Risk and Safety Assessment in Child Welfare: Instrument Comparisons

Amy D'Andrade, PhD
Michael J. Austin, PhD
Amy Benton, MSW

SUMMARY. The assessment of risk is a critical part of child welfare agency practice. This review of the research literature on different instruments for assessing risk and safety in child welfare focuses on instrument reliability, validity, outcomes, and use with children and families of color. The findings suggest that the current actuarial instruments have stronger predictive validity than consensus-based instruments. This review was limited by the variability in definitions and measures across studies, the relatively small number of studies examining risk assessment instruments, and the lack of studies on case decision points other than the initial investigation. doi:10.1300/J394v05n01_03 *[Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <http://www.HaworthPress.com> © 2008 by The Haworth Press. All rights reserved.]*

KEYWORDS. Risk assessment, safety assessment, child welfare, actuarial instruments, consensus-based instruments

INTRODUCTION

Before child welfare agencies intervene with families, they are generally required to identify maltreatment or the risk of maltreatment. As a

Amy D'Andrade is Research Director, Michael J. Austin is Professor, and Amy Benton is Doctoral Research Assistant; all are affiliated with Bay Area Social Services Consortium, School of Social Welfare, University of California, Berkeley.

[Haworth co-indexing entry note]: "Risk and Safety Assessment in Child Welfare: Instrument Comparisons." D'Andrade, Amy, Michael J. Austin, and Amy Benton. Co-published simultaneously in the *Journal of Evidence-Based Social Work* (The Haworth Press) Vol. 5, No. 1/2, 2008, pp. 31-56; and: *Evidence for Child Welfare Practice* (ed: Michael J. Austin) The Haworth Press, 2008, pp. 31-56. Single or multiple copies of this article are available for a fee from The Haworth Document Delivery Service [1-800-HAWORTH, 9:00 a.m. - 5:00 p.m. (EST). E-mail address: docdelivery@haworthpress.com].

Available online at http://jebsw.haworthpress.com

doi:10.1300/J394v05n01_03

result, the assessment of risk is a critical part of child welfare agency practice (Wald & Woolverton, 1990). Most states in the US formalize the process of assessing risk by using some type of structured decision-making process or tool. Risk assessment instruments generally include broad categories of areas related to abuse and neglect, behavioral descriptions, procedures to determine levels of risk, and standardized forms to record this information (Rycus & Hughes, 2003). After describing current approaches to risk and safety assessment and related issues, this review of the research literature describes instruments for risk assessment in terms of their reliability, predictive validity, outcomes, and use with children and families of color. The review concludes with implications for practice and research.

THE PURPOSE OF RISK ASSESSMENT IN CHILD WELFARE

One goal of risk assessment is to focus limited resources on the children who are at the greatest risk of maltreatment. Given the limited resources of social services agencies, risk assessment serves as a strategy for targeting scarce resources and services toward those who have the greatest need (Camasso & Jagannathan, 2000; Leschied, Chiodo, Whitehead, Hurley & Marshall, 2003; Rycus & Hughes, 2003; Wald & Woolverton, 1990). A second purpose of a structured risk assessment process is to facilitate an accurate, less biased decision-making process for determining which cases should receive services (Rycus & Hughes, 2003).

Researchers from a variety of academic fields, particularly psychology, have studied human decision-making and identified a number of common errors people tend to make in their predictions and decisions. For example, researchers have found that people tend to disregard information regarding the base-rate of a phenomenon in a population (such as child abuse or neglect) when attempting to predict or diagnose the presence of the phenomenon in a specific situation (such as a child abuse investigation) (Kahneman & Tversky, 1982). In addition, people tend to be overconfident of their ability to predict an event (Kahneman & Tversky, 1982), and to have difficulty weighing factors related to a decision (Grove & Meehl, 1996). Part of the difficulty for decision-makers may relate to the availability of too much information, some of which is likely unrelated to the outcomes and may thereby increase the chance that irrelevant information is used in the decision (Dawes, 1994).

Several studies focus on decision-making within the field of the child welfare. In a study by Schuerman, Rossi & Budde (1999), case vignettes were presented to social workers, expert practitioners and academics in the field. In considering the same vignette, workers made different choices from one another in deciding whether or not a child should be removed from the home. While this variability was less when cases were of very low or very high risk, "...similar cases in the midrange of severity of family problems are treated quite differently by different experts and workers" (Schuerman et al., 1999, p.609). In a second study, researchers conducted a content analysis of 45 "public inquiries" completed in England on two types of cases: cases of child deaths by parental actions, and sexual abuse cases in which the actions of professionals were suspected of being overzealous. The formal inquiries identified three general types of errors made by child welfare workers. First, "the most striking and persistent criticism was that professionals were slow to revise their judgments . . . (adherence to) the current risk assessment of a family had a major influence on responses to new evidence" (Munro, 1999, p. 748). Secondly, professionals were skeptical of new information when it conflicted with their initial view of a family, yet uncritical of new information when it supported their initial view. And third, evidence from some sources was more highly valued (e.g., doctor's statements regarding abuse and social worker's witnessing of injury) than other sources (neighbors, concerns of the public) (Munro, 1999). These findings suggest that child welfare workers are prone to the same difficulties in decision-making as those in other fields and that valid risk assessment processes are needed to aid social workers in decision-making.

DIFFERENT APPROACHES TO RISK ASSESSMENT

Currently, there are two major approaches to risk assessment in child welfare decision-making: a consensus-based model, and an actuarial model. Both involve a list of family or case characteristics believed to be associated with risk of maltreatment (a risk assessment "instrument"). However, the two approaches differ in the processes used to identify factors for inclusion in the instrument and how the instruments are utilized in practice.

Consensus-based instruments emphasize a comprehensive assessment of risk based upon various theories of child maltreatment, the research literature on maltreatment, and/or the opinions of with expert practitioners (English, 1999). Items on one instrument are often combined with items from another instrument, creating hybrid instruments that vary according to the needs or beliefs of the user. Sometimes factors are assessed numerically and families categorized by their total score, while other instruments simply describe areas that are to be assessed by the worker without necessarily providing direction in terms of that assessment. Either way, the worker considers the area identified and codes it high, moderate or low risk based upon his or her judgment. Consensus-based instruments tend to use the same instrument to predict all forms of maltreatment (English, 1999).

In contrast, actuarial instruments are developed using statistical procedures that identify and weigh factors that predict future maltreatment (Rycus & Hughes, 2003). Often the statistical analysis is done in the state or county in which the instrument is applied. Factors identified as predictive of maltreatment are incorporated into a checklist. Social workers score each factor, scores are summed into overall risk scores for each type of maltreatment, and families are categorized into low, moderate or high-risk groups. Actuarial instruments tend to use fewer factors than do consensus-based instruments and generally use different factors to predict the likelihood of physical abuse and of neglect.

