

CHILD ABUSE AND NEGLECT: IMPROVING CONSISTENCY IN DECISION MAKING

A Comparative Study of the Use and Effectiveness of Different Risk Assessment Models in CPS Decision Making Processes is a three year study funded by the National Center on Child Abuse and Neglect (Grant 90-CA-1550). The objectives of this study are to determine the relative reliability and validity of three approaches to risk assessment currently used in Child Protective Services (CPS).

NCCD's Children's Research Center has completed the reliability data collection and analysis portions of this three year study. This report describes the issues of reliability of the different approaches. These issues include: the strengths and weaknesses of methods used to measure reliability, the methodology selected to examine and estimate inter-rater reliability, and the results of the analyses.

I. INTRODUCTION

A. Problem Statement

Decisions that protective service workers make at the conclusion of a child abuse or neglect investigation are critical to the protection of children. The most important task workers perform at that point -- especially for families found to have abused or neglected children -- is the decision to either: 1) close the case; 2) open the case for protective service intervention or perhaps intensive in-home family preservation services; or 3) remove the child to foster care. Because this choice has important consequences for children, their families, and protective service agencies, it must be made as consistently and accurately as possible by the front line CPS workers who investigate abuse/neglect complaints. Like most child welfare workers, they must consider the best interest of the child, especially the risk of continued abuse/neglect absent service intervention when making this decision. However, if conclusions regarding the level of risk of a particular case vary widely depending on who does the rating, then some children will be left in situations with a high potential for continued maltreatment while others

who could have remained at home in relative safety will be placed in foster care. Mistakes can have enormous consequences, ranging from unnecessary expenditures to emotional upheaval and trauma to serious injury and even the death of a child.

There is no question that decision making in human services is difficult and that personal relationships in families served are dynamic and difficult to assess. It is also evident that those charged with protecting children - CPS investigative staff and managers - represent a wide spectrum of educational backgrounds, personal and professional experiences, and bring different values and perspectives to the job. This mix of conditions -- the potentially grave consequences of "error," the inherent difficulty of accurately assessing family situations and relationships, and the range of "skills" evident in the nation's CPS staff -- presents a near-perfect equation for widespread disparity in case decision making.

B. CPS Risk Assessment Systems and the National Center on Child Abuse and Neglect (NCCAN) Study

This brings us to the subject of the current study - CPS risk assessment systems. These are systems developed to help CPS workers accurately estimate the future risk of abuse/neglect and thus make better service decisions for families. The expectation is that risk assessment systems will improve worker decision making at CPS investigation and make it more consistent. A variety of systems for estimating risk have been developed to do this and adopted by various states or counties over the years, but the use of formal risk assessment in CPS is a relatively recent phenomenon (Tatara, 1987). The American Public Welfare Association (APWA) conducted a survey in 1996 of 54 states, territories, and large county child welfare agencies to determine use and satisfaction with CPS risk assessment systems. Of the 44 jurisdictions that responded, 38 had some risk assessment or safety assessment in place. Of these states, 26 first implemented their risk assessment after 1987 (Tatara, 1996).

While risk assessment models continue to be developed throughout the country, experts home expressed concern that the theoretical and empirical support for these systems is

inadequate (Cicchinelli, 1991). As noted by Murphy-Berman (1994), risk assessment procedures vary on a number of dimensions, and the task of comparing one to another is quite complex. Generally, CPS risk assessment systems fall into two basic types:

1. Consensus-Based Systems - In these systems, workers assess specific client characteristics identified by the consensus judgement of experts and then exercise their own clinical judgement about the risk of future abuse or neglect.
2. Actuarial Systems - These systems are based on an empirical study of CPS cases and future abuse/neglect outcomes. The study identifies items/factors with a strong association with future abuse/neglect and constructs an actuarial instrument which workers score to identify Low, Medium, or High risk families.

The NCCAN study described in this report was initiated to examine these two basic types of CPS risk assessment systems. Two questions were posed for this research:

1. How reliable are these systems, i.e., do different workers assign the same risk level to the same family at investigations?
2. How accurately do they estimate future risk of maltreatment, i.e., are the risk estimates made by these systems accurate indications of future maltreatment?

Only the first question is addressed in this report.

Consensus-Based Systems - Research in child abuse and neglect has served as a guide to the development of “expert” or consensus-based risk assessment systems. However, despite a volume of research considered by some as “quite formidable” (Doueck, Levine, and Bronson, 1993), these systems have rarely undergone rigorous evaluations. This fact adds to the significance of this research project.

The child protection philosophy that drives case service practice in child welfare has always required that the social worker perform a prospective evaluation of cases to estimate the risk of future maltreatment. Historically, caseworkers have used the case study method, a form of clinical assessment, to estimate the risk of future maltreatment. In this method of risk assessment, the investigating CPS worker relies almost entirely upon his or her clinical

experience, intuition, and interviewing skills to assess the future risk to the child. In many states, the clinical assessment is structured by an instrument or system that identifies specific case characteristics, often those identified after a review of the literature, for the worker to assess (Tatara, 1987). These instruments may be helpful to CPS workers in conducting a comprehensive case assessment, but they are typically not constructed from an empirical analysis of case outcomes in the jurisdiction where they are used (Wald and Woolverton, 1990). In other words, these instruments help organize the caseworker's clinical assessment of risk, but they are not based on research specific to the jurisdiction (Marks and McDonald, et. al., 1989).

