Risk Assessment in Child Protective Services: Consensus and Actuarial Model Reliability

Christopher Baird, Dennis Wagner, Theresa Healy, and Kristen Johnson

Three widely used child protective service risk assessment models (two consensus based, one actuarial) were examined to determine their reliability. Although no system approached 100% interrater reliability, raters employing the actuarial model made consistent estimates of risk for a high percentage of the cases they assessed, and interrater reliability for the actuarial model was much higher than that of the other systems.

Christopher Baird, M.A., is Senior Vice President; Dennis Wagner, Ph.D., is Director of Research; Theresa Healy, B.B.A. is Senior Research Associate; and Kristen Johnson, M.A., is Research Associate, National Council on Crime and Delinquency, Children's Research Center, Madison, WI. *This article reports on the first phase of a three-year study funded by the National Center on Child Abuse and Neglect (grant # 90-CA-1550), "Comparative Study of the Use and Effectiveness of Different Risk-Assessment Models in CPS Decision Making," to determine the relative reliability and validity of three approaches to risk assessment in child protective services.*

0009–4021/99/060723–26 $3.00

The decisions that child protective service (CPS) workers make at the conclusion of a child abuse or neglect investigation are critical to the protection of children. At that point, the worker must decide whether to: (1) close the case, (2) open the case for protective service intervention or intensive in-home family preservation services, or (3) remove the child to out-of-home care. Because this decision has important consequences for children, their families, and protective service agencies, it must be made as consistently and accurately as possible.

Frontline CPS workers must base their decisions on the best interest of the child, a determination that is largely dependent on the worker's estimation of the risk of future harm. If conclusions regarding risk in 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 out-of-home care. Mistakes can have enormous consequences, ranging from unnecessary expenditures to emotional upheaval and trauma, to serious injury to and even the death of a child.

Clearly, decisionmaking in human services is difficult—personal relationships in families served are dynamic and difficult to assess. Those charged with protecting children—CPS investigative staff and managers—represent a wide spectrum of educational backgrounds and 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 decisionmaking.

CPS Risk-Assessment Systems and the NCCAN Study

Risk-assessment systems are simply "formalized methods that provide a uniform structure and criteria for determining risk"

[Keller et al. 1988: 2]. CPS risk-assessment systems have developed to help workers accurately estimate the future risk of abuse/neglect and thus make better service decisions for families. The expectation is that such systems will increase the reliability and accuracy of CPS worker decisionmaking.

The use of risk-assessment systems, although widespread, is relatively recent. In 1996, the American Public Welfare Association (APWA) conducted a survey of 54 states, territories, and large county child welfare agencies to determine their use of 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, 26 first implemented their risk assessment after 1987 [Tatara 1996].

Although risk-assessment models continue to be developed, experts have 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* are those in which workers assess specific client characteristics identified by the consensus judgement of experts and then exercise their own clinical judgment about the risk of future abuse or neglect.
2. *Actuarial 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 that workers score to identify families at low, medium, or high risk.

This article describes the results from the first phase of a comparative study funded by the National Center on Child Abuse and Neglect to compare the use and effectiveness of different risk-assessment models in CPS. This phase addresses the question, "How reliable are these systems, that is, do different workers assign the *same* risk level to the same family at investigations?"

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. Although the volume of research on these systems is considered by some to be "quite formidable" [Doueck et al. 1993], these systems have rarely undergone rigorous evaluations.

Historically, caseworkers have used the case study method, relying almost entirely upon clinical experience, intuition, and interviewing skills to assess the future risk of abuse or neglect to a child. In many states, such clinical assessments are structured by instruments or systems that identify specific case characteristics, often 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 in which they are used [Wald & Woolverton 1990]. In other words, these instruments may help organize the caseworker's clinical assessment of risk, but they are not based on research specific to the jurisdiction [Marks et al. 1989].

An early study illustrating a critical problem for clinical judgment was conducted by Blenkner [1954], who found that three expert social workers did poorly in predicting case outcomes. Meehl [1954: 108] 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." Blenkner's work illustrated why many observers believe experts perform poorly on predictive tasks: 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 poorly when asked to predict case outcomes.

Rossi, Schuerman, and 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 should have a child(ren) placed in out-of-home 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 decisionmaking, even among the CPS experts, demonstrating that, in human service decisionmaking, there is considerable disparity in the way cases are handled.