There is some debate regarding the best approach to risk assessment in the field of child welfare. A consensus-based approach to risk assessment utilizes the underlying theoretical assumption that the causes of child maltreatment are multi-dimensional and complex, and therefore utilizes many related domains (English & Graham, 2000). These instruments can also help structure a worker's process of information gathering for clinical assessments of risk (English, 1999), and provide documentation of the reasoning underlying their decision-making (Doueck, English, DePanfalis & Moote, 1993; English, 1999). Some argue that the more comprehensive approach of consensus-based instruments provides better information for casework decisions (Nasuti & Pecora, 1993).

However, consensus-based models are criticized in the research literature for the following reasons: (1) poor conceptualization (according to Rycus and Hughes, "measures are often poorly defined, nebulous and ambiguous, overly global, illogical, and very subjective" (2003, p. 13)); (2) inconsistency in the type and number of variables included (Rycus & Hughes, 2003); (3) use of the same variables to predict physical abuse, neglect, and sexual abuse, even though contributing factors are often different for these types of maltreatment (Rycus & Hughes, 2003); and (4) reliance upon characteristics associated with maltreatment (rather than recurrence of maltreatment) (Wald & Woolverton, 1990) or upon characteristics for which there is no research support (McDonald & Marks, 1991).

Actuarial instruments help practitioners focus their risk assessments on a small set of case characteristics that have demonstrated a strong statistical relationship to future maltreatment (Ereth, Johnson & Wagner, 2003). In a meta-analysis of over 100 studies of various health and behaviorally oriented predictions that compared actuarial methods to clinical judgment, about 95% of the studies found actuarial processes to be equal to or superior to clinical judgment (Grove & Meehl, 1996). According to proponents, even a small set of factors generally does a better job predicting outcomes than does simply the use of clinical judgment (Dawes, 1979).

Critics of actuarial instruments assert that these instruments do not facilitate the clinical judgment of skilled practitioners. In addition, since the basis for including a factor on an actuarial instrument is its statistical association with a poor outcome (generally recurrence of maltreatment), factors may not appear to be causally related to the outcome. This perceived lack of a logical, theoretical connection between the items may lead to discounting the value of an actuarial instrument and objections to its use (Schwalbe, 2004).

It is not clear which approach to assessing risk is more commonly used. There is no national database regarding risk assessment approaches used by states (Lyons, Doueck & Wodarski, 1996). As of 1996, Lyons et al. (1996) reported that 15 states used the Illinois CANTS 17B instrument or some derivation of it, 4 used the CARF system, and 4 used WARM or some derivation, all consensus-based instruments. However, an increasing number of states are using an actuarial instrument as part of a "Structured Decision Making (SDM)" case management system developed by the Children's Research Center (CRC, n.d.).

USE ACROSS THE LIFE OF A CASE

Since a family's risk may change over time, it is important that risk be periodically reassessed (Munro, 2004). For example, few states have explicit guidelines for making screening decisions, and fewer still have formal instruments to guide this decision (Downing, Wells, & Fluke, 1990). The percentage of referrals that are screened out varies dramatically by state, from 5% in New Jersey to 78% in Vermont, yet states with higher investigation rates were just as likely to substantiate a referral as states with lower investigation rates (Tumlin & Geen, 2000). In addition, caseworkers have difficulty revising their assessments of families once they have been made (Munro, 1999). Therefore, a structured instrument could help workers attend to critical factors that would indicate changes in risk across the life of a case. However, using the same instrument to assess risk at different case points may be unwise (Wald & Woolverton, 1990). Factors that predict maltreatment at one point, such as at investigation and prior to services, may not be the same as those that predict subsequent maltreatment at another time point, such as at reunification after service provision. For example, one study examining two time points (within 24 hours of the initial investigation of a case and within 5 days of a case opening for in-home services) found that factors that predicted maltreatment recurrence at each time point were not always the same (Fuller, Wells & Cotton, 2001).

Unfortunately, there is very little research regarding the reliability or validity of instruments used at case time points other than the initial investigation (Rycus & Hughes, 2003; Zuravin, Orme & Hegar, 1995). For some relevant outcomes, validity is difficult to assess. For example, the safety assessment focuses on assessing imminent or current risk to a child and is usually concerned with maltreatment "...of a moderate to severe nature" (Fuller et al., 2001). While the likelihood of subsequent maltreatment within sixty days is fairly rare, the likelihood of subsequent maltreatment occurring within several days is even more so. When events are rare, it is very difficult to accurately predict them (Johnson, 2004; Munro & Rumgay, 2000).

IMPORTANT QUALITIES OF A RISK ASSESSMENT INSTRUMENT

When considering the value of any risk assessment approach, two psychometric qualities are of particular importance; namely, *predictive validity* which refers to the accuracy of the instrument in predicting a particular outcome and *inter-rater reliability* which involves the degree to which the use of an instrument leads to consistent worker decisions for similar cases (Rycus & Hughes, 2003). In child welfare, most studies focus on whether or not risk assessment instruments accurately predict the occurrence of subsequent maltreatment.

Most diagnostic or predictive instruments will produce some errors; that is, using the instrument, some cases will be identified as 'high risk' even though they are truly low risk (resulting in false positives) and some cases will be identified as low risk although they are truly high risk (resulting in false negatives). In child welfare, these classification errors are important because false negatives can be dangerous to the child and false positives can result in poor targeting of agency resources (Camasso & Jagannathan, 2000).

Some researchers assess the validity of an instrument in terms of the degree to which false negatives or false positives are minimized (Lyons et al., 1996). However, when considering a rare phenomenon like subsequent maltreatment, a low false negative and false positive rate can be achieved by simply predicting the event never happens (Baird, Ereth & Wagner, 1999; Dawes, 1994). Such a strategy, of course, is not useful for protecting children and/or providing services to families in need. An alternative strategy for assessing the validity of a prognostic instrument is to classify individuals into risk categories so that an individual can be categorized in terms of a low, moderate, or high risk of some adverse event. In medicine, this framework provides important information for a patient and doctor to make decisions about the most appropriate preventive approach to take (Altman & Royston, 2000; Baird & Wagner, 2000). Proponents of this approach to classification assert that it provides important information for case decisions and should be the basis for determining the validity of a risk assessment instrument, even if it produces more false results than does predicting that the event never happens (Baird, Ereth & Wagner, 1999; Rycus & Hughes, 2003).