An early study illustrating a critical problem for clinical judgment was conducted by Margaret Blenkner in 1954. Three expert clinical social workers were asked to perform several clinical assessments of 47 clients using information recorded at intake to a private social service agency. After reading each file, making the assessments and recording them on a data collection form, each clinician was asked to make a prognosis for future casework success (the cases were closed and outcomes had been established previously by other clinical judges).

In subsequent analysis, five of the assessment measures these clinicians had recorded demonstrated a strong relationship with case outcome and Blenkner added them together to create a summary score. Although the summary scores demonstrated a very high correlation with successful case outcomes (p. 52), the prognoses of the three clinical judges proved unrelated to outcomes and to one another. Meehl observed that, "Apparently these skilled case readers can rate relatively more specific but still fairly complex factors reliably enough so that an inefficient mathematical formula combining them can predict the criterion; whereas the same judges cannot combine the same data "impressionistically" to yield results above chance" (p. 108).

The Blenkner study illustrates why many observers believe clinical judges do poorly in a predictive setting; they differentially select and weight information about the subject. Despite the fact that the variables in Blenkner's formula had been developed by clinicians and required

clinical skill to observe, clinical judges performed very poorly when asked to predict case outcome.

A more recent study (Rossi, Schuerman, Budde, 1996) compared case decision making among identified "CPS experts" and CPS workers from four states. All were asked to read 70 case vignettes and to decide whether the case should be opened for in-home services or a child (or children) placed in foster care. In a second test of the same cases, a third option, family preservation services, was added to the equation. The researchers found a high degree of variance in decision making even among the CPS experts. Again, it was demonstrated that, in human service decision making, there is considerable disparity in the way cases are handled.

Actuarial Systems - The evidence now available from actuarial studies of child abuse/neglect suggests a conclusion endorsed several years ago in many other fields, namely that actuarial risk assessments derived from simple, empirically validated instruments can efficiently estimate the risk of future maltreatment and, therefore, may substantially improve the clinical risk assessment performed by an individual caseworker. Actuarial assessment methods, which require extensive longitudinal research, have only recently been introduced in Child Protective Services. In their summary of this area of research, Marks and McDonald (1989) cite only two actuarial risk assessment studies in abuse and neglect: an Alameda County study conducted by Johnson and L'Esperance (1984) and an Alaska study conducted by the National Council on Crime and Delinquency (NCCD) (Baird, 1988). However, since the 1989 Marks and McDonald publication, NCCD has conducted additional actuarial research in Oklahoma, Michigan, Rhode Island, and Wisconsin. A large body of research evidence in experimental psychology and corrections supports the view (see, for instance, Meehl, 1954; Sawyer, 1966; or Dawes, 1989) that actuarial instruments can estimate future behavior more accurately than an individual decision maker unaided by actuarial information (even decision makers who have had extensive clinical training). This study, however, may be the first to directly compare the reliability of decisions made with actuarial and consensus-based instruments.

C. Measuring Reliability

CPS agencies have engaged a variety of approaches to improve the consistency of risk assessment decisions including intensive staff training programs, case staffings or teaming concepts that bring a number of different perspectives into the decision making process, and most importantly, the use of structured guidelines for assessing cases. The latter strategy, generally defined as “Risk Assessment” is the subject of this research. Risk assessment systems are simply “formalized methods that provide a uniform structure and criteria for determining risk” (Keller, Cicchinelli, and Gardner, 1988, p.2).

Determining the level of reliability attained by risk assessment systems presents a variety of problems. In actual practice, assessments of families are generally dependent upon a variety of formal and informal activities and observations that include record reviews, personal contacts with the client, collateral contacts with law enforcement, school, medical and other social service personnel, and, in many instances, consultation with colleagues and supervisors. Actual practice, therefore, is nearly impossible to replicate in a test situation.

To measure the reliability of risk assessment systems (and decision making in general), two possibilities emerge: 1) reading social histories and all other documentation contained in case files and using these data to complete risk assessment forms, and 2) the creation of case vignettes (augmented, in some instances, with videotaped “interviews”) to serve as the basis for risk ratings. In both instances, a number of readers assess the same material and their ratings are then compared to determine the extent to which they agree or disagree on the level of risk each “case” represents.

Each approach to measuring reliability has its own strengths and weaknesses. Developing case vignettes gives the researchers greater control over the data provided to readers. This can help ensure that information needed to rate a case is adequate to the task, but it also can introduce an artificial dimension to the research. Vignettes often are based on actual cases, but “enhanced” to provide enough data to answer all (or most) questions

contained in the risk assessment. No matter how objective the researchers are, adding data to case files based on the information requirements of a particular system provides at least the potential for “leading” the rater to a particular conclusion.¹

The other option, reading case files, presents the opposite problem: while more representative of actual case practice, files may not contain all the data necessary to adequately assess cases. In the best-case scenario, researchers should expect that both the quality and quantity of information will vary depending on the skills of the worker involved, the agency’s policies regarding case recording, and the degree to which workers adhere to these policies.