Actuarial Systems

The evidence now available from actuarial studies of child abuse and 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 CPS. In summarizing 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 NCCD [Baird 1988]. Since the 1989 Marks and McDonald publication, however, NCCD has conducted additional actuarial research in Oklahoma, Michigan, Rhode Island, and Wisconsin. A large body of research evidence in experimental psychology and corrections (e.g., Dawes et al. [1989]; Meehl [1954]; Sawyer [1966]) supports the view that actuarial instruments can estimate future behavior more accurately than an individual decisionmaker unaided by actuarial information (even decisionmakers who have had extensive clinical training).

The present study takes the next step by directly comparing the reliability of decisions made with actuarial and consensus-based instruments.

Measuring Reliability

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, including record reviews; personal contacts with the client; collateral contacts with law enforcement, school, medical, and 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 decisionmaking 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) creating 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 strengths and weaknesses. Developing case vignettes gives the researchers greater control over the data provided to readers. This helps 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]

Reading case files presents the opposite problem: although more representative of actual case practice, case files may not contain all the data necessary to adequately assess cases. 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 those policies.

Neither approach will produce all of the data available to an actual decisionmaker. Information gained through personal contact and observation cannot be fully replicated, even when a caseworker's observations are discussed in the official record. There is, however, no evidence that introducing these data into the assessment process increases 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 interrater reliability [Clear 1988].

Several prior studies of the reliability of various risk-assessment systems have been conducted. Both the methods used to analyze results and the interpretation of findings, however, raise serious questions regarding the value of these 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 et al. [1993], used three videotaped case scenarios or three case vignettes. Although the number of raters was substantial, ranging from 30 to 214, 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 videotaped case vignettes.

The statistical methods chosen to rate reliability in a number of prior studies raise additional concerns. Some measures 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. Theoretically, two rat-

ers could 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.

Due to the limitations of prior studies, little is known about the reliability of various risk-assessment systems used by child protection agencies. A recent study of case decisionmaking, however, presents one model for assessing reliability. The Chapin Hall Study [Rossi et al. 1996] combines an adequate number of study cases (70) with an appropriate measure of interrater reliability—percent agreement subsequently adjusted for "chance," using Cohen's kappa. This simple and straightforward approach 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).

Method

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 [Berkowitz 1991] are the Washington Risk Assessment Matrix (WRAM), a risk-assessment system developed by practitioners in Washington State [Washington State Department of Social and Health Services, Division of Children and Family Services 1987], and the California Family Assessment Factor Analysis (CFAFA), a derivative of the Illinois Child Abuse and Neglect Tracking System (CANTS) system [California State University 1987]. Because it is the most widely used actuarial-based approach, the Michigan Structured Decision Making System's Family Risk Assessment of Abuse and Neglect (FRAAN) was selected for this study [NCCD/Children's Research Center 1995]. It is a research-based tool constructed during a study of 2,000 Michi-

gan families and recently revalidated on a cohort of 1,000 families [Baird et al. 1995]. This system is also used (or being implemented) in Georgia, Indiana, Minnesota, and Cleveland, Ohio.

The Washington Risk Assessment Matrix (WRAM). According to the Washington State Department of Social and Health Services "Risk Factor Matrix Guide," WRAM 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" [Washington State DSHS 1995: Appendix H: 1]. WRAM 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" [Washington State DSHS 1995: Appendix H: 1].

WRAM groups risk factors by child characteristics (five items), severity of child abuse/neglect (nine items), chronicity or recurrence of episodes of child abuse/neglect (one item), caregiver characteristics for primary and secondary caregivers (11 items each), caregiver-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, 2 for moderately low, 3 for moderate, 4 for moderately high, and 5 for high). The ratings system also includes 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, 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.

The California Family Assessment Factor Analysis (CFAFA). The second risk-assessment model, CFAFA, represents another consensus-based approach to risk assessment. The California Risk Assessment Curriculum for Child Welfare Services Resource Handbook [1995: 1] 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.

CFAFA recognizes five types of factors: precipitating incident (four factors), child assessment (five factors), caregiver 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.

