental mental health problems (DePanfilis & Zuravin, 1998; English, Aubin, Fine, & Pecora, 1993; Johnson & L'Esperance, 1984). Hindley, Ramchandani and Jones (2006) examined 15 studies from the United States and one study from Australia and found the four most consistent factors of the risk of recurring maltreatment included: the number of previous episodes of maltreatment; neglect (as opposed to other forms of maltreatment); parental conflict; and parental mental health problems. Although the majority of the research has been published in the United States, Wagner (1997) found that risk factors identified in protective service research projects in the United States appear to have very similar relationships to the reoccurrence of maltreatment when examined within a South Australian sample.

Utilization of Risk Assessment Tools in Child Welfare

During the 1980's risk assessment models for decision-making in child protection were developed to provide guidelines for practice, optimize the use of available resources, and provide a rationale for targeting scarce resources. Based on the potential for these standardized tools to improve the reliability, validity and objectivity of human service work, there has been a trend in child protection services agencies toward adopting standardized risk assessment instruments to help social service workers better estimate the likelihood of maltreatment recurrence (Munro, 2004). For example, a survey of 42 states in 1996 found that three quarters had used some form of risk assessment to determine the future likelihood of maltreatment for cases involved with child protective services (Tatara, 1996), and that number has probably increased over time (Gambrill & Shlonsky, 2000). Risk assessment procedures have also been initiated and evaluated to predict recurrence of child maltreatment in the UK, Australia, New Zealand and Canada (Trocmé, Barber, Goodman, Shlonsky & Black, 2007; Browne & Saqi, 1988, Stone, 1993; Dalgleish & Drew, 1989; Muir, Monaghan & Gilmore, 1989; Reid, Sigurdson, Wright, & Christianson-Wood 1996).

Some of the more commonly known models include the Washington Risk Assessment Matrix (WRAM); the California Family Assessment Factor Analysis (CFAFA or the "Fresno" model); the Alaska model, the Child at Risk Field (CARF) Matrix; the child Emergency Response Assessment Protocol (CERAP); the Child Well-Being Scales or the Family Risk Scales; Risk Assessment Model (RAM); and the actuarial Risk Assessment instruments developed by the Children's Research Center (CRC) (Lyons, et al., 1996; Pecora, 1991; National Resource Center on Child Abuse and Neglect, 1994). Yet many of the tools used in child welfare services lack a theoretical foundation and are not based on rigorous evidence (Cash, 2001). Aside from the actuarial approach, risk assessment developers did not exclusively use valid and reliable predictors of the recurrence of maltreatment and possibly paid little attention to the consequences of using such tools (see also English et al., 1993). Rycus and Hughes (2003) point to the continued lack of agreement regarding the proper scope and purpose of risk assessment technology in child protection assessment and case planning activities. Others have connected these problems to a broader systemic issue within the field of child protection; the weak connection, emphasis, development and application of scientific foundations to guide decision-making and program development (Melton & Flood, 1994; Wald & Woolverton, 1991).

Psychometric Properties of Maltreatment Risk Assessment Instruments

Although the evidence for specific risk assessment models is mounting, there remain few research studies that directly compare the relative validity of different systems (Baird & Wagner, 2000). Noteworthy exceptions include: Lyons, Doueck and Wodarski's (1996) study of 10 risk assessment models; Camasso and Jagannathan's (1995) study of 2 risk assessment models; and Baird and Wagner's (2000) study of 3 commonly used risk assessments.

Lyons, Doueck, & Wodarski (1996) reviewed 10 published risk assessment studies to determine the degree to which they were able to correctly predict which children would be maltreated and which would not. In general, they found varying rates of false positives and false negatives, but none of the tools had excellent psychometric properties. Camasso and Jagannathan (1995) assessed the 13-item Illinois CANTS 17B and the 32-item Washington State Risk Assessment Matrix (WRAM) and found that maltreatment cases could be correctly classified approximately 67 per cent of the time using these consensus-based models. In a subsequent study, Camasso and Jagannathan (2000) examined only the 32-item Washington Risk Assessment Matrix (WRAM) and found that its capacity to predict maltreatment was low, and many of its items were not significant predictors of maltreatment. Based on their findings, the authors suggested that risk assessment tools should avoid the "laundry list" approach (p. 893) and should instead be constructed with careful consideration of relevant risk factors.

Following this more parsimonious approach, Baird and Wagner (2000) compared the WRAM and a derivation of the CANTS 17B (the California Family Assessment Factor Analysis (CFAFA)), both consensus-based instruments, with Michigan's Family Risk Assessment of Abuse and Neglect (FRAAN), a CRC actuarial approach. A wide range of studies from many fields have found that actuarial (statistically-driven) prediction is at least as good, and most often better, than unassisted clinical prediction (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996), and this study was no exception. FRAAN's actuarial approach substantially outperformed the other tools in terms of correctly classifying high risk families who later

maltreated their children. Extending these findings further, Johnson's (2005) prospective validation of CRC's California Family Risk Assessment found that the instrument was able to correctly classify potentially maltreating families into low, moderate, high, and very high risk categories at levels beyond chance.

Based on this initial review, there remains little evidence regarding the predictive validity of risk assessment instruments measuring maltreatment recurrence in child protection cases. This systematic review will collect a comprehensive set of studies on the predictive validity of various risk assessment models and will synthesize the information, wherever possible, using statistical methods outlined in this protocol.

Definitions for the Review

Child Maltreatment
Definitions of child maltreatment vary (Giovannoni, 1989), but most include acts of commission or omission towards a child that result in, or can result in, harm to the child's physical or psychological wellbeing. Child maltreatment is often categorized into specific acts of physical abuse, sexual abuse, neglect, and emotional abuse including witnessing domestic violence.

Prospective Studies
Risk Assessment completed at the beginning of contact with family and then case followed over time.

Retrospective Studies
Researcher goes back in time and creates an artificial construct on information available at the time of first involvement (only what was known at the time of first involvement). Examples would be case notes, administrative data, collateral contact,etc.