Some have argued other forms of validity are also important to consider in risk assessment instruments. English & Graham (2000) assert that convergent validity, which involves the degree to which a measure corresponds to other measures of the same or similar constructs, is also relevant. In addition to considering the validity of the instruments overall, it is also important to consider the validity of the measures and/or outcomes assessed by an instrument. For example, there is some question regarding the appropriate indicator to use for maltreatment recurrence. If *substantiated maltreatment* is used, it could underestimate the future occurrence of maltreatment, especially if it is not detected (English & Graham, 2000). If *subsequent referral* is used, it could overestimate occurrence because some referrals will be unfounded. If substantiation decisions are themselves biased, using them as the basis for assessing the validity of a risk assessment instrument would be inappropriate (Morton, 1999). Some researchers believe that the severity of the maltreatment should be incorporated into the criterion (Morton & Salovitz, n.d.). Another critique of both consensus-based and actuarial risk assessment instruments is that they focus almost exclusively on the interpersonal factors of parents and rarely upon neighborhood, community or societal factors that may be associated with maltreatment (Galasso, 2001).

Inter-rater reliability refers to the degree to which an instrument results in similar decisions on similar cases when those cases are assessed by different workers (Rycus & Hughes, 2003; Schuerman et al., 1999). Assessing reliability also involves challenges. The goal is to determine whether multiple users of the same instrument would reach the same decision for the same situation. However, the practice environment of one person cannot be replicated for another person in order to determine if he or she would make the same decision. Two alternative strategies have been used to assess inter-rater reliability: (1) different workers read the same case file and their assessments or predictions of risk are compared, and (2) different workers read a hypothetical case vignette and their assessments or predictions of risk are compared. Both of these strategies are problematic for different reasons. A vignette is consistent, but artificial and limited in terms of the information that is provided to the decision-maker. Case files are complex and reflect the "real world" but could be missing critical information that was present in the real world situation (Baird, Ereth & Wagner, 1999). In addition, the statistical estimate used to assess reliability can be problematic; there could be a high correlation between two scores, even if the two coders have consistently different scores (Baird, Ereth & Wagner, 1999).

Child welfare agencies have an expectation that "risk assessment will have some effect on services provided" (Fluke, Wells, England, Walsh, English, Johnson, Gamble & Woods, 1993, p. 118). Therefore, another way to evaluate risk assessment strategies is to consider whether they have improved safety and risk decisions, facilitated services for high risk children, and thus improved outcomes (assuming services are effective). For example, one might expect to see fewer recurrences of maltreatment after implementation of a risk assessment instrument, as high risk cases would be targeted for services. Similarly, if workers were doing a better job assessing risk and safety in reunification decisions, one would expect to see a reduction in foster re-entry rates and/or maltreatment after reunification. In addition, some studies assess the use of various instruments with children and families of color because it is important to determine whether a particular instrument is equally valid for different racial/ethnic groups (Baird & Wagner, 2000; Baird, Ereth & Wagner, 1999; English, Marshall, Brummell & Orme, 1995; Johnson, 2004, 2005; Loman & Siegel, 2004).

Comparing the performance of various instruments is difficult because not all instruments have been assessed for their reliability, validity, or effects. Comparisons are further complicated by the variety of events used by different researchers to measure maltreatment recurrence. For example, many researchers use substantiated maltreatment (subsequent to the initial investigation) as the criterion against which to measure predictive validity. Others use subsequent referral, or a measure of the chronicity of subsequent referrals or substantiation, or placement of the child outside the home. Furthermore, the observation time frame varies (30 days, 60 days, 6 months, or 12 months or longer). Finally, different strategies are employed to assess the both reliability and validity and different kinds of statistics are used to estimate them.

Review of Studies

This review considered studies examining risk assessment instruments for reliability and validity, as well as studies examining the effects of the implementation of a risk assessment system on child and case outcomes and the effects for families of various different racial/ethnic groups. Specific search terms were used to search the social science and academic databases available through the University of California library, Websites specializing in systematic reviews, and publications of research institutes. Studies were excluded if they did not assess a particular identified instrument of risk assessment appropriate for use in a CPS investigation situation, or in the case of predictive validity studies, if the outcome assessed did not relate to maltreatment recurrence or case re-referral.

The search process identified studies examining the following seven instruments of risk and safety assessment: (1) the Washington Risk Assessment Matrix (WRAM), (2) the California Family Assessment Factor Analysis (CFAFA, the "Fresno" instrument), (3) the Child At Risk Field System (CARF), (4) the Child Emergency Response Assessment Protocol (CERAP), (5) the actuarial Risk Assessment instruments developed by the Children's Research Center, (6) the Risk Assessment Model of Child Protection from Ontario, and (7) the Utah Risk Assessment Scale. Findings from available studies related to predictive validity, convergent validity, inter-rater reliability, outcomes after implementation, and racial/ethnic group differences are summarized here.

WRAM

The WRAM was developed by Washington State social service agency in 1986 as a consensus-based instrument. Its contents are continuously updated based on new research evidence. Used at the initial investigation, the instrument currently includes 37 items based on seven major domains: (1) child characteristics, (2) severity of abuse/neglect, (3) chronicity of abuse/neglect, (4) caretaker characteristics, (5) caretaker/child relationship, (6) socio-economic factors, and (7) perpetrator access. To use the instrument, child welfare workers assess and rate the level of risk that they perceive for each item on a five point scale. Based on these ratings, families are categorized into risk levels. The instrument assesses risk of maltreatment in general rather than considering the risks for each kind of abuse (e.g., neglect, abuse, sexual abuse, etc.).

In tests of predictive validity, performance of the WRAM was mixed. One study of 1400 cases from four sites across the country found that while rates of subsequent investigation were higher for moderate or high risk families than for low risk families, rates of substantiated maltreatment for families in low, moderate or high risk groups were not significantly different (Baird & Wagner, 2000). When a slightly different outcome was used (the number of subsequent substantiated reports received by a family within two years), the same difficulty in accurately classifying families emerged (Baird & Wagner, 2000).

Another study of WRAM using different analytic techniques and outcomes produced similar findings. In this study of ten child welfare agency offices in New Jersey (n = 239), Camasso and Jagannathan (1995) adapted five of the seven major domains of the instrument into scales to measure the domains, and two domains were assessed individually. When these variables were entered into a regression model designed to predict subsequent maltreatment, the variables that measured the severity of abuse were found to be *negatively* correlated with maltreatment (i.e., the more severe the abuse the less likely the parent to maltreat again). The variables measuring child characteristics, caretaker characteristics, the parent-child relationship, and socio-economic status were not associated with subsequent maltreatment, while the variable measuring child behavior problems was positively associated with subsequent maltreatment. The model had poor predictive power overall, explaining approximately 6% of the variability in the outcome. In addition, based on a plot of the sensitivity and specificity of the instrument overall, authors concluded that the performance of the instrument ". . . . might be characterized as generally poor" (Camasso & Jagannathan, 1995).

In a different study of the WRAM in the state of Washington (n = 12,329), cases coded "low or no risk" were less likely to have a subsequent referral within 18 months than were cases coded as moderate and high risk. The cases coded as moderate risk, however, were not less likely to be re-referred than were high risk cases. Eleven items from the instrument were positively associated with re-referral, and eight items with recurrence of substantiated maltreatment. This study also showed that the average risk ratings of re-referred cases did not differ significantly from risk ratings of cases not re-referred (English, Marshall, Brummel & Orme, 1999).

