Risk assessment for child protective services: A review of the empirical literature on instrument performance

Peter Lyons, Howard J. Doueck, and John S. Wodarski

The development and use of systematic risk assessment models by child protective services agencies can be seen in part as a response to the dramatic increase in maltreatment referrals during the past two decades. However, despite the increasing popularity of such models, relatively few are empirically based. The authors reviewed the empirical literature on 10 risk assessment models and conclude that although each model contains generally sound psychometric properties, there is still a need for further model development and for process and outcome evaluation of model implementation.

Key words: **child protective services; evaluation; models; risk assessment**

***Peter Lyons, PhD, MEd,** is assistant professor, School of Social Work, University at Buffalo, State University of New York, 359 Baldy Hall, Buffalo, NY 14260-1050; e-mail: plyons@npiec.niagara.ca. **Howard J. Doueck, PhD, ACSW,** is associate professor, and **John S. Wodarski, PhD,** is Janet B. Wattles Research Professor, School of Social Work, University at Buffalo, State University of New York.*

The proliferation of child protection risk assessment models in more than 40 states (Berkowitz, 1991) can be seen in part as a response to the crisis engendered in child protection services (CPS) agencies by the growing number of intakes (National Center on Child Abuse and Neglect, 1988; U.S. Advisory Board on Child Abuse and Neglect, 1990) and a resource-depleted environment (Doueck, English, DePanfilis, & Moote, 1993). Risk assessment also reflects a legislative trend that views CPS intervention as justified "only when a child has suffered (or is likely to suffer) specific types of maltreatment" and that has defined maltreatment "more carefully in terms of particular physical harms or injuries, sexual conduct, or severe emotional damage" (Wald & Woolverton, 1990, p. 485).

RISK ASSESSMENT

Wald and Woolverton (1990) described three procedures as risk assessment: (1) categorization of cases on the basis of severity of risk; (2) "needs assessment," although the factors in this term are not necessarily those associated with the prediction of future abuse; and (3) a "process for assessing the likelihood that a given person (usually a parent) will harm a child in the future" (p. 486). Risk assessment is concerned "with predicting whether a child will be maltreated in the near future, absent intervention" (English & Pecora, 1994, p. 452). Although there is a strong practice assumption that risk assessment is used to predict a recurrence of maltreatment, it also refers "to an estimation of the likelihood that there will be an occurrence of maltreatment in a case where maltreatment has not occurred" (English & Pecora, 1994, p. 452).

Models

There are five basic types of risk assessment model (English & Pecora, 1994; Pecora, 1991):

1. matrix models—tables of risk factors rated in terms of their severity; for example, the Illinois Child Abuse and Neglect Training System

CCC Code:1070-5309/96 $3.00

(CANTS) 17B (Coler, 1982) or the Washington Assessment of Risk Model (WARM)
2. empirical predictor models—small sets of factors found to be predictive of substantiation or recurrence (for example, Baird, 1988; Johnson & L'Esperance, 1984; Weedon, Torti, & Zunder, 1988)
3. behaviorally anchored items or scales that assess levels of child or parent functioning—for example, the Family Assessment Scales (Magura, Moses, & Jones, 1987) and the Child Well-Being Scales (CWBS) (Magura & Moses, 1986)
4. comprehensive ecologically structured scales such as the Child At Risk Field (CARF) system (Holder & Corey, 1987, 1989), which uses 14 factors organized around five "force fields" (child, parent, family, maltreatment, and intervention) to predict maltreatment
5. computerized "expert systems," which use a blend of CPS expertise and artificial intelligence to derive computer-based decision rules (for example, Schuerman, Mullen, Stafner, & Johnson, 1989).

Criticism has been leveled at the field of risk assessment because of the confused and diverse way agencies use their risk assessment systems (Wald & Woolverton 1990). These concerns are especially relevant because most models have been developed through a process of consensus rather than through an empirically based process (Keller, Cicchinelli, & Gardner, 1988), and there is little unanimity about the definition of fundamental terms used in the models, such as "maltreatment" or "risk" (Berkowitz, 1991; Hutchison, 1990; Palmer, 1988; Zuravin, 1991). Furthermore, there is virtually no basic research on the correlates of maltreatment recurrence for any type of maltreatment, and although there is much research on maltreatment correlates, definitional and methodological problems make it difficult to draw conclusions with any degree of certainty.