Predictive validity
To ensure accurate predictive validity (prognostic accuracy)[1] of risk assessment tools, several prodecures can be used including: sensitivity, specificity, positive predictive value, negative predictive value, receiver operating characteristics (ROC) curves, and area under the curve (AUC)

Sensitivity
Sensitivity is an index of the performance of a procedure, calculated as the percentage of individuals who have subsequently maltreated a child and who were correctly classified as being at high risk for subsequent maltreatment. It is the conditional probability of a positive result given that the recurrence of maltreatment is present (Everitt, 1998). For families being

[1] Commonly used in the behavioral and social sciences, the term predictive validity refers to that ability of a measure to accurately predict a future state. Prospective studies are usually conducted to assess predictive validity. This is distinguished from concurrent validity, which refers to the simultaneous agreement between a measure being tested and a measure with established validity (Kazdin, 2003; Everitt, 1998). As used here, the term has the same meaning as prognostic accuracy.

investigated by child protection agencies, sensitivity refers to the instruments's accuracy in initially identifying families that actually have a subsequent recurrence of child maltreatment.

Specificity
Specificity is an index of the performance of an instrument, calculated as the percentage of individuals who have not subsequently maltreated a child who are correctly classified as having not subsequently maltreated a child. It is the conditional probability of a negative result given that the recurrence of maltreatment is absent (Everitt, 1998). For families with previous or current involvement with child protection services, specificity refers to the tool's accuracy in identifying families that do not return.

Positive predictive value
Positive predictive value refers to the proportion of people with a positive result on a risk assessment instrument who actually maltreat their child (Everitt, 1998).

Negative predictive value
Negative predictive value refers to the proportion of people with a negative result on a risk assessment instrument who do not maltreat their child (Everitt, 1998).

Receiver Operating Characteristics Curves (ROC)
A ROC curve is a plot of the sensitivity of an instrument against one minus its specificity as the cut-off criterion for indicating a positive result is varied (Everitt, 1998).

Area Under the Curve (AUC)
AUC is the (positive) area under the ROC curve and is expressed as the integral of the ROC curve over 1 – specificity. AUC can be interpreted as the probability that a randomly drawn score from one sample or population (e.g., reabusers' scores) is higher than a randomly drawn score from another sample or population (e.g., nonreabusers' scores) (Rice & Harris, 2005). This measure may be useful for measuring classification systems containing ordered categories of risk with more than two elements (i.e., low, medium, high).

Risk Assessment Instruments that may be included in the review

Several studies have assessed the validity of risk assessment instruments that have been implemented to measure subsequent maltreatment (Baird, 1988; Baird and Wagner, 2000; Browne & Saqi, 1988; Camasso & Jagannathan, 1995, 2000; Dalgleish & Drew, 1989; English, Marshall, Brummel, & Coghlan, 1998; Johnson & L'Esperance, 1984; Lyons, Doueck, & Wodarski, 1996; Marks & McDonald, 1989; Muir, Monaghan & Gilmore, 1989; Reid, Sigurdson, Wright, & Christianson-Wood 1996; Stone, 1993; Trocmé, Barber, Goodman, Shlonsky & Black, 2007; Weedon et al., 1988). In Stowman and Donahue's (2005) review of the literature, they included CPS risk assessment models, measures that focus on the environment of the child, parent self-report measures, and clinical interviews and observations. While it is not possible to estimate the number of instruments to be included in the review, the following instruments are reflective of the types of risk assessment tools and their associated studies that may be included in our review.

1) Risk assessment instruments (ie. Washington State Risk Assessment Matrix (Palmer, 1988; Child Endangerment Risk Assessment Protocol (Illinois Department of Children and Family Services, 1996; Child Well-Being Scales (Magura & Moses, 1986); The Ontario Child Neglect Index (Trocme, 1996); Child at Risk Field System (Holder & Corey, 1987 and Holder & Corey, 1989); Family Risk Scales (Magura & Moses, 1986); California Family Risk Assessment (Baird & Wagner, 2000); Michigan Family Risk Assessment (Baird, Wagner, Caskey & Nuenfeldt, 1995).

2) Measures that focus on the environment of the child (ie. Home Accident Prevention Inventory (Tertinger, Greene, & Lutzker, 1984; Checklist for Living Environments to Assess Neglect (Watson-Perczel, Lutzker, Greene, & McGimpsey, 1988); Home Safety and Beautification Tour (Donohue et al., 1997)

3) Parent self-report measures (ie. Child Abuse Potential Inventory (Milner, 1986); Conflict Tactics Scale, Parent to Child version (Straus, Hamby, Finkelhor, Moore, & Runyan, 1998).

4) Observation measures (Child Abuse and Neglect Interview Schedule (Ammerman, Hersen, & Van Hasselt, 1988); Childhood Level of Living Scale (Polansky, Chalmers, Buttenwieser, & Williams, 1981)

Contribution of this review

Although structured risk assessment has been shown in other fields to hold promise, a more extensive and systematic approach to the development and testing of child maltreatment risk assessment tools is needed to support child welfare practice (Knoke & Trocmé, 2005). This project will systematically compare the validity of risk assessment tools that have been used to estimate the likelihood of recurring child maltreatment. The review will assist child welfare agencies aspiring to improve their decision-making capacity with respect to preventing the recurrence of maltreatment.

Ultimately, the goal for this systematic review is to translate results from meta-analyses or other assessments of the collected studies into actionable recommendations for practitioners, decision-makers, and other potential audiences interested in real-world applications. By performing sub-group analysis, as outlined in this protocol, it is agencies, researchers, and policy-makers will be able to make an informed choice among available risk assessment instruments based upon which tool exhibits the highest predictive validity while considering differences in population and settings..

3. Objectives of the Review

The objective of this review is to assess the predictive validity (prognostic accuracy) of risk assessment instruments used to predict the recurrence of child maltreatment.

We aim to:
1) identify all relevant published and unpublished studies,

2) synthesize the evidence on the predictive validity of child maltreatment risk assessment instruments, and
3) identify major gaps in the literature in order to guide future research efforts.

4. Methodology

Criteria for inclusion and exclusion of studies in the review
The inclusion and exclusion criteria have been defined (below) in relation to the objectives of the systematic review. An affirmative answer to four key questions will determine whether the study will be included or excluded in the review:

1) Was a risk assessment instrument included as part of the study?
2) Was risk of child maltreatment recurrence assessed?
3) Was actual maltreatment recurrence (e.g., re-report, substantiated/indicated re-report, self-disclosure) assessed?
4) Was an appropriate statistical analysis used OR are there sufficient data reported to conduct such an analysis (as described below)?