Another study by English and Graham (2000) considered the convergent validity of the WRAM. In this study (n = 261), the instrument items were used as scales and correlated with other scales or items from well-known measures of the same constructs. A strong correlation would suggest the WRAM item did a good job of measuring that area. Alternative measures of constructs were available for only nine of the 37 items on the WRAM. Of these, items assessing child development and behavior problems were not associated with measures of similar constructs. Four of the 5 items related to the caregiver were found to be associated with measures of similar constructs, but the item related to stress and social support was not associated with related measures (English & Graham, 2000). As a result, the findings are mixed; some items do appear to have a degree of convergent validity, while others do not.

Two studies were identified that assessed the inter-rater reliability of the instrument. In the first study, four raters were asked to rate the same 80 cases using the instrument. All four workers classified families in the same way ess than 14% of the time, and three out of four workers did so just over half the time. Because some portion of these agreements could be due to chance, a statistical correction was done which produces a "kappa" score. Kappa varies from -1 to + 1; a kappa of 0 would mean the performance of the instrument was no better than chance. According to the authors of the study, a kappa in the range of .5 to .6 is generally considered acceptable. The score for the WRAM was 0.18 (Baird, Wagner, Healy & Johnson, 1999).

No studies were found that considered the effects of implementation of the WRAM on case outcomes. Several studies have considered the use of the WRAM with different racial/ethnic groups. One study of 8785 cases in Washington state found that African American and Native American families were more likely to be assigned to the highest risk level than their numbers in the referral population would suggest; Asian American families were under-assigned to the highest risk level

(English et al., 1995). However, it should also be noted that in a subsequent multivariate analysis of 12,329 cases in the same state, Native American families were in fact more likely to be re-referred, and Asian families less likely to be re-referred (English et al., 1999). Another study of 1400 cases in four sites found that approximately equal percentages for African American and White families were classified into each risk level by the WRAM (Baird & Wagner, 2000).

CFAFA (The "Fresno Model")

This consensus-based instrument, the CFAFA (California Family Assessment Factor Analysis) or the "Fresno Model," is derived from an instrument originally developed by the state of Illinois (the Child Abuse and Neglect Tracking System or CANTS 17B). The instrument is no longer used in Illinois, but has been used most recently in California. The instrument can be used throughout the life of a case. The CFAFA has 23 items that fit within five theoretical domains: (1) precipitating incident, (2) child assessment, (3) caregiver assessment, (4) family assessment, and (5) family-agency interaction. All types of maltreatment are considered together. A social worker rates each item as low, moderate or high risk and sums the number of items coded at each risk level in order to determine the overall level of risk.

In tests of predictive validity, the CFAFA did not perform well. While in a study of 1400 cases rates of subsequent investigation were higher for moderate or high risk families than for low risk families, rates of substantiated maltreatment for low, moderate or high risk families were not significantly different (Baird & Wagner, 2000). This was also true when the number of subsequent substantiated reports received by a family within two years was used as the outcome instead of the presence of any subsequent substantiated report (Baird & Wagner, 2000).

Another study examined the Illinois CANTS 17B that is the foundation of the CFAFA. When four items from this instrument were used as variables in a multivariate model predicting subsequent maltreatment (n = 239), none were associated with the outcome and the model had poor predictive power overall (explaining only 1% of the variability in the outcome). In addition, when authors plotted the sensitivity and specificity of the instrument, they concluded that "performance. . . . might be characterized as generally poor" (Camasso & Jagannathan, 1995).

One study attempted to determine the inter-rater reliability of the instrument. Four raters were asked to rate the same 80 cases using the instrument. Just over 16% of the time, all four workers classified families into the same risk groups; about 45% of the time, three out of four workers did so. The kappa for the CFAFA was 0.184 (Baird, Wagner, Healy & Johnson, 1999).

No studies were found that considered the effects of implementation of the CFAFA on case outcomes of the CFAFA. One study of 1400 cases considered use of the CFAFA with different racial/ethnic groups. In this study, the CFAFA classified approximately equal percentages of African American and White families into each risk level (Baird & Wagner, 2000).

CARF

The CARF (Child At Risk Field System) was one of the first risk assessment instruments to focus on safety as distinct from risk and was developed by ACTION for Child Protection. This consensus-based instrument can be used throughout the life of a case. It includes fourteen items within the following five domains: (1) child; (2) parent; (3) family; (4) maltreatment; and (5) intervention. Four "qualifiers" are also to be considered: (1) duration of a negative influence; (2) pervasiveness of a negative influence; (3) acknowledgement by parents of a negative influence; and (4) control of the negative influence. All types of maltreatment are considered together. Each item or qualifier is rated on a four point scale; the average of the 14 items plus the average of the four qualifiers is summed and divided by 2 to arrive at a final risk score. The family is then categorized into no risk, low risk, moderate risk, significant risk, or high risk groups.

The performance of CARF on tests of predictive validity was mixed. In one study of 207 cases in New York state, families assigned to the highest risk group were more likely to have a subsequent referral than families assigned the lowest risk group, though the relationship only "approached" statistical significance. Particular items were not found to be associated with subsequent maltreatment (Doueck, Levine & Bronson, 1993).

One study (Kolko, 1998) assessed the convergent validity of CARF (n = 90). The child welfare worker ratings of the "parent risk field" and the "family risk field" were not found to be associated with any of eight clinical measures of related constructs against which they were each tested. The child welfare worker ratings of the "child risk field" was found to be associated with one clinical measure of parent-reported "child to parent violence." However, the rating of the child risk field was also found to be negatively associated with the level of child PTSD (Post Traumatic Stress Disorder) which was the opposite direction than expected and no relationship was found between the ratings and five other clinical measures of related constructs (Kolko, 1998).

Only one study that assessed the inter-rater reliability of CARF was found. In this study (Fluke et al., 1993), 25-50 workers from several counties in Pennsylvania were asked to read case vignettes and assess the level of risk using the risk assessment instrument. The scores of all pairs of coders of a case vignette were correlated and those correlations averaged. Alpha coefficients for the CARF instrument overall risk scores for three different vignettes varied widely, ranging between .067 to .952 (Fluke et al., 1993).

In a study comparing substantiation rates before and after implementation of CARF in one New York county (n = 207), no differences were found. When maltreatment type was considered separately, physical neglect was found to be somewhat more likely to be substantiated before CARF was implemented than after it was implemented. No difference in "before and after" substantiation rates were found for physical maltreatment, sexual maltreatment, medical neglect, emotional maltreatment, or educational neglect (Doueck, Levine & Bronson, 1993).

No studies were found that considered the use of CARF with different racial/ethnic groups.