In any event, neither approach will produce all of the data available to an actual decision maker. Information gained through personal contact and observation can not be fully replicated, even when a caseworker’s observations are discussed in the official record. There is, however, no evidence that introducing these data to the assessment process would increase reliability. In fact, information theory strongly suggests that having more factors to consider, especially those resulting from subjective interpretation of client attitudes and personal characteristics, actually reduces inter-rater reliability (as well as the “quality” of the decision) (Clear, 1988).

Several prior studies of reliability of various risk assessment systems have been conducted. However, both the methods used to analyze results and the interpretation of findings raise serious questions regarding the value of some past efforts. In some instances, the number of cases assessed was too small to provide an adequate measure of reliability. For example, one study of the Child at Risk Field (CARF) utilized a single case (Allen, 1988). Other studies, cited by Doueck, English, DePanfilis, and Moote (1993), used three videotaped case scenarios or three case vignettes. While the number of raters was substantial, ranging from 30 (Texas) to 214 (Georgia), attempts to measure reliability using so few cases (particularly case scenarios or vignettes) is suspect in that it could well represent too controlled an experiment to reflect what actually transpires in the field. Further problems arise when the raters in the study have been trained to perform case assessments by viewing the same video case vignettes.

The statistical methods chosen to rate reliability in a number of prior studies raise additional concerns. In general, some measures that have been employed provide more of an estimate of association (or rater “patterns”) than actual agreement on risk ratings. Use of correlation coefficients, for example, are particularly problematic. It is theoretically possible to have two raters disagree on the risk level of every case in the study, yet their risk ratings could be perfectly correlated.² In addition, fairly high correlations are easily attained when the number of rankings possible (e.g., low, moderate, high) are limited.

It is sufficient to say that due to the limitations of prior studies, little is known about the reliability of various risk assessment systems used by child protection agencies. There is, however, a recent study of case decision making that presents an excellent model for assessing reliability. The Chapin Hall Study (1996), cited earlier, combines an adequate number of study cases (70) with an appropriate measure of inter-rater reliability - percent agreement subsequently adjusted for “chance” (using Cohen’s Kappa). This approach is simple, straightforward, and measures precisely what needs to be measured in this study, i.e., the degree to which each risk assessment method enhances the level of consistency among raters.

II. STUDY PARAMETERS

A. Selection of Risk Assessment Systems to be Tested

This study purposefully selected risk assessment models that typify both actuarial and consensus-based systems used in various jurisdictions nationwide.

Two of the more widely used versions of consensus-based systems are the Washington Risk Assessment Matrix, a risk assessment system developed by practitioners in the State of Washington, and the California Family Assessment Factor Analysis, a derivative of the Illinois CANTS system, also known as the Risk Assessment Observations/Recommendation system. The actuarial-based approach identified was the Michigan Structured Decision Making System’s Family Risk Assessment of Abuse and Neglect, a research-based tool constructed during a study of 2,000 Michigan families (NCCD, 1989) and recently revalidated on a cohort of 1,000

families (NCCD, 1995). This system is also used (or now being implemented) in Georgia, Indiana, Kentucky, and Washington, D.C.

1. The Washington Risk Assessment Matrix

According to the Washington State Department of Social and Health Services “Risk Factor Matrix Guide,” the Washington Risk Assessment Matrix identifies and organizes information needed to predict the risk of abuse/neglect. Overall risk is defined as “an assertion of the likelihood of Child Abuse/Neglect (CA/N) absent successful intervention.” The Washington matrix includes six overall risk categories that are defined by “the severity of child abuse and neglect which is likely to occur rather than by the degree of probability of CA/N, no matter how severe or minor. In other words, the overall risk level assumes that CA/N will occur and only asserts how severe the CA/N will be when it occurs.”

The Washington Risk Assessment Matrix groups risk factors by child characteristics (five items), severity of child abuse/neglect (nine items), chronicity or recurrent episodes of child abuse/neglect (one item), caretaker characteristics for primary and secondary caretakers (11 items each), caretaker-child relationship (six items), social economic factors (four items), and perpetrator access. These risk factors are rated from 0 to 5 (0 for no risk, 1 for low risk, 2 for moderately low, 3 for moderate, 4 for moderately high, and 5 for high risk). The ratings system also includes narrative and a summary assessment using the risk factors identified above. Workers use the highest risk element in the severity and chronicity groups to establish a risk assessment baseline and then use the other categories of the matrix to balance these factors against family strengths. If the risk factors outweigh the family strengths, the worker assigns a final risk classification no lower than the baseline risk indicated by the severity and chronicity groups. If the family strengths outweigh the risk factors, the risk indicated by the severity and chronicity groups becomes the upper limit for final risk classification. The worker then assigns a final risk rating.