Characteristics of risk assessment instruments
This review will include instruments designed to assess risk of recurrence of child maltreatment during and after involvement with child protection services. Risk assessment tools will include any instrument designed to predict the risk of recurrence of child maltreatment. This systematic review will exclude initial screening tools, diagnostic screening tools and risk assessments designed to predict events other than maltreatment (e.g., adoption disruption, juvenile delinquency). However, modified and unmodified instruments originally designed to screen for initial maltreatment and then used it to predict recurrence will be considered but these will be analysed separately

Characteristics of the target population
The population of interest will include families of children (under age 18) whose caregiver has previously been investigated for child maltreatment (regardless of the outcome of that investigation) and who continue to live with (or may resume living with) that caregiver.

Types of studies
Only studies using an instrument to predict recurring child maltreatment will be included. Studies will be excluded if they are implementation or construction studies lacking information about the instrument tested or its predictive validity. Both retrospective and prospective studies will be examined, though prospective studies will be considered to have met a higher standard.

For the overall predictive capacity of the instrument, studies will need to include sufficient data for calculation of one or more of the following measures of validity: sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratios, TAU, ROC curves, and area under the curve**.**

Study quality

Study quality will be assessed using the items in the Standards for Reporting of Diagnostic Accuracy (STARD) checklist and flow diagram. The purpose of the STARD initiative is to improve the quality of the reporting of diagnostic studies (Bossuyt et al., 2003a, 2003b). Specifically, we will evaluate study quality by considering the study population (the inclusion and exclusion criteria, setting and locations where the data was collected); participant recruitment (whether participants were recruited based on previous occurrences of maltreatment, all cases, random cases or other strategies); participant sampling (whether the study population was based on a consecutive series of participants defined by the selection criteria or other selection strategies); data collection (whether the data collection planned before the instrument was performed [prospective study] or after [retrospective study]); the reference standard and its rationale (whether actuarial or consensus); technical specifications of materials and methods involved (including how and when measurements were taken); definition of and rationale for the units, cutoffs and/or categories of the results of the risk assessment instrument and the reference standard; the number, training and expertise of the persons executing the instruments; whether the readers of the risk assessment instruments were blind to the results to other tests and any other clinical information available to the readers; methods for calculating or comparing measures of predictive accuracy and the statistical methods used to quantify uncertainty (i.e. 95% confidence intervals); methods for calculating test reproducibility; when the study ws done (including beginning and ending dates of recruitment); clinical and demographic characteristics of the study population; the number of participants satisfying the criteria for inclusion that did or did not undergo the risk assessment instrument; time interval from the risk assessment and any treatment administered between; distribution of severity of maltreatment; a cross tabulation and/or the distribution of the results of the risk assessment instrument (including indeterminate and missing results); any adverse events from performing the risk assessment instrument; estimates of predictive accuracy and measures of statistical uncertainty (i.e. 95% confidence intervals); how indeterminate results, missing responses and outliers of the results were handled; estimates of variability of predictive accuracy between subgroups of participants, readers or centers; estimates of test reproducibility; and whether there was discussion about the clinical applicability of study findings.

The critical appraisal of studies will be undertaken independently by two reviewers to minimize the risk of bias and error (see Details of Study Coding Categories).

Outcome measures

The primary outcome is the recurrence of child maltreatment. Acceptable measures include official reports and administrative data on the substantiation (or confirmation or indication) of an allegation of child maltreatment that occurred within at least 12 months of the initial investigation of maltreatment that brought the case into the study cohort. There are three ways to assess whether maltreatment occurred: Subsequent substantiated allegation (i.e., confirmed or indicated by the child welfare agency) of maltreatment (the allegation can come from single or multiple informants such as teachers, doctors, or other mandated reporters); caseworker assessment of whether maltreatment recurred (as distinguished from their response to an allegation); and self-report by a child, parent, or caregiver. For the purposes of this review, all three measures will be considered but substantiated maltreatment recurrence will be deemed the gold standard. Although the other measures may be valid indicators of maltreatment

recurrence, substantiation requires that the maltreatment conform to a set of legal standards that, while they may vary by jurisdiction, are probably less prone to bias (i.e., parents or children may not disclose maltreatment, only older children could be asked whether they were maltreated). Nonetheless, with proper specification, the other sources of information may be acceptable.

Risk assessments are conducted at both the individual and family-levels. That is, while most risk assessment instruments provide a risk rating or classification for the entire household, others give separate ratings or classifications for each child living in that household. Thus, it is important to specify whether the recurrence of maltreatment is related to the child who was first harmed, whether another child has been harmed and whether it is the same or different parent or guardian who has committed harm. If necessary, we will separate our synthesis based on the unit of analysis of the included studies based on the following criteria:

1) recurrence of maltreatment towards the child who was initially maltreated
2) recurrence of maltreatment towards any child in the same household

In addition, there may be differences in rates of recurrence by type of maltreatment, with child neglect likely associated with a higher rate of relapse (Hindley, Ramchandani, & Jones, 2006). Some instruments specify types of maltreatment (e.g., Baird & Wagner, 2000), while others do not. Therefore, it may be necessary to further divide the review by maltreatment type (i.e., sexual abuse, physical abuse, neglect, etc.).

Search strategy for identification of relevant studies

Literature search strategy for identification of appropriate studies

Both published and unpublished work will be considered eligible for the review. A Trial Search Coordinator will be responsible for coordinating this activity. The search will not be restricted to any single language or nationality.

Because risk assessments were first created and implemented in the child protection field in the 1980's[2], the search strategy will include only studies from 1980 to the present.

To ensure the search strategy has international coverage, the Nordic Campbell Center will be assisting in the searching of relevant studies for the review. The Nordic Campbell Center has agreed to fund a Nordic researcher, Ulla Jergeby of the Institute for Evidence-based Social Work Practice in Sweden, to review studies and to help with data synthesis.