Also, there is no reason to believe that the same variables predict both occurrence and recurrence. In their review of risk assessment models, McDonald and Marks (1991) found that only half of the 88 variables commonly used as factors predictive of maltreatment had been empirically validated. In addition, with few notable exceptions (for example, English, Aubin, Fine, & Pecora, 1993; Pecora & English, 1993), developers of child abuse screens have only recently begun to consider the relevance of cultural factors in the definitions of maltreatment, family functioning, and abusive parenting. Although risk assessment models may have some degree of face validity and concurrent validity, the empirical evidence to support their predictive validity is currently lacking (Camasso & Jagannathan, 1995; Pecora, 1989).

State Use of Risk Assessment Instruments

Despite the variety of risk assessment models and their use in 42 states (Berkowitz, 1991), there is no comprehensive national database currently available documenting what specific systems are used by the various states (T. Tatara, director, Research and Demonstration Department, Public Welfare Association, Washington, DC, personal communication, July 31, 1995). To complicate matters further, a single state may use more than one system or modify an existing system, or states may use the same system for different purposes. Of the 42 states using some form of risk assessment, 15 use a matrix approach, 13 use a scale for entering ordinal values, and 11 use risk assessment in certain counties or districts or use different approaches in different parts of the state. Fifteen use the Illinois CANTS 17B model or a derivation of it, four use the CARF system, one uses a combination of the CWBS and the Family Risk Assessment, and four use WARM or a modification of it (Berkowitz, 1991). However, most of the states' risk assessment instruments are a variation of a few basic models (Berkowitz, 1991; Doueck et al., 1993).

METHOD

Although a great deal of excellent research has been undertaken by state departments of social services on risk assessment models, unless that research has undergone the scrutiny of the peer review process through publication or presentation at a national conference (for example, the Texas and Washington models), it is difficult to determine the general relevance of the studies. Given the dynamic nature of the field and the importance of this material to child protection practitioners and researchers, it is appropriate to examine risk assessment model development and evaluation research and to bring this information together in one place. This article has two purposes: (1) to examine the psychometric properties of existing risk assessment models, including their reliability and validity, as reported in published evaluations and (2) to examine the outcomes of implementing these models.

The selection of studies was guided by two criteria: (1) The evaluation was published in the professional literature or presented at a national conference or published in conference proceedings, and (2) the evaluation had been conducted by an independent evaluator not connected to the agency that used the model. These criteria were adopted because they were expected to

result in the most generally relevant and useful data for examination. However, because so few models were developed empirically, results from two additional studies were included in the review, although they did not meet the independence criterion. Both studies reported on model development and predictive validity, one of the more important criteria in model development (Doueck, 1995) and one with limited empirical support (Camasso & Jagannathan, 1995; English & Pecora, 1994).

To cast a broad net during the search, we used both general and specific terms related to risk assessment. For example, the PsycLIT and ERIC databases were searched for the years 1975 to 1995 using the following search terms: risk assessment, risk analysis, risk measurement, risk management, decision making, psychometric, child abuse, child welfare, child maltreatment, child protection, and child protective services. In addition, relevant journals were searched manually for recent publications on risk assessment, and all of the *Proceedings of the Risk Assessment Roundtables* convened by the American Public Welfare Association and the American Humane Association were examined. The search resulted in studies on eight different models that met the two criteria: the Philadelphia model (a derivative of the Illinois CANTS 17B) (Fluke et al., 1993), the CARF system (Doueck et al., 1993; Fluke et al. 1993), the Alaska model (Baird, 1988), the Utah family assessment model (Nasuti & Pecora, 1993), WARM (Camasso & Jagannathan, 1995; Fluke et al., 1993; Marks & McDonald, 1989), CWBS (Fanshel, Finch, & Grundy, 1994; Gaudin, Polansky, & Kilpatrick, 1992), the Illinois CANTS 17B model (Camasso & Jagannathan, 1995), and the New York Child Protective Services Review Document (CPSRD) (Fanshel et al., 1994). The two additional models reviewed that did not meet the independence criterion were the Alameda County model (Johnson & L'Esperance, 1984) and the Vermont model (Weedon et al., 1988).