2. The California Family Assessment Factor Analysis

The second risk assessment model, the California Family Assessment Factor Analysis represents another consensus-based approach to risk assessment. The California Risk Assessment Curriculum for Child Welfare Services “Child Welfare Training” manual subscribes to the following risk definition:

Risk assessment is a process used to assess the level of risk to a child who is reported for alleged abuse and or neglect... It is also a tool which measures and organizes factors present in abuse and neglect situations, and which are considered as important in describing the current safety and in predicting the future safety of the child. (Harris, 1987)

The California risk assessment system recognizes five types of factors: precipitating incident (four factors), child assessment (five factors), caretaker assessment (seven factors), family assessment factors/stressors (five factors), and family/agency interaction (two factors). Each child in the family is rated for each factor into one of five categories: not applicable, insufficient information, low risk, moderate risk, or high risk. These individual assessments are then summarized to reflect the highest risk code for each factor. In addition, each type of factor includes narrative observations. Also included in this risk assessment model is a family assessment narrative of strengths and problems. Based on all of the above, the case is rated as low, moderate, or high risk.

3. The Michigan Family Risk Assessment of Abuse and Neglect

The third risk assessment system included in this study is the Michigan Family Risk Assessment of Abuse and Neglect, which is based on the statistical relationship between behavior and case characteristics and subsequent abuse and neglect. The Michigan model is designed to classify families into four different risk categories based on the likelihood of abuse or neglect in the future. It does not predict future behavior; rather, it is meant to assign cases to different categories based on observed rates of behavior. Like the two consensus based models, the Michigan risk assessment system recognizes that no system can substitute entirely

for the judgment or skill of CPS workers. It can, however, help practitioners focus their assessment on the relatively small set of case characteristics that have demonstrated a strong statistical relationship to future child maltreatment.

The Michigan system consists of three parts. The first two parts are the Neglect Risk Assessment (11 factors), and the Abuse Risk Assessment (12 factors). The third part is a separate Caretaker Strengths and Needs Assessment comprised of 13 factors. In practice, case workers score each individual item on the Neglect and Abuse scales, total the results, and rate the family into a low, moderate, high, or intensive risk classification. The Michigan needs assessment is used in case planning, assisting workers in establishing priorities and in referring cases to needed services.

B. Site Selection

The objective of the site selection process was to identify four distinct sites that would offer: 1) a broad geographical representation; 2) a significant representation of ethnic and racial minorities, including African Americans, Hispanics, and Native Americans; and 3) a mixture of urban and rural sites. The need for a high volume of cases dictated that large urban centers be chosen as primary sites, and the need for geographical diversity demanded that cases from surrounding rural sites be systematically added to the study cohort, when possible.³

The four sites selected for inclusion in the study are:

Alameda County (Oakland), California
Dade County (Miami), Florida
Jackson County (Kansas City), Missouri
Four counties in the state of Michigan

Macomb County
Muskegon County
Ottawa County
Wayne County

C. Case Reader Training

In order to collect necessary data from each site, case reading teams of three people were selected and trained: one case reader for each risk assessment model. Every attempt was made to recruit prospective readers who had child welfare experience and/or education.

Once the teams had been identified, the 12 case readers gathered for a three-day intensive training session. One member of each team was thoroughly trained by an expert to complete one risk assessment model. In addition, all training sessions included inter-rater reliability testing to ensure that case readers understood the system thoroughly.

In addition to the extensive training in the risk assessment models, case readers participated in a half-day training session on cultural sensitivity and awareness. This interactive session assisted readers in recognizing and eliminating personal bias based on race or ethnicity from their work and motivated them to base their responses on facts provided in the case files.

D. Reliability Case Sample Selection

Upon returning to their home site, the case readers commenced the case reading/data collection process. One reader first summarized case information on a case survey document and then each reader completed his/her risk assessment instrument, based on both the summary document and the file contents. From these initial readings, 20 cases from each site were selected to be included in the inter-rater reliability study, for a total of 80 cases.

Once the reliability case samples were identified, copies of the case files were stripped of identifying information and sent to the case reading teams at the other three sites. Each team member read each case and completed his or her respective risk assessment instrument, producing four independent ratings for each of 80 reliability study cases.

III. RESULTS OF THE RELIABILITY STUDY

A. Reliability Measures

Two measures of inter-rater reliability were used in this analysis. The first was simple percent agreement among all raters. The second was Cohen's Kappa. This, as explained earlier, is simply a measure of percent agreement between each pair of raters adjusted for chance. For example, when risk ratings are limited to three choices, even random assignment to risk levels should result in 33% agreement. A positive kappa represents the level of agreement obtained beyond chance. The following explanation is offered by Rossi, Schuerman, and Budde (1996):

The expected agreement among pairs of persons is determined by the "marginal distributions" of each, that is, the percentages of all the cases that they put in various categories. For example, if two experts (or workers) both say that 90% of the cases should be placed, they will have a higher expected agreement than two raters who each thought that only 50% of the cases should be placed. The adjusted measure of agreement is called Cohen's Kappa, sometimes referred to below as Cohen's K or simply K (or kappa).¹ It can range from -1 to +1, with 0 indicating actual agreement equal to expected agreement. Negative kappas indicate that the degree of actual agreement is less than expected by chance and that two sets of judgements disagree more than can be expected by chance.