Since studies regarding risk assessment have been published in a variety of journals ranging from social work to medicine, a comprehensive review of the databases will be performed. The search for the published literature will be conducted using the following databases:

1. Psychological Abstracts (PsycINFO, PsycLIT, ClinPsyc-*clinical subset*)
2 MEDLINE
3. EMBASE

[2] The first known validation study of a risk assessment instrument for predicting maltreatment recurrence was Johnson and L'Esperance, 1984.

4 Database of reviews of effectiveness (DARE online),
5. ChildData (child health and welfare)
6. ASSIA (applied social sciences)
7. Caredata (social work)
8. Social Work Abstracts
9. Child Abuse, Child Welfare & Adoption
10. Cochrane Collaboration
11. C2-SPECTR
12. Social Sciences Abstracts
13. Social Service Abstracts
14. Social Science Citation Index
15. Dissertation Abstracts

To ensure maximum sensitivity and specificity, subject headings and word text will be searched using a systematic process. Search terms related to child abuse have been adapted from Barlow et al.'s (2006) review of parenting programmes for the treatment of physical child abuse and neglect

1. Child/
2. Child Abuse/
3. "exp" Child Abuse
4. Risk Assessment/
5. "exp" Risk Assessment
6. "2" and "4"
7. "3" and 6"
8. (infan$.mp. or child$.mp. or teen$.mp. or adolesc$.mp. or minor$.mp. or toddler$.mp. or baby.mp. or babies.mp.)
9 (abus$.mp. or maltreat$.mp. or neglect$.mp. or harm.mp. or physical abuse.mp. or sex$ abuse.mp. or witness$.mp. or child$ protect$.mp. or child$ welfare.mp. or threat.mp. or danger.mp.)).tw.
10. (risk assessment$ or risk tool$ or risk measure$ or risk evaluat$ or risk analys$ or assessment of risk).mp.
11) ("8 AND "9" and "10")
12) "11" and ("1" or "2" or "3" or "8")

A Systematic Information Retrieval Coding Sheet (SIRC) has been developed (see Appendix 1) to record each search for the review. The SIRC will be used to log results for each database and grey literature searched and will include:
1) The date(s) of the search;
2) The name of the researcher
3) The database used for the search;
4) The specific search terms used in combination (including limiters and expanders)
4) The number of results for each search strategy.

The SIRC will allow for replication of the search strategy because each search will be recorded. Furthermore, the search strategy will be saved and "copied and pasted" into the review to avoid editing errors.

In addition to C2-SPECTR and other bibliographic databases noted above, the following sources will also be searched for relevant studies:

Reference lists
Reviewers will check the reference lists of all relevant articles that are obtained, including those from previously published reviews. Potentially relevant articles that are identified will be retrieved and assessed for possible inclusion in the review.

Personal communication
Face-to-face discussions at meetings, emails, requests on list-serves, and formal letters of request for information from authors, presenters and experts in the child welfare field will be solicited to assist the review team to locate relevant studies. A list of the inclusion criteria for the review, along with a sample of relevant articles, will be sent to these key informants along with the request for studies.

Handsearching journals
Journals relevant to child welfare and risk assessment will be handsearched by trained researchers to uncover relevant studies not found by electronic database searches. In addition, trained reviewers will search reference lists of relevant articles. These include, but are not limited to:

1) Child Welfare
2) Children and Youth Services Review
3) Social Service Review
4) Child Maltreatment
5) Child Abuse and Neglect
6) Journal of Social Services Research
7) Social Work
8) Research on Social Work Practice
9) Social Work Research
10) Child Abuse Review

Grey Literature
Special attention will be made to search and collect relevant studies captured in the grey literature. Specifically, the review will include the following strategies to locate articles: 1) Conference Proceedings; 2) Research Reports; 3) Government Reports; 4) Book Chapters; 5) Dissertations; 6) Policy Documents; 7) Personal Networks; and 8) Research Organizations' Web Sites. Grey Literature web-based sites will be searched to uncover this unpublished literature, such as Grey.Net (http://www.greynet.org/index.html) and GrayLit Network (http://graylit.osti.gov/). In addition, a search will be performed within child welfare clearinghouse websites such as the USDHHS funded Child Welfare Information Gateway at http://www.childwelfare.gov/index.cfm. Given that much of the early research on risk

assessment instruments have been described in conference proceedings or governmental reports but not published in peer reviewed journals, it will be likely that grey literature documents will be located via these clearinghouses but not located elsewhere. Search terms will also be entered into the internet search engine *Google* (www.google.com) to find studies relevant to risk assessment that may not be accessible through peer-reviewed journals.

Description of methods used in primary research.
The methods used by the studies in this review are likely to vary in research design, number and types of participants, and construction of outcome measures. Given that there are various ways to obtain data documenting the recurrence of maltreatment (e.g., child/youth report, parent report, child protection worker report) as well as various gradations of certainty (e.g., confirmed or substantiated maltreatment compared with report only), it will be important to discuss these differences and, in all likelihood, to conduct separate analyses for each type of reporter and outcome.

Criteria for determination of independent findings.
The outcome measure for this review will be the recurrence of maltreatment. Where individual studies report multiple outcome measures of maltreatment, each of these outcomes will be coded and analyzed separately. Where outcomes are common across several studies, attempts will be made to synthesize the data.

Key decisions
Guided by Trialstat! Version 4.0, the data abstraction forms will cover three separate stages for studies to pass to be included in the final review.

The first stage will consist of an initial screening to be used to quickly determine whether a study might be appropriate for the review based on the study's title, abstract and bibliographical information. The purpose of this initial screening stage is to include all possible relevant studies of risk assessment tools related to recurring maltreatment.

The second stage will consist of a strict screening form where reviewers will be given full copies of studies to determine whether studies should remain in the review. For studies to pass this stage, each must: 1) include a risk assessment instrument as part of the study; 2) assess risk of child maltreatment recurrence; 3) measure actual maltreatment recurrence (e.g., re-report, substantiated/indicated re-report, self-disclosure); and 4) use an appropriate statistical analysis OR provide sufficient data to conduct such an analysis.

The third stage will consist of the Critical Appraisal Form (see Appendix 2) to extract information from the articles that have made it past the two previous screenings.