The Michigan Family Risk Assessment of Abuse and Neglect (FRAAN). The third risk-assessment system included in this study is FRAAN, which is based on the statistical relationship between behavior and case characteristics and subsequent abuse and neglect. FRAAN 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, FRAAN 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.

FRAAN 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 that comprises 13 factors. In practice, caseworkers 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. FRAAN is used in case planning, in assisting workers in establishing priorities, and in referring cases to needed services.

Site Selection

The site selection process sought 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, Latinos, and American Indians; 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 were Alameda County (Oakland), California; Dade County (Miami), Florida; Jackson County (Kansas City), Missouri; and four counties in the state of Michigan (Macomb County, Muskegon County, Ottawa County, and Wayne County).

Case Reader Training

To collect the 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 interrater reliability testing to ensure that case readers understood the system thoroughly.

In addition to receiving extensive training in 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 the facts provided in the case files.

Case Sample Selection

Upon returning to their home sites, 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 interrater reliability study, for a total of 80 cases.[4]

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.

Reliability Measures

Two measures of interrater reliability were used in this analysis. The first was percent agreement among all raters. The second was Cohen's kappa, which measures the 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 Cohen's kappa represents the level of agreement obtained beyond chance:

> 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... as Cohen's K or simply K (or kappa).[5] 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. [Rossi et al. 1996: 16–17]

In this study, the highest degree of overall interrater reliability would be reached in instances where all four readers reach the same conclusion. Reaching 100% agreement based on information contained in case files may be too high an expectation, however, so a 75% agreement rate was also set as an acceptable level of interrater 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. The number of final level of risk categories varies among these systems, however: CFAFA has three, FRAAN has four, and WRAM has six.

Because fewer risk levels are designated in CFAFA, it could be expected that, everything else being equal, its reliability would be the highest. Therefore, risk levels designated by FRAAN and WRAM were combined before comparing their results to those

of CFAFA. FRAAN's high and very high risk levels were combined under the designation "high risk." WRAM' six levels were combined into low risk (formerly no risk and low); moderate risk (formerly moderately low and moderate); and high risk (formerly moderately high and high).

Although results using all levels of risk for FRAAN and WRAM are reported, all comparisons *among* systems are based on three risk-level designations: low, moderate, and high.

Results

Comparisons of Rater Agreement

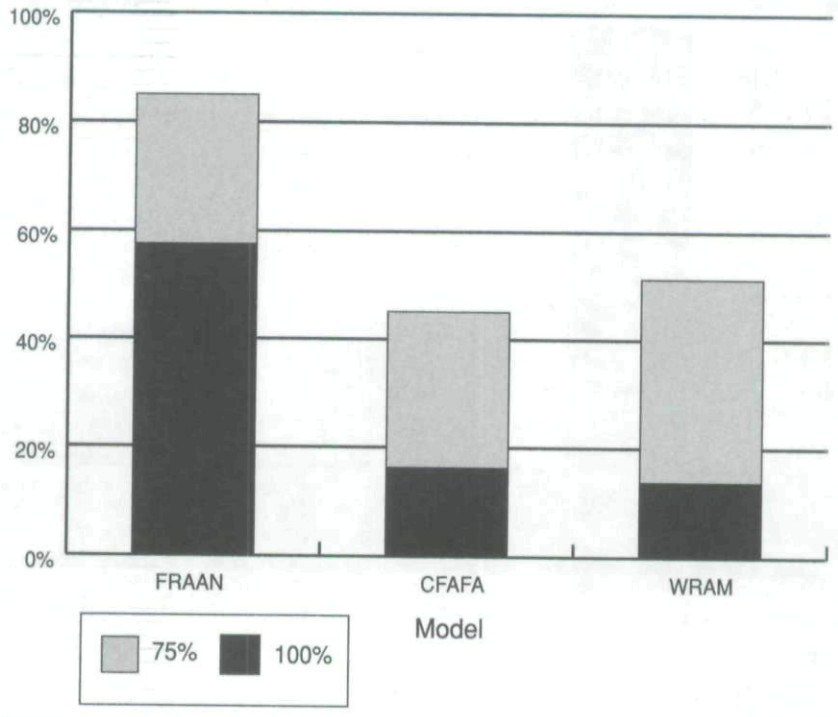
Variance among raters in the risk level assigned to cases was evident in each of the systems examined in the study. The level of agreement attained, however, was significantly higher for FRAAN than for either CFAFA or WRAM. Although case readers using FRAAN attained 100% agreement on 46 of 80 (57.5%) cases, raters using CFAFA agreed on only 13 of 80 (16.3%) cases, and those using WRAM agreed on only 11 of 80 (13.8%) cases (see figure 1).