In this study, the highest degree of overall inter-rater reliability would be reached in instances where all four readers reach the same conclusion. However, because reaching 100% agreement based on information contained in case files may be too high an expectation, 75% agreement was also set as an acceptable level of inter-rater reliability. The 75% agreement rate is achieved when three out of four raters independently arrive at the same conclusion. Both levels are presented in the discussion of results.

Each risk assessment system categorizes cases into an overall risk level. However, the number of risk categories varies among these systems as follows: California has three, Michigan has four, and Washington has six levels of final risk classification.

Because fewer risk levels are designated in the California system it could be expected, everything else being equal, that its reliability would be the highest. Therefore, it was necessary to combine risk levels designated by the Michigan and Washington systems before comparing results. Michigan's high and very high risk levels were combined under the designation "high risk." Washington's six levels were combined as follows:

No Risk & Low Risk	=	Low
Moderately Low & Moderate	=	Moderate

Moderately High & High = High

Although results using all levels of risk for the Michigan and Washington systems are reported, all comparisons among systems are based on three risk level designations: low, moderate, and high.

B. Comparisons of Rater Agreement

Variance among raters in the risk level assigned to cases were evident in each of the systems examined in the study. However, the level of agreement attained was significantly higher for the Michigan system than for either the California model or the Washington system. Figure 1 presents the percentage of cases for which all raters assigned the same risk level. While case readers using the Michigan system attained 100% agreement on 46 of 80 cases, all raters agreed on only 13 of 80 cases and 11 of 80 cases respectively when the California and Washington models were used.

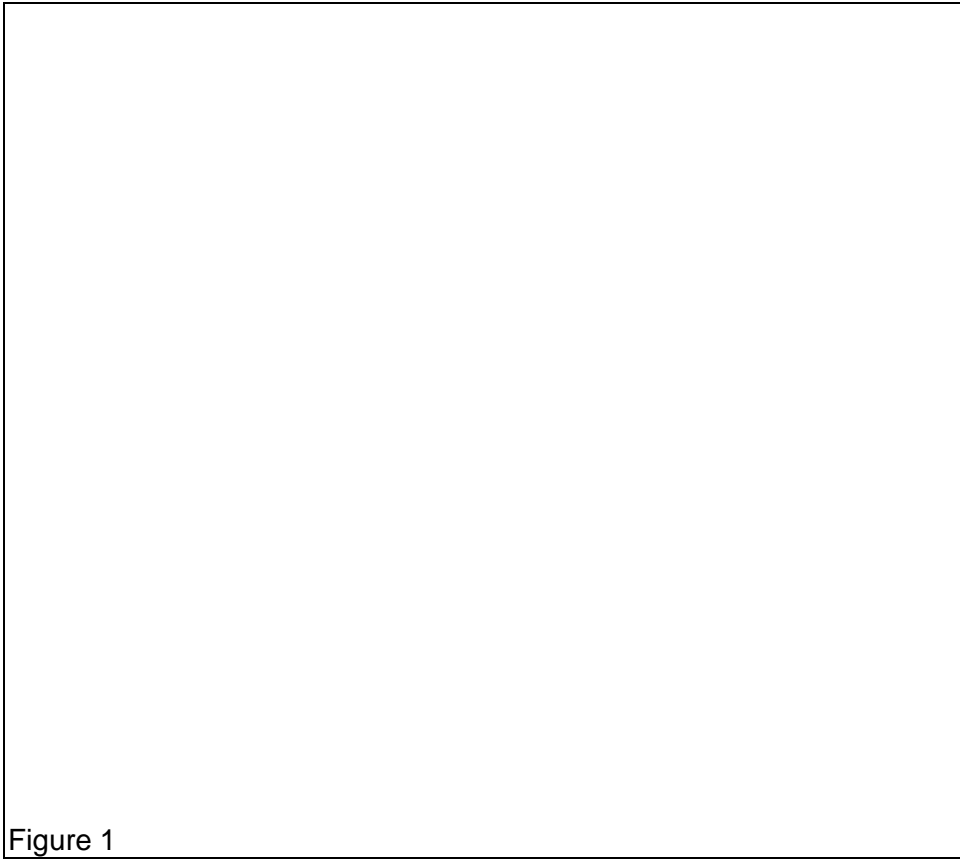


Figure 1

The very low reliability levels found for the Washington and California systems indicate that given the difficulties inherent in reliability studies (and the difficulties encountered in case readings), 100% agreement may be too high a criterion for measuring reliability. Hence, a second, lower standard was applied and the results from each model were again compared. In essence, this analysis asks: In what proportion of cases did three of four readers agree on the risk level assigned.¹

As Figure 2 illustrates, when 75% agreement is used as a reliability threshold, reliability increases significantly. In 85% of all cases rated using the Michigan system, at least three of the four raters scored cases at the same risk level. Three or more raters agreed on 51% of the Washington system ratings and 45% of California system risk designations.