RESULTS

The research reported findings on seven models for predictive validity, five for interrater reliability, six for internal consistency, one for concurrent validity, and three for implementation effect (Table 1).

Predictive Validity

In risk assessment research, the test for predictive validity consists of determining the degree to which an identified risk assessment instrument is capable of differentiating individuals who are likely to maltreat their children in the future from those who are unlikely to maltreat. Related to the issue of predictive validity is an assessment of the degree to which an instrument has acceptable levels of specificity and sensitivity (Brown & Hollander, 1977), as well as acceptable rates of false positives (cases incorrectly labeled high risk) and false negatives (cases labeled low risk in which maltreatment subsequently occurs). In the context of risk assessment, *sensitivity (SE)* can be defined as the probability of correctly identifying people who will maltreat in the future:

$$SE = tp/(tp + fn),$$

where tp = true positive, fp = false positive, tn = true negative, and fn = false negative.

Specificity (SP) is the rate of correctly identifying people who are unlikely to maltreat in the future:

$$SP = tn/(fp + tn)$$

(Chmura-Kraemer, 1992). To calculate the rate of false positives, the number of people incorrectly predicted as having a recurrence is divided by the total number of people predicted as having a recurrence. The false negative rate is calculated by dividing the total number of cases incorrectly predicted as not having a recurrence by the total number of cases predicted as not having a recurrence. Although it is important to maximize both sensitivity and specificity, this discussion focuses on false negatives because they most represent failure to predict child maltreatment.

Alameda County Model. The Alameda County model was moderately successful in minimizing false negatives; in the validation group false negatives were 30.4 percent, although rereport was equated with reabuse (Johnson & L'Esperance, 1984). In a separate study, the Alameda model had false negatives of 38.3 percent using discriminant analysis and 76.6 percent using logit analysis (Marks & McDonald, 1989). The difference between scores is probably an artifact of the diagnostic cutoff point used in each of the programs.

The Alameda County model was also reviewed for prediction of recidivism by type of maltreatment, including physical abuse, physical neglect, sexual abuse, and combined abuse and neglect. False negative rates resulting from discriminant analysis and logit analysis, respectively, were as follows: physical abuse, 22.7 percent and 86.4 percent; physical neglect, 31.1 percent and 69.9 percent; and sexual abuse, 45 percent and 90.9 percent (Marks & McDonald, 1989).

Vermont Model. Although the Vermont model was found to be 68 percent accurate in overall prediction in one study, it inaccurately classified almost two of every three cases in prediction of "freedom from reabuse" (Weedon et al., 1988), highlighting the problem of dependence on overall prediction scores. These

TABLE 1—Summary of Information in Studies on Risk Assessment Models

Model	Authors and Date	Type of Model and Purpose	Information Available	Results[a]
Alameda County	Johnson & L'Esperance (1984)	Actuarial: prediction of reabuse	Predictive validity: reabuse (physical)	*fp* = 18.0 percent; *fn* = 30.4 percent; total correct = 74.4 percent
	Marks & McDonald (1989)		All types of abuse and neglect, sexual abuse, physical abuse and neglect	All types of child abuse and neglect *fp* = 20.8 percent or 2.4 percent; *fn* = 38.3 percent or 76.6 percent; total correct = 76.0 percent or 76.0 percent
Alaska	Baird (1988)	Actuarial: prediction of reabuse	Predictive validity	Abuse scale: *fp* = 21.9 percent; *fn* = 31.0 percent; total correct = 76.0 percent Neglect scale: *fp* = 26.0 percent; *fn* = 31.0 percent; total correct = 72.4 percent
Vermont	Weedon, Torti, & Zunder (1988)	Actuarial: prediction of reabuse and neglect	Predictive validity	*fp* = 14 percent; *fn* = 62 percent; total correct = 68 percent
			Interrater reliability	*r* = .75
			Internal consistency	Cronbach's α = .77
Utah	Nasuti & Pecora (1993)	Child Well-Being Scales derivative: family assessment	Internal consistency	All: Cronbach's α > .940
			Interrater reliability	All: Cronbach's α > .970
Child Well-Being Scales (segment)	Gaudin, Polansky, & Kilpatrick (1992)	Family assessment scales, family assessment	Concurrent validity	*fp* = 13 percent; *fn* = 21 percent
			Internal consistency	Supported
	Fanshel, Finch, & Grundy (1994)		Predictive validity	Significant correlation with worker perception of risk
Child At Risk Field system\	Doueck, English, DePanfilis, & Moote (1993)	Risk assessment and decision making	Process	Imperfectly implemented
	Fluke et al. (1993)		Interrater reliability	Group mean good, single rater mean poor to moderate
			Internal consistency	Variable (see text)
			System effects	Marginal
			Service effects	Marginal