The very low reliability levels found for CFAFA and WRAM 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 second analysis asks: "In what proportion of cases did three of four readers agree on the risk level assigned?"[6]

As figure 1 illustrates, when 75% agreement is used as a threshold, reliability increases significantly. In 85% of all cases rated using FRAAN, at least three of the four raters scored cases at the same risk level. Three or more raters agreed on 51.3% of the WRAM ratings and 45.1% of the CFAFA risk designations.

When analysis of FRAAN and WRAM is expanded to incorporate all risk levels used in the actual application of each sys-

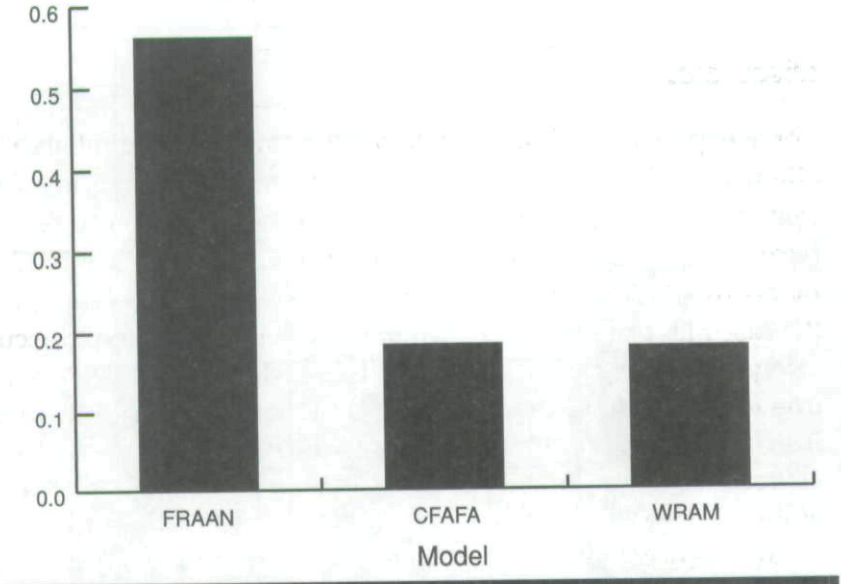
FIGURE 1
Percentage of Cases with 75% or 100% Agreement



tem, the reliability of WRAM shrinks to a marginal level. The 75% agreement threshold was achieved for only 11 of 80 cases (13.8%). FRAAN contains only four risk designations; adding a single risk level to the analysis made little difference in the reliability levels attained. Using FRAAN's four risk categories, three or more raters assigned the same risk level to 65 of 80 cases (80%) in the study.

The WRAM results are instructive: Attempts to be too precise with risk designations can be extremely problematic. In actual practice, if designations of "low risk" or "no risk" lead to similar decisions and actions, then the impact of such low levels of

FIGURE 2
Cohen's Kappa Among Raters for Overall Risk



interrater reliability will be minimal. If decisions and actions are not different, however, there is no reason for multiple risk categories that have demonstrably low levels of reliability.

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. The difference in kappas computed for FRAAN compared to CFAFA or WRAM are substantial, however. As figure 2 indicates, the overall kappa computed for CFAFA was .184; it was .562 for FRAAN, and .180 for WRAM.

There is no definitive kappa threshold that designates an acceptable level of reliability, but kappas below .3 generally indicate very weak reliability. Although researchers vary on what is considered adequate, 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 differing risk levels when using CFAFA or WRAM than when using FRAAN.

Discussion

Three explanations are possible for the low levels of reliability attained for CFAFA and WRAM: (1) 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); (2) the case files used in the test did not contain data needed to accurately complete the CFAFA and WRAM (the low reliability was due to a lack of needed data); or (3) the manner in which these instruments analyze factors and categorize families into risk levels renders these systems inherently unreliable (the lack of reliability is a systems problem).