Data extraction
Both the principal reviewer and the systematic review specialist will pilot test the data abstraction forms. Inter-rater reliability on the inclusion and exclusion of studies will be

evaluated at each stage. Specifically, blocks of 10 studies will be separately coded until reliability at or above .8 is reached as measured by Cohen's Kappa (k) statistical test.[3] When assessing inter-rater reliability, working with simple percentages tends to cause a problem with accuracy. For this reason, SRS uses Cohen's Kappa to calculate inter-rater reliability when dealing with the consequential conflicts (conflicts that impact the inclusion and exclusion at the liberal and strict levels of references) between reviewers. The formula for Cohen's Kappa (k) is as follows:

$k= (Po-Pc)/(1-Pc)$

Po- is the observed proportion of agreement
Pc- proportion predicted by chance

Kappa scores of less than 0.8 are highlighted in red since 0.81-1.00 is generally considered to be "almost perfect agreement" (Sim, & Wright, 2005).

Differences in coding will then be resolved in order to enhance inter-coder reliability. This approach will also allow the review team to make any necessary modifications to the quality study form to ensure all relevant data are being captured within the template (for example, demographic data found within the studies). Following the pilot test, research assistants will be trained to reliably use the same procedure. Inter-coder reliability will also be checked at regular intervals by examining Cohen's Kappa among pairs of raters for each item in the abstraction forms. If either measure falls below our threshold for high correspondence (0.8), corrective action will be taken (i.e., the item will be evaluated and, if necessary, further defined and coders re-trained to acceptable levels of reliability).

The study details will be extracted by two independent reviewers using a standardized data extraction sheet (see Appendix 2). Any discrepancies will subsequently be resolved by referral back to the source of the material. The results will be presented in standardized, structured tables. The studies will be examined to ensure that all the relevant data for each are recorded.

The reviewers will attempt to contact the authors of studies that are missing key data deemed essential for the review.

Details of study coding categories
1) Study: Information will be collected regarding the author(s); year of publication; source; country; and language
2) Description of Risk Assessment Tool: the name and acronyms of the risk assessment tool; the nature and purpose of the tool (to assess risk of recurrence of maltreatment); the method of the risk assessment (clinical, consensus, actuarial, other); the process of administering the tool (who by and length of time); the structure of the tool (number of questions, number of scales); and the content (factors and number of items for each scale)

[3] In general, significant Kappa scores of 0.3 to 0.49 indicate fair correspondence between raters, 0.5 to 0.69 indicate moderate correspondence, and 0.7 and above indicate high correspondence. The more conservative threshold of 0.8 will be used for this review.

3) Population and Setting: data will be collected on key demographic variables (age, sex, family structure, SES, race/ethnicity, urbanicity, etc); population (child welfare agency caseload, treatment agency sample, other); sample size; setting; sample recruitment procedures (via records, letter to all subjects asking to take part); duration of the study; and length of follow-up
4) Outcomes: Confirmation that the tool is being used to predict maltreatment recurrence; source of the reported outcome, and whether maltreatment is indicated/confirmed/substantiated.
5) Results related to Risk Assessment Tool: Results related to the risk assessment tool will be evaluated based whether there is sufficient data to calculate sensitivity, specificity, positive predictive value, negative predictive value, ROC, and area under the curve.

Software
All studies and data extraction sheets will be stored electronically in TrialStat, which is a web-based data extraction software that allows multiple reviewers to code studies while keeping a web-based log of all activities during the review process. Endnote will also be used as the bibliographic management system to store, locate, and track references.

Statistical procedures and conventions.

The effect size of interest for this systematic review is the accuracy of the screening instrument. Two important indices that represent the accuracy of a test are the sensitivity and the specificity, which are defined on p.8. These two indices are calculated in the form of percentages, which means they are "scale-free" and are comparable across studies. Therefore, the metric of the outcome measures in most diagnostic test studies is usually not an issue (see flowchart for data analysis plan). However, the challenge of this kind of study is the availability of information necessary for obtaining the sensitivity and the specificity of a given instrument.

We will also document other common metrics used to evaluate the predictive validity of risk assessment instruments (positive predictive value, negative predictive value, and ROC/AUC).

For analyses of overall performance of risk assessment instruments to predict recurrence of maltreatment, a comparison of common metrics of risk assessment tools will be conducted. Subsequently, the review will synthesize the data by using meta-analytic techniques to determine the overall predictive capacity of each risk assessment instrument. Advances in the synthesis of studies of the accuracy of diagnostic and prognostic instruments discussed in the Cochrane and Campbell Collaborations will be incorporated into this review.

Flowchart of Analyses:

Extract the accuracy indices from each study.
-Sensitivity (True Positive Rate: TPR);
-Specificity (True Negative Rate:TNR);
-Any frequency count and ratio that can give us TPR and TNR;
-ROC curve and/or AUC

↓

Obtain the estimation of a summary ROC (SROC), which takes the linear model format: D=α+ βS.
The intercept α is an odds ratio (as defined on the right), which can be used as a summary index of the accuracy; the regression coefficient β shows the extent to which the ORs from studies are homogeneous.

Note: Both D and S can be calculated based on the effect sizes extracted at the previous step:
D=logit(TPR)-logit(1-TNR)=log(OR);*
S= logit(TPR)+logit(1-TNR).
**The odds ratio (OR) is defined as:*

$$OR=\frac{\frac{TPR}{1\text{-}TPR}}{\frac{1-TNR}{TNR}}=\frac{TPR}{1\text{-}TPR}*\frac{TNR}{1-TNR}$$

β is close to zero ↓

When β is close to zero, the accuracy of risk assessment instruments across studies is homogeneous. Different point estimates of the accuracy can be obtained, such as the OR (α) and the averaged sensitivity and specificity (as well as their confidence intervals).

β is not close to zero ↓

When β is not close to zero, the accuracy across studies is heterogeneous. Subgroup analyses using moderators (study characteristics) can be performed.

More complicated methods will be adopted to take into account the variability of the accuracy between studies.

One possible method would be the hierarchical SROC (HSROC) model proposed by Rutter and Gatsonis (2001) to quantify variation within and between studies.

Since any diagnostic or prognostic measure will vary in sensitivity and specificity, depending on where one sets the threshold, the review requires the use ROC curves (or at least plotting sensitivity and 1-specificity in ROC space) to assess the predictive accuracy of the measure. If studies report sensitivity and specificity for a particular threshold (one point on the ROC curve), the first step in the synthesis of results of multiple studies can be to plot these values in ROC space.