(continued)

Washington Assessment of Risk Matrix	Fluke et al. (1993)	Matrix: family assessment	Interrater reliability	Group mean good, single rater mean poor to moderate
			Internal consistency	Good
			System effects	Marginal
			Service effects	Marginal
	Camasso & Jagannathan (1995)		Predictive validity	Generally poor (two studies)
	Marks & McDonald (1989)		Predictive validity	*fp* = 19.6 percent or 2.9 percent; *fn* = 14.3 percent or 85.7 percent; total correct = 81.0 percent or 70.5 percent
Philadelphia	Fluke et al. (1993)	Matrix: family assessment (Illinois CANTS 17B derivative)	Interrater reliability	Group mean good, single rater mean poor to moderate
			Internal consistency	Good
			System effects	Marginal
			Service effects	Marginal
Child Protective Services Review Document (New York)	Fanshel, Finch, & Grundy (1994)	Family assessment: risk categorization and case planning	Predictive validity (prediction of case workers' perception of risk)	Segment on perpetrator alcohol or drug use had significant correlation with workers' perception of risk
Illinois CANTS 17B	Camasso & Jagannathan (1995)	Matrix: risk assessment	Predictive validity (case substantiation, closing, and recidivism)	Generally poor

NOTE: *fp* = false positive; *fn* = false negative.

[a]First percentage is result by discriminant analysis; the second percentage is result by logit analysis.

overall scores may be misleadingly encouraging while masking high rates of false positives and false negatives. In the study, a nonrandom sample of child protection workers was asked to apply the instrument to their entire caseload on the basis of the information available after the initial investigation. However, because the cases had been open from three weeks to eight years, the task of trying to decide retrospectively what was known at that time was extremely difficult. Furthermore, it is entirely possible, as the authors suggested, that in their effort to ensure face validity, the integrity of the concept of risk was compromised by the introduction of the concept of family dysfunction. This may also account for the fact that several of the individual scale items had predictive validity scores higher than that of the total scale.

Alaska Model. In the Alaska study (Baird, 1988) cases of unconfirmed abuse were included in the maltreated sample, because workers used this category when they were unable to confirm but had a strong suspicion that abuse had taken place. The results indicated that neglect and abuse would be better predicted using two separate scales and that the rate of false negatives was 31 percent for both scales.

CWBS and CPSRD. The Child Well-Being Scales were examined by Fanshel et al. (1994) along with the New York Child Protective Services Review Document and the Problems and Conditions in the Life of the Family Scale (Beck & Jones, 1973). Because the latter was not designed as a risk assessment instrument and was found to be "largely superfluous for this population" (Fanshel et al., 1994, p. 1082), it is not included in this review. A segment of the CWBS based largely on the parental disposition subscale proved significantly correlated with workers' perceptions of risk, and the two scales contributed to several of the indexes of perceived risk developed by Fanshel et al. This research was carried out in a community populated predominantly by people of color, suggesting the scales may be useful in assessing this population. However, the authors also suggested that the scales did not adequately capture levels of extreme poverty, thus lending credence to early criticisms of the scales by Seaberg (1988): The scales lacked clarity of definition of child well-being, there were substantial missing data, the indicators lacked validity, and there were no clinical cuttoff points. The CPSRD defined an index of economic deprivation and an index of parental substance abuse. However, Fanshel et al. examined worker perceptions of risk, which are not necessarily predictive of actual risk. As a result, the two instruments may be more relevant for predicting worker decision making than subsequent maltreatment.