CERAP

A consensus-based instrument, the CERAP (Child Emergency Response Assessment Protocol) was developed as a "safety assessment" by Illinois Department of Child and Family Services, the American Humane Association, the University of Illinois, and experts in the field. All types of maltreatment are considered together. The instrument includes fourteen items where and can be used throughout the life of the case. The child welfare worker notes the presence or absence of each item; if any of the items are present, the worker decides whether the child is "safe" or "unsafe." If the worker decides the child is unsafe, a safety plan is developed. The training for using the instrument includes a rigorous testing and certification process.

One study attempted to assess the predictive validity of the CERAP. Since the CERAP focuses on safety assessment, the outcome used was subsequent substantiated maltreatment within 60 days. The use of the CERAP was assessed at two different points in time in the case; namely, initial investigation (n = 380) and within five days of case opening (n = 350). At initial investigation, neither overall safety assessment nor number of safety factors identified were associated with subsequent maltreatment within 60 days, either in bi-variate or multivariate tests that controlled for CERAP completion, prior reports, total number caregiver problems, and service receipt (Fuller, Wells & Cotton, 2001). Within five days of opening the case, both the safety assessment and the number of safety factors identified were found to be associated with subsequent maltreatment within 60 days in bi-variate tests, but these relationships did not remain in multivariate tests once other factors had been controlled for. Completion of the instrument, regardless of safety rating, was negatively associated with subsequent maltreatment (Fuller et al., 2001).

In considering changes in the 60 day maltreatment recurrence rates after implementation of CERAP compared to the year before implementation in one state, a series of studies found a reduction in the maltreatment recurrence rates that has been maintained for six years following implementation (Garnier & Nieto, 2002; Nieto & Garnier, 2001). Several alternative explanations for the reductions (increased use of out of home placement, another policy, nationwide trend) were considered and ruled out in a follow-up study (Fluke, Edwards, Bussey, Wells & Johnson, 2001).

No studies were found that considered the inter-rater reliability of the CERAP, or the use of the CERAP with different racial/ethnic groups.

CRC ACTUARIAL INSTRUMENTS FOR RISK ASSESSMENT

Slightly different versions of this actuarial risk assessment instrument have been developed by the Children's Research Center (CRC) for various jurisdictions. These instruments are based upon the statistical association of variables with substantiated maltreatment injury, foster care placement, and reinvestigation within two years in each location. The instrument is used at the initial investigation and includes two subscales of ten items each; one subscale assesses risk of neglect and the other risk of physical or sexual abuse. Each item is scored with a 0, 1, or 2 as indicated on the instrument and each subscale is summed. Based on the highest subscale score, a family is classified into a low, moderate, high, or very high risk category. In most jurisdictions, workers can override the risk classification and increase the risk rating by one level.

A number of studies have assessed the predictive validity of the CRC risk assessment instruments and found that the instruments are able to distinguish between low, medium and high levels of risk of subsequent maltreatment. Families categorized as high risk by the instrument have a distinctly higher rate of subsequent maltreatment than do families categorized as moderate risk. Similarly, moderate risk families have a distinctly higher rate of subsequent maltreatment than do families categorized as low risk. This was found to be true for subsequent maltreatment within 6 months (Johnson, 2004), 18 months (Baird & Wagner, 2000), and 2 years (Johnson, 2004; Loman & Siegel, 2004), and for the total number of subsequent substantiated reports received by a family at 18 months (Baird & Wagner, 2000) and 24 months (Loman & Siegel, 2004). Additionally, in a multivariate study, families coded as higher risk showed a stronger association with subsequent maltreatment (controlling for ethnicity, county size, service receipt, and safety finding) than did families coded as moderate or low risk (Johnson, 2004).

The CRC risk assessment instruments performed fairly well in reliability studies. In one study (Baird, Wagner, Healy & Johnson, 1999), four raters were asked to rate the same 80 cases using the instrument. Over half of the time, all four workers classified families in the same way; 85% of the time, three out of four workers did so. The kappa score for the CRC Risk Assessment instrument was .562 (Baird, Wagner, Healy & Johnson, 1999). In a second study that involved coding case vignettes, most workers scored the subscales within 4 points of one another (scores ranged from 0-20). Somewhat lower consistency was realized when scores were combined for an overall risk score.

No study was found that assessed the use of the CRC risk assessment instrument with respect to outcomes. One study (Wagner, Hull & Luttrell, 1995) assessed outcomes in one state following the implementation of an *array* of CRC instruments that included the actuarial risk assessment. Compared to a demographically matched set of counties that were not implementing the array of instruments, counties implementing the CRC array had: (1) lower rates of re-referral or substantiation for cases closed without services, (2) families received more services, particularly if they were high risk, and (3) and referral rates, substantiation rates, removal rates, and injuries were lower (Wagner, Hull & Luttrell, 1995).

The findings from studies that assess the use of the instruments with different racial/ethnic groups are mixed. Some studies found that the instrument classifies approximately equal percentages of all ethnic groups into each risk level (Baird & Wagner, 2000; Baird, Ereth & Wagner, 1999; Johnson, 2004), that rates of recurrence for different risk categories are consistent across ethnic groups (Baird, Ereth & Wagner, 1999), and that the association of scores with subsequent maltreatment does not differ by ethnic group (Johnson, 2005; Johnson, 2004). However, studies have also found that white families were somewhat more likely to be coded higher on more items than families of color (Johnson, 2005; Johnson, 2004); African American families scored slightly higher on neglect scale, and white families scored slightly higher on physical abuse/sexual abuse scale (Baird, Ereth & Wagner, 1999), Southeast Asian and Hispanic families in Minnesota were slightly more likely to be coded as low risk than other ethnic groups, and Native American families were slightly more likely to be coded High and Intensive Risk (differences were greatest on the neglect scale) (Loman & Siegel, 2004). Additionally, in the Minnesota study, when predictive validity was considered separately by ethnic group, distinctions between maltreatment rates of risk groups were smaller for Native American families due to the high referral rate in the Low Risk group (Loman & Siegel, 2004).

Risk Assessment Model of Child Protection (Ontario)

Based upon scales originally developed by Magura and Moses (1986), this consensus-based instrument was modified by a research team from the University of Toronto in consultation with the Ontario Association of Children's Aid Societies. Twenty-two items within five domains (caregiver, child, family, intervention, and abuse/neglect) are assessed by means of a 4-point Likert-type scale scoring system. The child welfare worker determines a total overall assessment of risk (low to high) and the cumulative risk score (the total of the ratings from the five domains). The instrument is used to decide whether or not a child should be removed and placed into foster care. All types of maltreatment are considered together. No studies were found that assessed predictive validity (using outcomes of subsequent referral or maltreatment), outcomes, or racial/ethnic differences of the Ontario Risk Assessment Model.

One study (Lescheid et al., 2003) was found that assessed the inter-rater reliability of the Ontario Risk Assessment Model. A reliability score of r = .92 is reported for the cumulative risk score, and of r = .96 for the overall risk (Lescheid et al., 2003). However, it is not clear how many coders were used in the reliability assessment, how many cases were assessed for reliability, or by what means reliability was determined.