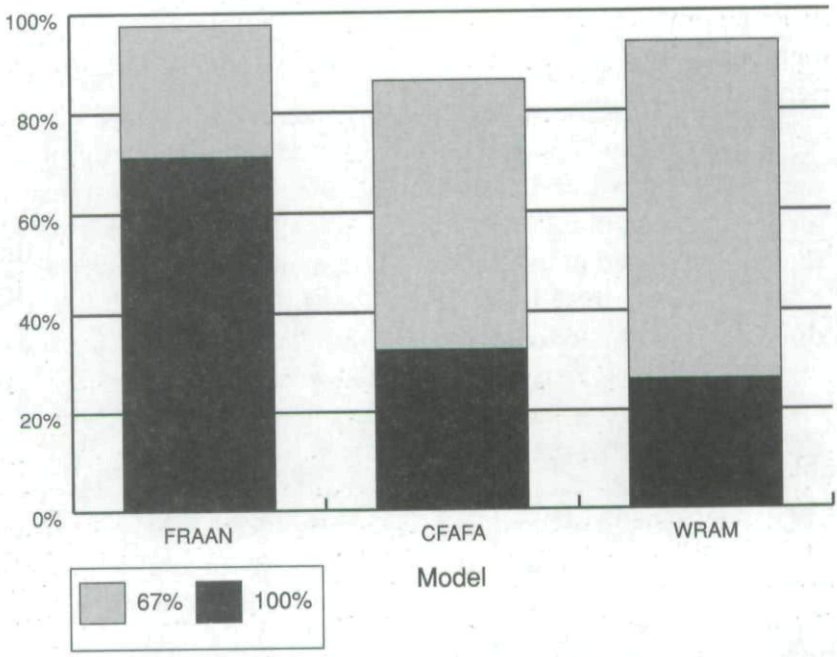
In essence, the lack of reliability could be due to factors outside of the system or it could be due to problems inherent to each system. Although it is difficult to determine with total confidence the source of the problem, it is possible to gain some insight as to what caused the lack of reliability 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 FRAAN, 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 with the risk-assessment scale.

In similar fashion, a means analysis of raters was conducted for CFAFA and WRAM. While no pattern as clear as that noted for FRAAN 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

FIGURE 3

Percentage of Cases with 67% or 100% Agreement (excludes three outlying raters)



dropped and overall risk ratings for the three remaining raters were compared.

As figure 3 illustrates, interrater reliability improved as expected. At this level of analysis, however, 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 WRAM and CFAFA remains below acceptable levels. Median kappas were .211 for CFAFA, .245 for WRAM, and .635 for FRAAN.

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, it could be argued that excluding an "outlying" rater may distort what actually occurs in the field. In actual practice, risk instruments will be completed by CPS staff members with different backgrounds and varying levels of experience. In many jurisdictions, CPS workers are not required to have degrees in social work and with the turnover often experienced by child protection agencies, the case readers 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.

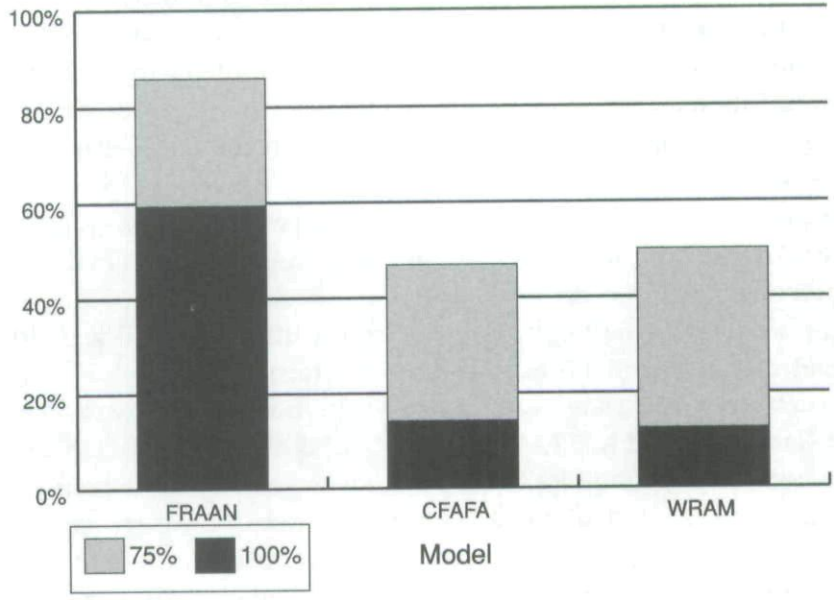
The second issue investigated was whether the low level of reliability found in CFAFA and WRAM could be attributed to a lack of data in case files. When there was not enough data available to rate a particular risk item, case readers were instructed to enter "I" (insufficient information) in the appropriate 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 4 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.