WARM and Illinois CANTS 17B. The WARM instrument was the subject of two studies (Camasso & Jagannathan, 1995; Marks & McDonald, 1989). Marks and McDonald found a false negative rate of 14.3 percent using discriminant analysis and 85.7 percent using logit analysis. Their study focused on the prediction of recidivism for all types of child abuse and neglect.

Camasso and Jagannathan focused on the prediction of three factors: case closing, case substantiation, and recidivism. They found that the model predicted all three at a level greater than chance. Their study also introduced receiver operating characteristic (ROC) curve analysis to the study of CPS risk assessment. ROC analysis allows for comparison of specificity and sensitivity rates at various diagnostic cutoff points. The procedure may have much to offer the field because it compensates for the arbitrariness of single cutoff points, which may contribute to the type of discrepancy between discriminant and logit analysis results noted above.

The WARM and Illinois CANTS 17B models were examined to determine their capacity to predict the three case events of interest to the study (Camasso & Jagannathan, 1995; Marks & McDonald, 1989). The researchers determined that both WARM and the Illinois CANTS 17B model were able to discriminate at levels better than chance, although both models were described as generally poor in terms of their predictive capacity.

Factors in Predictive Studies. These predictive studies present a rather pessimistic picture of the predictive validity of empirically based risk assessment instruments. However, the overall findings do not necessarily detract from the relative merits of the models or the research that led to their development. Nevertheless, the serious consequences of mistakes in either direction demand minimal error. So far there is too high a rate of error or uncertainty to recommend exclusive reliance on any risk assessment instrument for predictive purposes.

Table 2 shows the factors identified in the predictive studies. No factors were common to all models. However, perpetrator access and age or ability of child appeared the most frequently, in four of the models. Of the 31 factors, 13 are unique to individual models, and each of the studies derived a different number of predictors, ranging from 13 (Vermont) to five (Alameda County).

The criterion variable also varied depending on which model was examined. For example, the Alameda County model was developed to predict physical reabuse (Johnson & L'Esperance, 1984), whereas other

TABLE 2—Comparison of Predictors in Studies on Predictive Validity

Predictor	Baird (1988): Alaska model (abuse)	Baird (1988): Alaska model (neglect)	Johnson & L'Esperance (1984): Alameda County model	Weedon, Torti, & Zunder (1988): Vermont model	Marks & McDonald (1989): WARM	Marks & McDonald (1989): Alameda County model
Ability of child or age of child				Yes	Yes	Yes
Number of children in home and number of children abused	Yes	Yes	Yes			Yes
Caretaker history of abuse or neglect	Yes	Yes		Yes		
Caretaker depression or stress	Yes			Yes		
Drug and alcohol abuse	Yes	Yes		Yes		
Caretaker criminal history	Yes					
Parental expectations			Yes			
Parenting skill, quality of care			Yes	Yes		Yes
Number of adults in home or substitute caretaker	Yes			Yes		
Age of youngest caretaker		Yes				
Negative social relationships	Yes					
Previous referrals	Yes	Yes				Yes
Prior placement of child	Yes	Yes				
Problems noted in current referral	Yes	Yes				
View of nonperpetrating caretaker	Yes					
Perpetrator access to child			Yes	Yes	Yes	Yes
Adult motivation		Yes				
Cooperation with agency	Yes			Yes		
Ability to use agency resources			Yes			
Severity or frequency of abuse or neglect				Yes		
Caretaker abilities and control				Yes		
Parent–child relationships				Yes		
Environmental conditions in home, utilities in service				Yes		Yes
Strength of family support				Yes		
Dangerous acts					Yes	
Provision for basic needs					Yes	
Protection of child or mother					Yes	
Sexual contact					Yes	
Adequacy of supervision					Yes	
Lack of adequate shelter						Yes
Neglect reported						Yes

NOTE: WARM = Washington Assessment of Risk Matrix.

models were examined for prediction of all types of child maltreatment (Marks & McDonald, 1989); worker perception of risk (Fanshel et al., 1994); and case closing, case substantiation, and recidivism (Camasso & Jagannathan, 1995). It is apparent that there has been no clear uniformity of methods or universal definition of independent or criterion variables during model development.