Figure 2

When analysis of the Michigan and Washington systems is expanded to incorporate all risk levels used in the actual application of each system, the reliability of the Washington system shrinks to a marginal level. The 75% agreement threshold was achieved for only 11 of 80 cases (14%). The Michigan system contains only four risk designations (rather than the six risk categories in the Washington system) and consequently adding a single risk level to the analysis made little difference in the reliability levels attained: Overall, three or more raters assigned the same risk level to 65 of 80 (80%) cases in the study.

The Washington results are instructive: Attempts to be too precise with risk designations can be extremely problematic. In actual practice, however, if designations of “low risk” or “no risk” lead to similar decisions and actions, then the impact of such low levels of inter-rater reliability will be minimal. Even so, this begs the question “why bother?” If precise designations can not be made reliably and differences between certain levels have minimal impact on actions taken, it makes little sense to introduce such specificity into the system.

In addition to comparing “percent agreement” among raters, Cohen’s Kappa was computed for each set of raters. The overall kappa was computed as the median value for all sets of raters for each model. The computed kappas indicate reliability is above chance for all systems. However, the difference in kappas computed for Michigan compared to California or Washington was substantial. As Figure 3 indicates, the overall kappa computed for the California model was .184, Michigan was .562, and the Washington model kappa was .180.



Although there is no definitive kappa threshold that designates an acceptable level of reliability, kappa’s below .3 generally indicate very weak reliability. While researchers vary on what is considered an adequate kappa, (depending upon the types of questions posed, number of potential responses, etc.), a kappa above .5 to .6 is generally deemed acceptable. In effect, these results indicate that workers assessing the same family are much more likely to assign a different risk level using the California or Washington systems.

C. Discussion of Findings

There are three possible explanations for the low levels of reliability attained for the California and Washington models. These are:

The raters received insufficient training and/or were not equipped (educationally or experimentally) to accurately complete the risk assessment instruments (the lack of reliability was, in effect, a rater problem);

The case files used in the test did not contain data needed to accurately complete the California and Washington instruments (low reliability was due to a lack of needed data); or

The manner in which these instruments analyze factors and categorize families into risk levels renders these systems inherently unreliable (lack of reliability is a systems problem).

In essence, the lack of reliability could be due to factors outside of the system or it could be due to problems inherent to each system. Although it is difficult (at best) to determine with total confidence the source of the problem, it is possible to gain some insight as to what caused the lack of reliability in the two models.

First, by comparing sets of raters, outliers or extreme values can be identified and removed from the analysis. For example, when comparing scores on the Michigan instrument, there were 21 instances in which one of the four raters was outside the range established as the criterion for reliability. Sixteen of the 21 deviations can be attributed to a single rater, indicating the problem may be more with the rater than the risk assessment scale.

In similar fashion, a means analysis of raters was conducted for the California and Washington models. While no pattern as clear as that noted for the Michigan system emerged, one rater systematically scored cases differently than other raters for each of these systems. Therefore, the rater with the lowest level of agreement in each system was dropped and overall risk ratings for the three remaining raters were compared.

As Figure 4 illustrates, inter-rater reliability improved as expected. However, at this level of analysis the 100% agreement threshold takes on added importance. Because a rater was

dropped from the analysis of each system, the lower threshold becomes agreement by two of three raters (67%) rather than agreement by three of four raters (75%). When “chance” is added to the equation (using Cohen’s Kappa), reliability for the Washington and California systems remains below acceptable levels. Median kappas were .211 for the California model, .245 for the Washington system and .635 for the Michigan risk assessment.



The marginal degree of improvement gained when the “outlying” rater was dropped from the analysis, indicates that the lack of reliability is, in all likelihood, not due to a lack of expertise among the case readers. In fact, a case could be made that excluding an “outlying” rater may distort what actually occurs in CPS practice. In actual practice, risk instruments will be completed by a wide variety of CPS staff with different backgrounds and different levels of experience. In many jurisdictions CPS workers are not required to have degrees in social work and with the level of turnover often experienced by child protection agencies, the “case readers”

utilized in a study of this nature may, in fact, represent a best-case scenario. Given the careful selection process used to hire case readers, coupled with the training provided, it seem doubtful that CPS field staff, in general, represent a higher quality of raters.

The second issue investigated was whether the low level of reliability found in the California and Washington systems could be attributed to a lack of data in case files. When there was not enough data available to rate a particular risk item, case readers were instructed to enter "I" (insufficient information) in the appropriate column on each risk form. To ascertain the degree to which "missing" data affected reliability, cases with missing information were systematically included and excluded during a series of analyses. The overall impact of excluding cases was minimal. Figure 5 provides an example of how missing data had little effect on the degree of reliability attained. In this phase of the analysis, the 16 cases (20% of the sample) with the most data missing were dropped from the analysis. Results for the remaining 64 cases were not significantly different than those obtained for the entire sample.