Utah Risk Assessment Scales

The Utah Risk Assessment Scales are based upon Family Risk Scales and Child Well-Being Scales originally developed by Magura and Moses (1986) along with additional scale items developed by members of the Utah Department of Social Services and the Utah Child Welfare Training Project. These additions reflected the practice experience of the staff members involved in the development of the instrument (Nasuti, 1998). This consensus-based Likert-type instrument is composed of 32 items within five domains: parent, child, family, maltreatment, and intervention. The instrument was designed for all types of maltreatment and can be used by child welfare intake and investigative workers. No studies were found that considered the predictive validity, outcomes, or racial/ethnic differences for the Utah Risk Assessment Scales.

One study (Nasuti, 1998; Nasuti & Pecora, 1993) assessed the inter-rater reliability of the Utah Risk Assessment Scales. To assess inter-rater reliability, eight vignettes were developed describing cases of varying severity. Child welfare workers, supervisors and child welfare experts were asked to review each case vignette and provide a risk rating using the Utah Risk Assessment Scales. These risk scores were then correlated to determine reliability scores. Pearson's r coefficients ranged from .568 to .855 for the eight vignettes, each of which was assessed by 22 to 28 raters. A "Spearman-Brown prophecy formula" was applied to these scores to provide a "stepped-up" reliability estimate that was perceived to provide a more accurate estimate of inter-rater reliability. Stepped-up estimates were all above .970 (Nasuti, 1998; Nasuti & Pecora, 1993).

DISCUSSION AND IMPLICATIONS

Overall, the available research suggests that the CRC risk assessment instruments appear to have greater predictive validity than the available consensus-based instruments. This may be related to their development via the statistical identification of the strongest predictors of a particular outcome in that state or county. Unless the sample used to develop that model was different from the typical population referred to child welfare agencies in that jurisdiction (or there were major changes in the local context or population demographics), it would be reasonable to assume that those variables in the model would continue to be predictive of outcomes experienced by subsequent cohorts. The processes for identifying factors for consensus-based instruments may simply be less accurate at identifying the strongest predictors of maltreatment.

Convergent validity was not formally assessed for the CRC risk assessment instruments, and the performance of consensus-based instruments in this area was generally poor. These instruments may be unreliable and/or include measures that do not adequately reflect underlying concepts. It is difficult to draw conclusions across studies on inter-rater reliability. Only one study compared several instruments at the same time and found that the consensus-based instruments performed less effectively than the CRC actuarial instruments. Other studies of a consensus-based instrument have found variable or high inter-rater reliability. Higher inter-rater reliability might be expected from actuarial instruments because items in these instruments are more often objective while items in consensus-based instruments are more often subjective and less precise. For example, in a question related to prior CPS history using the consensus-based WRAM, the child welfare worker is required to determine whether past incidents were 'isolated' or 'intermittent,' and whether there is evidence of 'minor' abuse and neglect or 'moderate' abuse and neglect. In comparison, the actuarial CRC instrument asks 'whether or not' there was a prior injury to a child from abuse or neglect, or 'whether or not' there was a prior investigation. Generally, well-defined, objective, and clearly articulated measures are more likely to be reliable because differences of opinion about the meaning or coding of factors are minimized (Rycus & Hughes, 2003). In support of this conclusion, English & Graham (2000) noted that in a study of the CRC actuarial instrument (Wood, 1997), objective items on the instrument, such as the age of a child, had higher reliability than subjective items, such as "was child inadequately supervised?" In addition, some consensus-based instruments often require coders to use their judgment regarding the level or risk related to an area. That is, rather than asking workers "whether or not" maltreatment previously occurred, a consensus-based instrument asks workers to assign a level of risk to the broad area of previous maltreatment. Because different workers could perceive or define risk differently, different levels of risk could be assigned to similar situations.

In terms of outcomes, studies of the CERAP and the CRC instruments suggested that implementation resulted in improved outcomes. These instruments may be improving the accuracy of worker assessments of risk, resulting in fewer high risk children being left at home to be re-abused. The findings regarding the use of the instruments with different racial/ethnic groups are mixed.

It is important to note that the available research is limited. For any particular instrument, there were only a few studies available and sometimes only a single study was found. Therefore, conclusions about the risk assessment instruments should be considered preliminary and in need of further study.

IMPLICATIONS FOR PRACTICE

The debate about the best approach to assessing risk and safety may be related to a lack of clarity regarding the purposes of "risk assessment." Distinctions between *risk assessment* and *family assessment* can be somewhat unclear (Lyons et al., 1996), and a number of researchers have argued that they have often been confused (Rycus & Hughes, 2003; Wald & Woolverton, 1990). "While risk assessment is designed to accurately estimate the likelihood of future incidents of maltreatment, the purpose of family assessment is to identify and explore, in considerable depth, the unique complex of developmental and ecological factors in each family and their environment that may contribute to or mitigate maltreatment" (Rycus & Hughes, 2003, p. 11). Some researchers assert that risk assessment and family assessment are separate and distinct, and that neither activity is served by attempting to use a single instrument to do them both, or even by attempting to do them at the same time (Rycus & Hughes, 2003; Wald & Woolverton, 1990).

If the goal of an assessment is to *predict the likelihood of the recurrence of maltreatment* in order to provide services to the families at greatest risk, this is clearly a risk assessment. The research evidence suggests that the actuarial instrument will produce a more accurate and reliable prediction than the consensus-based instruments. On the other hand, if the goal of an assessment activity is to *gain a comprehensive understanding of the service needs of a family or individual,* a family needs assessment instrument incorporates more items and thus provides more information. However, consensus-based instruments did not have high convergent validity, suggesting they may not accurately measure the relevant characteristics and thus would not necessarily be helpful in family assessments.

Finally, most research on risk assessment acknowledges that the use of any kind of risk assessment instrument, actuarial or consensus-based, requires good clinical skills (Doueck et al., 1993; Johnson, 2004). For example, the CRC actuarial instrument contains numerous items that require clinical judgment to score, and allow for a clinical over-ride based on family characteristics or dynamics that are likely to affect risk but are not included on the actuarial instrument. As Ereth et al. have noted, "... A caseworker can sense things that an actuarial instrument would ignore or could not employ . . . Many characteristics of human subjects simply cannot be quantified empirically and actuarial models cannot easily account for rare events" (2003, p. 3). Therefore, clinical judgment can never be eliminated from any risk assessment process. In fact, many researchers in child welfare stress that the instruments for risk and safety assessment should be understood as decision aids to enhance or expand upon clinical judgment, rather than as a competing approach (Ereth et al., 2003; Fuller et al., 2001; Munro, 1999). As Munro observed, "... Errors can be reduced if people are aware of them and strive consciously to avoid them. The challenge is to devise aids to reasoning that recognize the central role of intuition and do not seek to ignore or parallel it but, by using our understanding of its known weakness, offer ways of testing and augmenting it" (1999, p. 756).