Type of Analysis. The predictive studies generally used some combination of multiple regression and discriminant function analysis (Table 3). Such multivariate analyses may not be appropriate with low-base-rate phenomena, small sample sizes, and dichotomous variables, because the assumptions of multivariate normality and equal covariance matrixes on which these procedures are based are likely to have been violated

TABLE 3—Summary of Research Methods in Studies on Risk Assessment Models

Author and Date	Design	Sample	Measures	Analysis
Johnson & L'Esperance (1984)	Retrospective model development	Development sample: $n = 55$, $n = 85$ Random assignment to validation sample: $n = 81$, $n = 39$	105-item data abstraction form administered by 10 graduate students	Multiple linear discriminant analysis
Weedon, Torti, & Zunder (1988)	Prospective pre-experimental design	$n = 147$	24 workers and supervisors administered Vermont model to the workers' open cases	Regression analysis, discriminant analysis
Baird (1988)	Retrospective model development	$n = 550$ randomly selected cases: 262 Alaska natives, 217 white people, 71 other	Case abstraction of 100 or more variables Alaska abuse and neglect models	Correlation, multiple linear regression, cross tabulations, chi-square, Pearson's r
Nasuti & Pecora (1993)	Prospective test of measurement properties	$n = 15$, $n = 56$ Random assignment: no	Utah model (CWBS derivative)	Cronbach's alpha, Pearson's r, Spearman Brown prophecy
Gaudin, Polansky, & Kilpatrick (1992)	Prospective field trial	Neglect group: $n = 53$ Control group: $n = 80$	Segment of the CWBS	Correlation, bivariate analysis, discriminant analysis
Doueck, English, DePanfilis, & Moote (1993)	Retrospective multimethod (qualitative and quantitative), pre and post	Pre: $n = 118$ Post: $n = 89$ Interview: $n = 17$	85-item data abstraction form; page-for-page replication of the CARF system 26-item semistructured interview schedule	chi-square, bivariate analysis, principal components analysis
Fluke et al. (1993)	Retrospective multimethod (qualitative and quantitative)	Reliability analysis: $n = 180$, $n = 29$ Service effects: $n = 340$ (survey), $n = 47$ (interview)	CARF; WARM; Philadelphia model (Illinois CANTS 17B derivative) Structured interview protocol; mail survey	Cronbach's alpha, chi-square, bivariate analysis

(continued)

Camasso & Jagannathan (1995)	Comparative analysis of predictive ability	86 in-home cases 51 out-of-home cases 102 intake cases (random)	Data abstraction performed by seven trained coders with experience in law or social work	Receiver operating characteristic curve; logistic regression; Wilcoxon
Fanshel, Finch, & Grundy (1994)	Prospective test of predictive validity	*n* = 72 families	CWBS, CPSRD, and Problems and Conditions in the Life of a Family, completed by 12 volunteer CPS workers	Multiple regression
Marks & McDonald (1989)	Retrospective test of predictive validity	543 cases from Washington state 468 cases from Alameda County	Washington state case record review performed by individual staff members; Alameda County: case record review	Discriminant analysis and logit analysis

NOTE: CWBS = Child Well-Being Scales; CARF = Child At Risk Field system; WARM = Washington Assessment of Risk Matrix; Illinois CANTS 17B = Illinois Child Abuse and Neglect Training System 17B; CPSRD = Child Protective Services Review Document (New York); CPS = child protective services.

(Cooley & Lohnes, 1971; DePanfílis, 1993; Klecka, 1985; Pedhazur, 1982).