Figure 5

With relatively clear indications that the reliability problems encountered with the California and Washington systems were not due to lack of data or problems with case readers, the analysis focused on the basic design of these systems. Two potential explanations are presented below.

First, the overall risk rating assigned to each case is not directly related to individual risk elements. The Washington model, for example, contains a number of separate elements on which children and/or caretakers are rated. However, there is no clearly defined relationship between these factors and the overall risk level assigned. (The design of the California model is similar.) Each element could assume a different level of importance for each rater (and for each case for that matter). The level of “structure” provided by these elements may be less than is required to attain an acceptable level of reliability. Even if individual factors could be reliably rated, the differential weighting of each factor noted by Blenkner in 1954 (described earlier in this report) is a likely contributor to the lack of reliability in these models.

Second, there may also be a problem with how each factor is rated; in effect, a “dimensions” issue may exist. For instance, both the California and Washington systems rate the history of caretakers as prior “victims of abuse/neglect.” However, workers do not simply determine if the caretakers were maltreated as children, but are instead required to assign a current risk level to this (and every) factor. Hence, the level of consistency that might be obtained by simply answering the question may be jeopardized by adding the dimension of “current” risk to the rating of individual items. Obviously, many adults who were abused or neglected as children do not mistreat their own children. While individual factors may incrementally contribute to a caretaker’s potential to abuse or neglect his/her children, assigning risk levels to each item distorts the real relationship between that item and subsequent caretaker behavior.

When assessment systems are applied in field settings (child protection, probation, domestic violence, etc.) where they are used by a wide variety of staff with many responsibilities and limited time available, reliability is highly dependent on the simplicity of the instrument(s),

the degree of structure imposed, and the overall clarity of system design (Baird, 1991). Improvements in all three of these areas may be required to obtain a greater degree of reliability in the Washington and California models.

IV. CONCLUSIONS AND IMPLICATIONS

A. Conclusions

In sum, this study demonstrates the following:

First, while none of the systems approached 100% inter-rater reliability, raters employing the Michigan system made consistent risk estimates for a high percentage of the cases they assessed and inter-rater reliability for the Michigan model was much higher than that achieved by the other systems.

Second, the level of inter-rater reliability attained by the California and Washington risk assessment models were well below what could be considered adequate. To the extent these systems are representative of other “consensus-based” or “expert” systems, the problem noted here may apply to other systems as well.

Finally, it appears that the reliability of both the California and Washington systems could be enhanced by:

simply requiring answers to questions raised, rather than assigning a risk level to each item;

adding structure to the manner in which the overall risk level is determined

B. Implications

Since the reliability of risk assessment systems can have a profound affect on the efficacy of decision making, it is important to view the implications of this study in the context of the current state of Child Protective Services nationwide. Child Protective Services in the United States are generally in a state of disarray, as evidenced by lawsuits in over 30 states, the recent assault on family preservation programs, the tragedies that are frequently the subjects of

media exposés and the push toward new strategies for dealing more effectively with the burgeoning problem of abuse and neglect.

Much of the current debate surrounding child protection across the country has focused on the choice between increased use of foster care and the family preservation movement. Over the last decade, family preservation has been embraced by child welfare agencies throughout the country. Now, however, professionals question if the emphasis on keeping families intact has left too many children in high risk situations, resulting in increases in child abuse and neglect, serious injuries, and even child deaths. When tragedies like the Elisa Izquierdo death in New York City occur, it is hard to defend a system that leaves children in such desperate circumstances. But social service administrators know that foster care is no panacea either. Funds for foster family recruitment, training, licensing, and monitoring of foster parents are rarely adequate to the task. As a result, children often move from foster home to foster home, trapped in a perpetual state of transition as they slip through legal and bureaucratic cracks in the system.

The debate over family preservation and foster care or other resources, while clearly useful, misses the bigger point. The issue facing child protection really centers on decision making: studies have clearly demonstrated that decisions regarding the safety of children vary significantly from worker to worker (Chapin Hall, 1996). As a consequence, actions taken are often inappropriate and sometimes completely indefensible. In far too many agencies, child protection can best be described as a loosely affiliated group of workers asked to make extremely difficult decisions with very little guidance or training. Their actions are rarely monitored, data related to program effectiveness are not available, and computer technology is virtually nonexistent. As a result, case decisions are based on the expertise, education, intuition, and biases of individual workers. Even the most experienced and talented social workers find it difficult to deal with increasing levels of poverty, substance abuse, and despair. As Time Magazine noted in its December 11, 1995 edition:

Dispatched into unfamiliar, often dangerous surroundings, they are expected to make instant predictions about tomorrow, based largely on a sixth sense about the data their five senses gather today. Certainly many people outrank them in the child welfare hierarchy, yet their views carry the greatest weight. Only they “walk up the drug-filled staircase, sit on the dirty couch, and talk to the teenage mother,” says Marc Parent, who spent four years as a caseworker in New York City. As the Elisa Izquierdo case demonstrates, “if you get a caseworker who goes to somebody’s home and says it’s fine, then it’s fine,” notes Parent. “That’s how important their voice is.”