IMPLICATIONS FOR RESEARCH

Most of the available research on risk and safety assessment instruments is limited to five well-known instruments: the WRAM, CERAP, CARF, CFAFA (the "Fresno" instrument), and the CRC actuarial instrument. Clearly, further research in the area of risk assessment is needed. One next step in this area might be an on-going survey of the utilization of risk assessment instruments across all the states, as currently there is no process by which such information is gathered, updated and made available to the practice and research community. As a result, it is not clear how many jurisdictions use actuarial instruments, consensus-based instruments, or none at all, nor which instruments are used. In addition, much of the research on various instruments has been conducted by researchers who are associated with the development of those instruments. While this research is of high quality and has been published in respected peer-reviewed publications, studies conducted by independent researchers are also needed.

It is also important to note that current research focuses primarily on one decision point in the case; namely, the initial investigation. There is growing interest in utilizing instruments that can assist child welfare workers in making decisions at other points in the life of a case. Therefore, there is a clear need for the development of research-based instruments that can be validated for other decision points such as placement and reunification.

While predictive validity studies are needed for any instrument that attempts to assess the likelihood of future maltreatment, some researchers have suggested that other relevant outcomes besides recurrence need to be considered, such as severity of abuse (Wald & Woolverton, 1990). On-going validity and reliability studies are necessary so that the instruments can be continually refined and improved over time. Many researchers have also suggested that the effects of services provision need to be taken into consideration (English & Aubin, 1990; Milner, 1994; Nasuti, 1998). Wald and Woolverton assert that a risk assessment instrument is "truly useful only if it identifies the likelihood of re-abuse *given specific interventions*" (1990, p. 491). Since the provision of services may reduce risk, predictions of future abuse and neglect that fail to take services into account may possibly overestimate risk. For example, the lack of consideration of child-caregiver interactions (Morton, 2004b) or neighborhood factors in current instruments need to be taken into account in future research. Lastly, more research should consider the quality and nature of the implementation process, especially worker acceptance or resistance to the use of risk assessment instruments (English & Aubin, 1990).

CONCLUSION

This review of the available research literature on instruments of risk and safety assessment in child welfare suggests that CRC actuarial instruments have stronger predictive validity than available consensus-based instruments. This structured review was limited by: (1) the lack of studies on decision points other than initial investigation, (2) the variability in definitions and measures across studies, and (3) the relatively small number of studies examining risk assessment instruments. Nonetheless, the findings should be useful to practitioners and researchers evaluating the various approaches to risk and safety assessment in child welfare.

REFERENCES

Albers, M. & Roditti, M. (2004). *Management values and decision-making instruments: Nuts and bolts of risk assessment in California.* Bay Area Academy, Berkeley, CA.

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

American Federation of State County and Municipal Employees (1999). *Liability and child welfare workers.* Available online at: http://www.afscne.org/publications/child/*cww99205.htm.*

Baird, C., Ereth J. & Wagner, D. (1999). *Research-based risk assessment: Adding equity to CPS decision-making.* Madison, WI: Children's Research Center.

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, C. Wagner, D. Healy, T. & Johnson, K. (1999). Risk assessment in child protective services: Consensus and actuarial instrument reliability. *Child Welfare, 78*(6), 723-748.

Camasso, M. J. & Jagannathan, R. (2000). Instrumenting the reliability and predictive validity of risk assessment in child protective services. *Children and Youth Services Review, 22*(11/12), 873-896.

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

Children's Research Center (n.d.). SDM: Structured decisions made in child welfare. Retrieved August 15, 2005 from www.nccd-crc.org/crc/c_sdm_about.html.

Child Welfare Services Stakeholders Group (2003). *CWS Redesign: The future of California's child welfare services–final report.* State of California Health and Human Services Agency, Department of Social Services.

Cicchinelli, L. (1990). Risk assessment: Expectations, benefits and realities. *Fourth National Roundtable on CPS Risk Assessment.* San Francisco, CA: American Public Welfare Association.

Costello, T. (1995). Why is it so hard to implement risk assessment? *Ninth National Roundtable on CPS Risk Assessment.* San Francisco, CA: American Public Welfare Association.

Curran, T. F. (1995). Legal issues in the use of CPS risk assessment instruments. *From the APSAC Advisor, 8*(4), 1-4.

Davidson, H. (1991). Risk assessment and the law: Unsettled issues. *Fifth National Roundtable on CPS Risk Assessment.* San Francisco, CA: American Public Welfare Association.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision-making. *American Psychologist, 34,* 571-582.

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

DePanfilis, D. (1996). Implementing child mistreatment risk assessment systems: Lessons from theory. *Administration in Social Work, 20*(2), 41-59.

Doueck, H. J., English, D. J., DePanfalis, D. & Moote, G. T. (1993). Decision-making in child protective services: A comparison of selected risk assessment systems. *Child Welfare, 72*(5), 441-452.

Doueck, H. J., Levine, M. & Bronson, D. E. (1993). Risk assessment in child welfare services: An evaluation of the Child at Risk Field System. *Journal of Interpersonal Violence, 8*(4), 446-467.

Downing J. D. Wells S. J. & Fluke J. (1990). Gatekeeping in child protective services: A survey of screening policies. *Child Welfare, 69* (4), 357-369.

English, D. J. (1999). Evaluation and risk assessment of child neglect in public child protection services. In H.Dubowitz, (Ed.), *Neglected children: Research, practice, and policy* (pp. 191-210). Thousand Oaks, CA: Sage Publications, Inc.

English, D. & Aubin, S. W. (1990). Outcomes for screened-out and low-risk cases within a child protective services risk assessment system. *Fourth National Roundtable on CPS Risk Assessment* (pp. 41-58). San Francisco, CA: American Humane Association.

English, D. J. & Graham, J. C. (2000). An examination of relationship between children's protective services social worker assessment of risk and independent LONGSCAn measures of risk constructs. *Children and Youth Services Review, 22*(11-12), 897-933.

English, D. J., Marshall, D., Brummel, S. & Orme, M. (1995). A preliminary examination of similarities and differences I the assessment of risk for different ethnic groups. *Ninth National Roundtable on CPS Risk Assessment,* pp. 195-218. San Fransisco, CA: American Human Association.

English, D. J., Marshall, D., Brummel, S. & Orme, M. (1999). Characteristics of repeated referrals to child protective services in Washington state. *Child Maltreatment* 4(4), 297-307.

Ereth, J., Johnson, K. & Wagner, D. (2003). *New Mexico Children, Youth and Families Department Foster Provider Risk Assessment Study.* Madison, WI: Children's Research Center.