Next, following Moses, Shapiro, and Littenberg (1993), summary ROC (SROC) curves can be generated from the relationship between predictive accuracy (= logit(true positive rate) – logit(false positive rate)) and a proxy threshold (logit(true positive rate) + logit (false positive rate)). The relationship between predictive accuracy and proxy threshold is expressed in the regression equation: accuracy = constant + coefficient(proxy threshold). A fixed effect model is fitted using linear regression. The model can be weighted using inverse variance methods.

SROCs can be generated for each risk assessment method and these curves can be compared in terms of their accuracy and shape (asymmetry). One assessment method may be more accurate than another at a certain point (sensitivity and 1-specificity) while another can be more accurate at another point.

One problem with this approach is its lack of consideration of heterogeneity of results across studies (or heteroskedasticity in the regression) using the same instrument. Heterogeneity of predictive accuracy can be due to client characteristics (e.g., different base rates of abuse/neglect in different samples).

While SROC regression is currently the most commonly used method, it does not account for within and between study variability, and confidence intervals and p-values may be incorrect (Macaskill, 2006).

Multilevel (hierarchical) models are needed to take within-study and between-study variability into account. Two hierarchical SROC models have been proposed and they produce statistically equivalent results. The HSROC model includes random effects for both predictive accuracy and threshold. The focus is on estimating the SROC. The area under the curve (expected sensitivity for a given specificity, etc.) can be derived from this model. The HSROC model has 5 parameters: mean & variance of both threshold and accuracy (=4) + shape of the curve (scale, asymmetry). The first level models within-study variation, the second level models between-study variation. Rutter & Gatsonis (2001) used Bayesian methods to fit the model, which requires a third level (prior probabilities/distributions). However, this procedure is difficult to do and is rarely used. The 2-level HSROC can be fitted in SAS using PROC NLMIXED (not the linear version) or STATA (though we understand that the Cochrane Diagnostic Test Accuracy group is writing code to share with reviewers for both SAS and STATA).

The second approach uses a bivariate hierarchical model of the relationship between sensitivity and specificity (after logit transformation), including random effects for both. This method can then be used to generate an underlying SROC curve this 2-level model can be fitted using NLMIXED in SAS or gllamm in STATA.

The methods described above use a single point (sensitivity + 1-specificity) from each study. When studies provide an entire ROC curve, an attempt will be made to use all of this information.

We will explore heterogeneity of results across multiple studies of the same instrument and compare results across risk assessment instruments. Specifically, threats to internal validity that are unique to studies of prognostic accuracy (e.g., Bossuyt et al., 2003a, 2003b) will be considered. First, we will explore the degree to which spectrum (refers to the distribution of the target "recurrence of child maltreatment" construct in the participants who do and do not exhibit a recurrence of maltreatment) varies across studies by analyzing between-studies heterogeneity and its influences the generalizability of results to particular populations;. We will also consider "verification bias", which is concerned with the possibility that the reference standard of maltreatment recurrence is applied differentially according to the outcome of the initial diagnosis (e..g., if participants deemed at higher risk of recurrence are monitored more vigilantly or longer than lower risk participants. In addition, we will explore error in the reference standard, which can distort the evidence against which initial diagnosis is compared and subsequently yield biased estimates of prognostic accuracy. This latter issue will be explored by stratifying differential attrition and length of follow-up.

Independence of Results

Based on our preliminary review of the literature, there will be studies that will report results for two (or more) different instruments used to assess the same participants; such a study by itself provides a strong basis for comparing the two instruments since between-studies confounds are eliminated, but including both instruments in a statistical analysis would violate the assumption of an independent sample for each instrument.

When appropriate (i.e., key assumptions are met), we will use "type of instrument" as a moderator variable to perform subgroup analyses. Specifically, we can group the effect sizes based on the instrument, and calculate and plot the SROC curve for each type of instrument. The curves will then be used to compare the validity of the instruments. When the "type of instrument" is used as a moderator to separate the effect sizes, it reduces the bias arising from the use of multiple instruments tested in a single study. Methods used for analyzing the data will depend on the availability of the threshold values in the included studies. For studies evaluating instruments with multiple thresholds (cutoff points), the effect size with either the most commonly used threshold will be employed or thresholds will be entered as covariates to form an SROC curve.

Sensitivity Analysis

Another likely source of dependence arises if the same participants contribute more than one sensitivity-specificity pair, such as when multiple points are read from a reported ROC curve, or when participants' risk levels are classified into more than two graded categories. We will be aware of these concerns regarding dependencies and we will conduct sensitivity analyses by first performing separate analysis and only combining studies without dependency concerns.

Publication Bias

Irwig, Macaskill, Glasziou and Fahey (1995) suggest that publication bias may have an even more serious influence on studies of the diagnostic accuracy of an assessment.
Since the determinants of publication bias are likely to be different for investigations of prognostic accuracy (Deeks, Macaskill, Irwig, 2005), we will undertake funnel plot investigations to examine the possibility of publication and other sample size related effects using plots of InDOR against $1/ESS^{1/2}$, and test for asymmetry using related regression or rank correlation tests as suggested by Deeks and colleagues.

Treatment of qualitative research.
Qualitative research will not be included in this systematic review.

5. Timeframe

Activity	**Expected Completion**
Development of protocol and specific review questions	February 2007
Searches for published and unpublished studies	July 2007
Pilot testing of inclusion criteria	August 2007
Pilot testing of study codes and data collection	August 2007
Extraction of data from research reports	December 2007
Statistical Analysis	June 2008
Preparation of initial report	September 2008

6. Plans for Updating the Review

The review will be updated every two years.

7. Acknowledgements

Jessie Ball Dupont, Bell Canada, Royal Bank, and the Campbell reviewers.

8. Statement Concerning Conflict of Interest

None known.

9. References

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine, 19(4),* 453– 473.

Baird, S. 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, Occasional Monograph Series No. 2*. 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(11/12),* 839-871.

Baird, S. C., Wagner, D., Healy, T., & Johnson, K. (1999). Risk Assessment in Child Protective Services: Consensus and Actuarial Model Reliability. *Child Welfare, 78,* 723 - 748.