This is at the heart of the child protection crisis in America and it is naive to think that increases in resources or changes in agency philosophy will turn the tide. The simple fact is that some children are better off with their parents and some children need foster care. Some families require intensive services and monitoring; others do not. Until decisions are based on the best available information consistently applied to all cases, families that could be saved will be split up and children that should be removed will remain at home and again be abused or neglected.

Until valid, reliable decision support systems are fully utilized, debate over which programs and strategies work and which do not is fruitless. Designing programs to solve the CPS problem without adequately addressing the issue of decision making is analogous to building a house on a weak foundation. Regardless of the quality of the carpentry, the house will eventually fall apart.

REFERENCES

- Allen, T.C., "Field Testing of the Child-At-Risk Field System." Presentation for the American Public Welfare Association, Second National Roundtable on CPS Risk Assessment; Denver, CO: Action for Child Protection, 1988.
- Baird, S.C., "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, 1988.
- Baird, S.C., "Validating Risk Assessment Instruments Used in Community Corrections." National Council on Crime and Delinquency, Madison, WI, January 1991.
- Blenkner, M., "Predictive Factors in the Initial Interview in Family Casework." Social Science Review, 28, (1954).
- Cicchinelli, F. (ed.), Proceedings from the Symposium on Risk Assessment in Child Protective Services, National Center on Child Abuse and Neglect, Washington, D.C., 1991.
- Clear, T., "Statistical Prediction in Corrections." A monograph (1988).
- Dawes, R.M., Faust, D., and Meehl, P.E., "Clinical Versus Actuarial Judgment." Science, 243, (1989), 1668-1674.
- Doueck, H., English, D., DePanfilis D., and Moote, G., "Decision Making in Child Protective Services: A Comparison of Selected Risk Assessment Systems." Child Welfare, 72(5), 441-452, 1993.
- Doueck, J., Levine, M., Bronson, D., "Risk Assessment in Child Protective Services: An Evaluation of the Child at Risk Field System," Journal of Interpersonal Violence, December, 1993.
- Johnson, W., and L'Esperance, J., "Predicting the Recurrence of Child Abuse." Social Work Research and Abstracts, 20 (2), 21-26, 1984.
- Keller, R.A., Cicchinelli, L.F., Gardner, D., "Comparative Analysis of Risk Assessment Models: Phase I Report." Applied Research Associates, Denver, CO (1988).
- Marks, J., McDonald, T., Bessey, W., Palmer, M., "Risk Factors Assessed by Instrument-based Models: A Review of the Literature." Risk Assessment in Child Protective Services, National Child Welfare Resource Center for Management and Administration, (1989).
- Marks, J., and McDonald, T., "Predicting Recurrence of Child Maltreatment." Risk Assessment in Child Protective Services, National Child Welfare Resource Center for Management and Administration, (1989).
- Meehl, P., "Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence." Minneapolis, University of Minnesota Press, (1954).
- Murphy-Berman, V., "A Conceptual Framework for Thinking About Risk Assessment and Case Management in Child Protective Services." Child Abuse and Neglect, Vol. 18, No. 2, 1994.
- Rossi, P., Schuerman, J., Budde, S., "Understanding Child Maltreatment Decisions and Those Who Make Them." Chapin Hall Center for Children, University of Chicago, June 1996.

Sawyer, J., "Measurement and Prediction, Clinical, and Statistical." Psychological Bulletin, Vol. 66 No. 3, 178-200, 1966.

Tatara, T., "An Overview of Current Practices in CPS Risk Assessment and Family Systems Assessment in Public Child Welfare," Summary of Highlights of the National Roundtable on CPS Risk Assessment and Family Systems Assessment, American Public Welfare Association, Washington, D.C., 1987.

Tatara, T., "A Survey of States on CPS Risk Assessment Practice: Preliminary Findings," Presented at the Tenth National Roundtable on CPS Risk Assessment, American Public Welfare Association, Washington, D.C., 1996.

Wald, M.S., Woolverton, M., "Risk Assessment: The Emperor's New Clothes?" Child Welfare, Vol. LXIX No. 6, 483-511, November-December 1990.

ENDNOTES

¹ Even when vignettes are carefully constructed, they may not contain all information deemed necessary for decisions by workers. In a recent study of case decision making conducted by Chapin Hall (Rossi, Schuerman, and Budde, June 1996), about one-third of all ratings of the adequacy of information provided in the vignettes were deemed “somewhat adequate” to “inadequate” by “experts” in the study.

² For example, if one rater always rates a case one level higher than another rater, a correlation coefficient of 1 is obtained, yet they never actually agree on a risk level.

³ Despite the objective of including cases from rural areas, the complexities encountered in the data collection phase limited this effort. It should be noted that the vast majority of cases in the study come from urban settings.

¹ The formula for Cohen’s kappa is: $K = (\text{actual agreement} - \text{expected agreement}) / (1 - \text{expected agreement})$.

¹ In essence, this analysis accounts to a degree for problems with reliability that may be outside the system, i.e., lack of data, difficulty in finding information in files, and problems with raters.