Fluke, J., Edwards, M., Bussey, M., Wells, S. & Johnson, W. (2001). Reducing recurrence in child protective services: Impact of a targeted safety protocol. *Child Maltreatment, 6*(3), 207-218.

Fluke, J., Wells, S., England, P., Walsh, W., English, D., Gamble, T. & Woods, L. (1993). Evaluation of the Pennsylvania approach to risk assessment. *Seventh National Roundtable on CPS Risk Assessment* (pp. 116-170). San Francisco, CA: American Humane Association.

Fuller, T. L., Wells, S. J. & 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.

Galasso, L. B. (2001). *Toward the prevention of child maltreatment through risk assessment:* Evaluation of an ecological, prospective instrument of risk for child abuse potential. Dissertation, Michigan State University Department of Psychology.

Garnier, P. & Nieto, M. (2002). *Illinois Child Endangerment Risk Assessment Protocol evaluation: Impact on short-term recurrence rates–year six +* . Children and Family Research Center, University of Illinois at Urbana-Champaign.

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

Hollinshead D. & Fluke J. (2000). What works in safety and risk assessment for child protective services. In M. Kluger G. Alexander and P. Curtis (Eds.), *What Works in Child Welfare* (pp. 67-74). Washington D.C.: CWLA Press.

Johnson, W. (2004). *Effectiveness of California's child welfare Structured Decision-making Instrument (SDM): A prospective study of the validity of the California Risk Assessment.* Unpublished report.

Johnson, W. (2005). Effects of a research-based risk assessment on racial/ethnic disproportionality in service prevision decisions. In D. Derezotes, J. Poertner & M. F. Testa (Eds.), *Race Matters in Child Welfare: The Overrepresentation of African American Children in the System.* Washington, D.C.: CWLA.

Kahneman, D. & Tversky, A. (1982). On the psychology of prediction. in D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 50-98). New York, NY: Cambridge University Press.

Kanani, K., Regehr, C. & Bernstein, M.M. (2002). Liability considerations in child welfare: Lessons from Canada. *Child Abuse and Neglect,* 26, 1029-1043.

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(3), 262-275.

Leschied, A. W., Choido, D., Whitehead, P. C., Hurley, D., & Marshall, L. (2003). The empirical basis of risk assessment in child welfare: The accuracy of risk assessment and clinical judgment. *Child Welfare, 82*(5), 527-540.

Loman L. A. & Siegel G. L. (2004). *An evaluation of the Minnesota SDM Family Risk Assessment.* St. Louis, MO: Institute for Applied Research. Available from: http://www.iarstl.org.

Lyle, C. G. & Graham, E. (2000). Looks can be deceiving: Using a risk assessment instrument to evaluate the outcomes of child protection services. *Children and Youth Services Review, 22*(11/12), 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*(3), 143-155.

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

Milner, J. S. (1994). Assessing physical child abuse risk: The Child Abuse Potential Inventory. *Clinical Psychology Review, 14*(6), 547-583.

Morton, T.D. (1999). The increasing colorization of America's child welfare system: The overrepresentation of African American children. *Policy & Practice,* 57(4), 23-30.

Morton, T. D. (April, 2004a). Where assessment fails. *Commentary.* Duluth, GA: Child Welfare Institute.

Morton, T. D. (Sept, 2004b) *Caretaker and child interactions in child maltreatment.* Duluth, GA: Child Welfare Institute.

Morton, T. D. & Salovitz, B. (n.d.) *Evolving a theoretical instrument of child safety in maltreating families.* Unpublished manuscript.

Munro, E. (1999). Common errors of reasoning in child protection work. *Child Abuse and Neglect, 23*(8), 745-758.

Munro, E. (2004). A simpler way to understand the results of risk assessment instruments. *Children and Youth Services Review, 26*(9), 873-883.

Munro, E. & Rumgay, J. (2000). Role of risk assessment in reducing homicides by people with mental illness. *British Journal of Psychiatry,* 176, 116-120.

Nasuti, J. P. (1998). Risk assessment in child protective services: Challenges in measuring child well-being. *Journal of Family Social Work, 3*(1), 55-70.

Nasuti, J. P & Pecora, P. J. (1993). Risk assessment scales in child protection: A test of the internal consistency and interrater reliability of one stateside system. *Social Work Research and Abstracts, 29*(2), 28-33.

Nieto, M. & Garneir, P. (2001). *Illinois Child Endangerment Risk Assessment Protocol evaluation: Impact on short-term recurrence rates–year five.* Children and Family Research Center, University of Illinois at Urbana-Champaign.

Nohejl C. Doueck H. J. & Levine M. (1992). Risk assessment implementation and legal liability in CPS practice. *Law & Policy,* 14(2 & 3),185-203.

Repetosky C. & Bailey D. (1988). The risk "fit": Integrating risk assessment with case management and practice decisions in child protection and child welfare services. *Second National Roundtable on CPS Risk Assessment.* San Francisco, CA: American Public Welfare Association.

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

Scheurman, J., Rossi, P. H. & Budde, S. (1999). Decisions on placement and family preservation. *Evaluation Review, 23*(6), 599-618.

Schwalbe, C. (2004). Re-visioning risk assessment for human service decision making. *Children and Youth Services Review, 26*(6), 561-576.

Sheets, D. (1992). How Texas Got S.M.A.R.T.: A description of the rapid application design and development process used by Texas to design and implement a statewide risk assessment system. *Sixth National Roundtable on CPS Risk Assessment.* San Francisco, CA: American Public Welfare Association.

Squadrito, E., Neuenfeldt, D., & Fluke, J. (1994). Findings of Rhode Island's two-year research and development efforts. *Eighth National Roundtable on CPS Risk Assessment.* San Francisco, CA: American Public Welfare Association.

Thompson L., Brown J. & Pecora P. J. (1989). Implementation of risk assessment systems: Training and quality assurance issues. *Third National Roundtable on CPS Risk Assessment.* San Francisco, CA: American Public Welfare Association.

Tumlin, K. C. & Geen, R. (2000). The decision to investigate: understanding state child welfare screening policies and practices. *No.A-38 in New Federalism Issues and Options for States, Series A.* Washington, D.C.: The Urban Institute.

Wagner, D., Hull, S. & Luttrell, J. (1995). Structured decision making in Michigan. *Ninth National Roundtable on CPS Risk Assessment,* pp. 167-191. San Francisco, CA: American Humane Association.

Wald, M.S. & Woolverton, M. (1990). Risk assessment: The emperor's new clothes? *Child Welfare, 69*(6), 483-511.

Wood, J.M. (1997). Risk predictors for re-abuse or re-neglect in a predominantly Hispanic population. *Child Abuse and Neglect,* 21(4), 379-389.

Zuravin, S. J., Orme, J. G. & Hegar, R. L. (1995). Disposition of child physical abuse reports: Review of the literature and test of a predictive instrument. *Children and Youth Services Review 17*(4), 547-566.

doi:10.1300/J394v05n01_03