With relatively clear indications that the reliability problems encountered with CFAFA and WRAM 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 di-

FIGURE 4

Percentage of Cases with 75% or 100% Agreement (excludes 20% of sample cases with most missing data)



rectly related to individual risk elements. WRAM, for example, contains a number of separate elements on which children and/or caregivers are rated. There is, however, no clearly defined relationship between these factors and the overall risk level assigned. CFAFA's design is similar. Each element could assume a different level of importance for each rater, and for each case. 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 [1954] is a likely contributor to the lack of reliability in these models.

Second, there may be a problem with *how* each factor is rated. For instance, both CFAFA and WRAM rate the history of care-

givers as prior "victims of abuse/neglect." Workers, however, do not simply determine if the caregivers 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. Although individual factors may incrementally contribute to caregivers' potential to abuse or neglect their children, assigning risk levels to each item distorts the real relationship between that item and subsequent caregiver behavior.

When assessment systems are applied in field settings (child protection, probation, domestic violence, etc.) where they are used by a variety of staff with many responsibilities and limited time, 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 WRAM and CFAFA.

Conclusions and Implications

Conclusions

Although none of the systems approached 100% interrater reliability, raters employing FRAAN made consistent risk estimates for a high percentage of the cases they assessed and interrater reliability for FRAAN was much higher than that achieved by the other systems. In addition, the levels of interrater reliability attained by CFAFA and WRAM 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 CFAFA and WRAM could be

enhanced by requiring answers to questions raised, rather than assigning a risk level to each item, and by adding structure to the manner in which the overall risk level is determined.

Implications

Since the reliability of risk-assessment systems can have a profound affect on the efficacy of decisionmaking, this study should be viewed in the context of the current state of child protective services nationwide. Much of the ongoing debate surrounding child protection across the country has focused on the choice between increased use of out-of-home care and the family preservation movement. Over the last decade, family preservation has been embraced by child welfare agencies nationwide. 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 [Schorr 1997]. But social service administrators know that out-of-home care is no panacea either. Funds for foster family recruitment, training, licensing, and monitoring of foster parents are rarely adequate to the task [Traglia et al. 1997]. As a result, children often move from family foster home to family foster home, trapped in a perpetual state of transition as they slip through legal and bureaucratic cracks in the system [Schorr 1997].

The debate over family preservation, out-of-home care, and the use of other resources, while clearly useful, misses the bigger point. The primary issue facing child protection really centers on decisionmaking: studies have clearly demonstrated that decisions regarding the safety of children vary significantly from worker to worker [Rossi et al. 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 [Schorr 1997]. Their actions are rarely monitored, data related to program effec-

tiveness 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 [Rossi et al. 1996]. 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:

> 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." [Smolowe 1995: 40]

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 decisionmaking is analogous to building a house on a weak foundation. Regardless of the quality of the carpentry, the house will eventually fall apart.◆

Notes

1. Even when vignettes are carefully constructed, they may not contain all information deemed necessary for decisions by workers. In a study of case decisionmaking conducted by Chapin Hall [Rossi et al. 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 came from urban settings.

4. Ten substantiated and ten unsubstantiated cases were selected based on the following: the first three physical abuse cases, the first two sexual abuse cases, and the first three neglect cases, followed by two cases of any maltreatment type.

5. The formula for Cohen's kappa is: K = (actual agreement - expected agreement) / (1 - expected agreement).

6. 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.

References

Allen, T. C. (1988). *Field testing of the Child-At-Risk Field System.* Paper presented at the American Public Welfare Association, Second National Roundtable on CPS Risk Assessment, Denver, CO.