Trocmé, N., Barber, J., Goodman, D., Shlonsky, A., & Black, T. (2007). The reliability and predictive validity of a consensus-based risk assessment. University of Toronto.

Barlow, J., Johnston, I., Kendrick, D., Polnay, L. & Stewart-Brown, S. (2006). Individual and group-based parenting programmes for the treatment of physical child abuse and neglect. *Cochrane Database of Systematic Reviews* 2006, Issue 3.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Moher, D., Rennie, D., de Vet, H. C. W., & Lijmer, J. G. (2003a). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry, 49,* 7-18.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, CA., Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D., & de Vet, H. C. W. (2003b). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326,* 41-44.

Browne, K., & Saqi, S. (1988). Approaches to screening for child abuse and neglect. In, K. Browne, P. Stratton, & C. Davies (Eds.), *Early Prediction and Prevention of Child Abuse* (pp. 57-85). Chichester: John Wiley & Sons.

Cash, S. J. (2001). Risk assessment in child welfare: The art and science. *Children and Youth Services Review, 23(11),* 81 l-830.

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,* 174-1 83.

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

Child Welfare League of America (2005). A comparison of approaches to risk assessment in child protection and brief summary of issues identified from research on assessment in related fields. Retrieved May 5, 2006 online at http://www.pacwcbt.pitt.edu/Organizational%20Effectiveness/Practice%20Reviews/RevisedRAArticleCWLA11-05.DOC

Cooper, E. (1997). Identifying those at risk for physically abusing children: Literature review. BC Institute on Family Violence. Retrieved online on August 11, 2006 at http://www.bcifv.org/pubs/Identifying%20Those%20At%20Risk.pdf.

Dalgleish, L. I., & Drew, E. C. (1989). The relationship of child abuse indicators to the assessment of perceived risk and to the court's decision to separate. *Child Abuse & Neglect*, *13,* 491-506.

Dawes, R. M. (1993). Findings guidelines for tough decisions. *Chronicle of Higher Education, 39*.

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.

Deeks, J. J., Macaskill, P. & Irwig, L. (2005). The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy. *Journal of Clinical Epidemiology, 58(9),* 882-93.

DePanfilis, D., & Zuravin, S.J. (1998). Rates, patterns, and frequency of child maltreatment recurrences among families known to CPS. *Child Maltreatment, 3,* 27-42.

Doueck, H., English, D., DePanfilis D., & Moote, G. (1993). Decision making in child protective services: A comparison of selected risk assessment systems. *Child Welfare*, *72,* 441-452.

Doueck, J., Levine, M., & Bronson, D. (1993). Risk assessment in child protective services: An evaluation of the Child at Risk Field System. *Journal of Interpersonal Violence, [volume number?]*, 446-467.

English, D. J., Aubin, S. W., Fine, D., & Pecora, P. J. (1993). Improving the accuracy and cultural sensitivity of risk assessment in child abuse and neglect cases. Seattle, WA: University of Washington, School of Social Work.

English, D. J. & Pecora, P. J. (1994). Risk assessment as a practice method in child protective services. *Child Welfare, 73(5),* 451-473.

Everitt, B. S. (1998) *The Cambridge dictionary of statistics*. United Kingdom: Cambridge University Press.

Gambrill, E., & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth Services Review, 22(5–6),* 813– 837.

Gambrill, E., & Shlonsky, A. (2001). The need for a comprehensive risk management system in child welfare. *Children and Youth Services Review, 23(1),* 79– 107.

Giovannoni, J. (1989). Definitional issues in child maltreatment. In D. Cicchetti & V. Carlson (eds.), *Child maltreatment: Theory and research on the causes and consequences of child abuse and neglect* (pp. 3-37). New York: Cambridge University Press.

Gottfredson, S.D. & Moriarty, L.J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency, 52(1),* 178-200.

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

Hindley, P. G., Ramchandani, D. P., & Jones, H. (2006) Risk factors for recurrence of maltreatment: A systematic review. Archives of Diseases in Childhood Press Releases. Retrieved online on August 10, 2006 at http://press.psprings.co.uk/adc/june/ac85639.pdf

Irwig, L., Macaskill, P., Glasziou, P. & Fahey, M. (1995). Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology,48,* 119-130.

Johnson, W. (2004). Effectiveness of California's child welfare structured decision-making (SDM) model: A prospective study of the validity of the California Family Risk Assessment. Oakland, CA: Alameda County Social Services Agency.

Johnson, W., & L'Esperance, J. (1984). Predicting the recurrence of child abuse. *Social Work Research and Abstracts, 20(2),* 21–26.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston: Allyn and Bacon

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, 143-155.*

McDonald, T. & Marks, J. (1991). A review of risk factors assessed in child protective services. *Social Service Review, 65,* 112-132.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Melton, G. B. & Flood, M. F. (1994). Research policy and child maltreatment: Developing the scientific foundation for effective protection of children. *Child Abuse & Neglect, 18,* l-28.

Moses, L. E., Shapiro, D., and Littenberg, B. (1993). Combining independent studies of diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine,* 12, 1293–1316.

Muir, R. C., Monaghan, S. M., & Gilmore, R. J. (1989) Predicting Child Abuse and Neglect in New Zealand. *Australian and New Zealand Journal of Psychiatry*, *23,* 255-260.

Munro, E. (2004). A simpler way to understand the results of risk assessment instruments. *Children and Youth Services Review, 26,* 873-883.

Murphy-Berman, V. (1994). A conceptual framework for thinking about risk assessment and case management in child protective services. *Child Abuse & Neglect*, *18,* 193-201.

Pecora, P. (1991). Investigating allegations of child maltreatment: The strengths and limitations of current risk assessment systems. *Child and Youth Services Review, 15,* 73-92.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*, 737-748.

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

Reid, G., Sigurdson, E., Wright, A., & Christianson-Wood, J. (1996). Risk assessment: Some Canadian findings. *Protecting Children*, *12(2):* 24-31.

Rutter, C. M. & Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20, 2865-2884.

Rycus, J. S., & Hughes, R. C. (2003). *Issues in risk assessment in child protective services: Policy white paper.* Retrieved May 5, 2006, from the North American Resource Centre for Child Welfare Web site, http://www.narccw.com

Shlonsky, A., & Wagner, D. (2005). The next step: Integrating actuarial risk assessment and clinical judgment into evidence-based practice framework in CPS case management. *Children and Youth Services Review, 27,* 409-427.