Baird, S. C. (1988). Development of risk assessment indices for the Alaska Department of Health and Social Services. In T. Tatara (Ed.), *Validation research in CPS risk assessment: Three recent studies* (Occasional Monograph Series No. 2). Washington, DC: American Public Welfare Association.

Baird, S. C. (1991). *Validating risk assessment instruments used in community corrections.* Madison, WI: National Council on Crime and Delinquency.

Baird, S. C., Wagner, D., Caskey, R., & Neuenfeldt, D. (1995). *Michigan Department of Social Services structured decision making system: An evaluation of its impact on child protection services.* Madison, WI: National Council on Crime and Delinquency, Children's Research Center.

Berkowitz, S. (1991). *Key findings from the state survey component of the Study of High Risk Child Abuse and Neglect Groups.* Rockville, MD: Westat.

Blenkner, M. (1954). Predictive factors in the initial interview in family casework. *Social Science Review, 28,* 65-73.

California Risk Assessment Curriculum for Child Welfare Services Resource Handbook. (1995). Fresno, CA: The Child Welfare Training Project, California State University-Fresno, School of Health and Social Work.

California State University-Fresno, School of Health and Social Work. (1987). *Family Assessment Factor Analysis* (adapted from the state of Illinois *Factor Worksheet*, Illinois Department of Children and Family Services, Springfield, IL, 1985). Fresno, CA: Author.

Cicchinelli, F. (Ed.). (1991). *Proceedings from the Symposium on Risk Assessment in Child Protective Services.* Washington, DC: National Center on Child Abuse and Neglect.

Clear, T. (1988, March). Statistical prediction in corrections. *Research in Corrections*, 1–39.

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

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

Doueck, J., Levine, M., & Bronson, D. (1993, December). Risk assessment in child protective services: An evaluation of the Child at Risk Field System. *Journal of Interpersonal Violence, 8,* 446–447.

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

Keller, R. A., Cicchinelli, L. F., & Gardner, D. (1988). *Comparative analysis of risk assessment models: Phase I Report.* Denver, CO: Applied Research Associates.

Marks, J., McDonald, T., Bessey, W., & Palmer, M. (1989). *Risk assessment in child protective Services: Risk factors assessed by instrument-based models: A review of the literature.* Portland, ME: National Child Welfare Resource Center for Management and Administration.

Marks, J., & McDonald, T. (1989). *Risk assessment in child protective services: Predicting recurrence of child maltreatment.* Portland, ME: National Child Welfare Resource Center for Management and Administration.

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

Murphy-Berman, V. (1994). A conceptual framework for thinking about risk assessment

and case management in child protective services. *Child Abuse and Neglect, 18,* 193-201.

NCCD/Children's Research Center. (1995). *Michigan Assessment of Family Risk.* Madison, WI: Author.

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

Sawyer, J. (1966). Measurement and prediction, clinical, and statistical. *Psychological Bulletin, 66, 3,* 178-200.

Schorr, L. B. (1997). *Common purposes. Strengthening families and neighborhoods to rebuild America.* New York: Anchor Books/Doubleday.

Smolowe, J. (1995, December 11). Making the tough calls, *Time,* 40-45.

Tatara, T. (1987). An overview of current practices in CPS risk assessment and family systems assessment in public child welfare. In *Summary of Highlights of the National Roundtable on CPS Risk Assessment and Family Systems Assessment* (pp. 415–459), Washington, DC: American Public Welfare Association.

Tatara, T. (1996). *A survey of states on CPS risk assessment practice: Preliminary findings.* Paper presented at the American Public Welfare Association Tenth National Roundtable on CPS Risk Assessment, Washington, DC, 1996.

Traglia, J. J., Pecora, P., Paddock, G. B., & Wilson, L. (1997). Outcome-oriented case planning in family foster care. *Families in Society: The Journal of Contemporary Human Services,* 453-462.

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

Washington State Department of Social and Health Services, Division of Children and Family Services. (1995). *Risk Factor Matrix guide.* Seattle, WA: Author.

Washington State Department of Social and Health Services, Division of Children and Family Services. (1987). *Washington Risk Factor Matrix.* Olympia, WA: Author.

(Address requests for a reprint to Christopher Baird, National Council on Crime and Delinquency, Children's Research Center, 426 South Yellowstone Drive, Suite 250, Madison, WI 53719.)