Shlonsky, A. (2007). Initial construction of an actuarial risk assessment measure using the National Survey of Child and Adolescent Well-Being (NSCAW). In Wulczyn, F. & Haskins, R. (Eds). *Child Protection: Using Research to Improve Policy and Practice.* Washington, DC: Brookings Institution.

Sim, J. & Wright, C. C. (2005) "The Kappa Statistic in Reliability Studies: Use,

Interpretation, and Sample Size Requirements" in *Physical Therapy*. 85, 257-268

Stone, M. (1993). *Child Protection: A Model for Risk Assessment in Physical Abuse/Neglect*. Surrey County Council, Surrey.

Tatara, T. (1996). *A survey of states on CPS risk assessment practice: Preliminary findings.* Paper presented at the 10th Annual National Roundtable on CPS Risk Assessment, San Francisco, CA.

Wagner, D. (1997). Risk Assessment Validation Study. South Australia Department of Family and Community Services. Retrieved online on August 10, 2006 at http://www.nccd-crc.org/crc/pubs/so_aus_1997_risk_val.pdf

Wald, M. S., & Woolverton, M. (1990). Risk assessment: The emperor's new clothes*? Child Welfare, 69,* 483-5

Wiggins, J. S. (1981). Clinical and statistical prediction: Where are we and where do we go from here? *Clinical Psychology Review, 1,* 3-18.

Tables and Figures

Appendix 1 -Systematic Information Retrieval Coding Sheet (SIRC)

Project:

_____

Reviewer:

_____

Date(s) of Search:

_____

Search Method: Electronic Database: Name: _____
Grey Literature: Name: _____
Other Name: _____

Language(s):

_____

Date Range Searched within Database:

_____

Description of Search:

_____
_____
_____

Search Term Combinations (including all limiters and expanders)	**Results**

Appendix 2 - Critical Appraisal Form

Administrative

1. Author(s): _____

2. Year of Publication: _____

4. Source: _____

5. Country:_____

6. Language: _____

7. Record Number: _____

Description of Risk Assessment Tool:

8. Risk Assessment Tool (Name and Acronym)

a) Risk assessment instruments
- Washington State Risk Assessment Matrix
- Child Endangerment Risk Assessment Protocol
- Child Well-Being Scales
- The Ontario Child Neglect Index
- Child at Risk Field System
- Family Risk Scales
- California Family Risk Assessment
- Michigan Family Risk Assessment
- Other (please specify):_____

b) Measures that focus on the environment of the child
- Home Accident Prevention Inventory
- Checklist for Living Environments to Assess Neglect
- Home Safety and Beautification Tour
- Other (please specify):_____

c) Parent self-report measures
- Child Abuse Potential Inventory
- Conflict Tactics Scale, Parent to Child version
- Other (please specify):_____

d) Observation measures
- Child Abuse and Neglect Interview Schedule
- Childhood Level of Living Scale
- Other (please specify):_____

6. Overall aim/purpose of

_____
_____
_____
_____

7. Focus of Risk Assessment

Assess risk of recurrence of maltreatment

8. Method of Risk Assessment

Clinical
Consensus
Actuarial

9. Person administering instrument

Protection worker
Other (please specify):_____

10. Content: (label names of the questions/factors and number of items for each scale)

_____
_____
_____
_____

Population

11: Number of Participants: _____

12. Age: a. Range:_____
b. Mean: _____
c. SD: _____

13. Gender: a. Male
b. Female

14. Sample recruitment procedures: Via records
Letter to all subjects asking to take part
Other (please specify):_____

15. Types of Recurring Maltreatment: Physical abuse
Sexual abuse
Neglect
Emotional abuse (including witnessing domestic violence)

16. Unit of Analysis

- □ recurrence of maltreatment towards the child who was initially maltreated
- □ recurrence of maltreatment towards a subsequent child in the same household

Setting and Location

17. Location: _____

18. Setting: _____

Length and Retention

19. Duration (in weeks)

20. Follow-up (in weeks)

21. Number of participants not included in the analysis:_____

22. Attrition rate:

Outcome Measures

19. Outcome Measures:

subsequent investigation of maltreatment within 18 months of the investigation that brought each case into the study cohort

substantiation resulting from these investigations over the 18-month follow-up period

Quality of Study

20. STARD checklist for the reporting of studies of diagnostic accuracy[4]

Manuscript number and/or corresponding author name:			
Section and Topic	**Item #**		**On page #**
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS		Describe	
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations	

[4] Reprinted from Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al, for the STARD Group. Towards complete and accurate
reporting of studies of diagnostic accuracy: the STARD initiative. Ann Intern Med 2003;138:40–4.

		where the data were collected.	
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.	
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	
RESULTS		Report	
Participants	14	When study was done, including beginning and ending dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	
	22	How indeterminate results, missing responses and outliers of the index tests were handled.	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, typically imprecision (as CV) at 2 or 3 concentrations.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

Analysis

21. Results related to Risk Assessment Tool

Effectiveness of the Tool Extract the accuracy indices from each study including TPR, TNR, Frequency count and Ratio (see A, B, C, D, below)	**Calculations and Results**
A. Sensitivity (True Positive Rate:TPR);	
B. Specificity (True Negative Rate:TNR);	
C. Frequency count	
D. Ratio	
The values of threshold (cutoff score) that the study used to decide the outcome (positive or negative).	
Estimation of a summary ROC (SROC), which takes the linear model format: D=α+ βS. *Note: Both D and S can be calculated based on the effect sizes extracted at the previous step:* *D=logit(TPR)-logit(1-TNR)=log(OR*);* *S= logit(TPR)+logit(1-TNR).* **The odds ratio (OR) is defined as:* $OR=\frac{\frac{TPR}{1-TPR}}{\frac{1-TNR}{TNR}}=\frac{TPR}{1-TPR}*\frac{TNR}{1-TNR}$	
Positive predictive value:	
Negative predictive value:	

Area under the curve:	
Other measures of classifying accuracy.	

22. Comments:

_____

_____

_____

_____

_____

_____