Interrater Reliability

A frequently cited benefit of risk assessment has been the assumption that implementation of a risk assessment procedure would lead to greater standardization of evaluation among child protection workers. In fact, the need for confidence that the same case would have been evaluated in the same manner by other workers is fundamental to the success of any risk assessment procedure. The interrater reliability of five models—CARF, WARM, the Philadelphia model (a derivation of the Illinois CANTS 17B), the Vermont model, and the Utah family assessment model (a derivation of the CWBS)—were examined in a number of studies.

Fluke and his colleagues (1993) tested three models: WARM, the Philadelphia model, and the CARF system. Each model was evaluated using case vignettes to determine the interrater reliability for risk by type of maltreatment at different decision points during the case. Space does not allow for a full reporting of the results; the overall scale reliability scores of all the models were considered good. Even so, none of the three models scored as well on mean single-rater correlations when compared with the overall model scores. Although the interrater reliability was high, actual performance did not live up to that potential.

Examination of the interrater reliability of the Utah family assessment scales resulted in a stepped-up score for the Spearman Brown prophecy formula above .970 (Nasuti & Pecora, 1993). However, the authors pointed out that there was no certainty that the caseworkers rated the vignettes in the same way that they would have rated an actual case. Finally, the Vermont model scale score for interrater reliability was $r = .75$ (Weedon et al., 1988). However, only four of the 13 item scores were over .7, suggesting that there may be too much subjectivity built into this scale.

The relatively good scores received by most models on interrater reliability are not surprising. A difficulty inherent in the use of interrater reliability scores is that child protection workers may be able to differentiate high-risk from low-risk cases but often have difficulty differentiating between cases in the middle range. A valid test of interrater reliability, therefore, should include an examination of the entire range of cases, from no risk to high risk, across the different types of maltreatment.

Internal Consistency

A commonly used method to measure internal consistency is Cronbach's alpha coefficient, which focuses on the interrelationships between scale items and their ability to measure the same concept (Brown, 1983). It can be interpreted as the average expected correlation among all items measuring the concept.

The internal consistency of WARM, the Philadelphia model, and the CARF systems was reviewed by Fluke and his colleagues (1993) by initial assessment, follow-up, and type of maltreatment. Each of the models appeared acceptable, with Cronbach's alphas ranging from .72 to .97. The one exception was the CARF model, in which the score for initial assessment of physical abuse and neglect was .57, below a generally accepted standard (Carmines & Zeller, 1979).

Internal consistency scores for the Utah family assessment model were quite high: Cronbach's alphas were .972, .959, and .947 for the parent, child, and household scales, respectively (Nasuti & Pecora, 1993). Internal consistency for the Vermont model was also acceptable at .77 (Weedon et al., 1988).

Concurrent Validity

The test for the existence of maltreatment, concurrent validity, is an examination of identifying factors that help differentiate people who have maltreated a child from those who have not. In a test of concurrent validity, Gaudin et al. (1992) found that the CWBS was able to discriminate successfully in relation to neglect. Discriminant analysis scores on three combined factors (those identified by Magura & Moses, 1986) correctly classified 79 percent of neglectful families and 87 percent of control families (the false negative rate was 21 percent). The authors concluded that the scales were able to discriminate between neglectful and nonneglectful families, providing some support for the scales' concurrent validity. However, one of the difficulties associated with the complexities of concurrent validity is the potential that positive findings for concurrent validity will be misconstrued as positive findings for predictive validity (Caldwell, Bogat, & Davidson, 1988).

Service and System Effects of Implementation

Two studies examined the service and system effects of model implementation for the CARF system and the Philadelphia model. Despite imperfect implementation of the CARF system, Doueck, Levine, and Bronson (1993) found a significant association between data collected and some subsequent caseworker decisions. However, the researchers were unable to determine to what extent the system was being used as a guide for worker decisions or as a tool to document worker decisions already made.

One of the more striking findings in the Philadelphia study (Fluke et al., 1993) was that the use of a specific, or even of any, risk assessment model made little difference in the type of caseworker decisions made. In addition, almost half of the responding intake workers reported using a risk assessment model after a decision had been reached. Furthermore, there was only a marginal drop in the volume of cases seen by workers across the state and a general sense that risk assessment was largely peripheral to their work with clients. In addition, despite attempts to measure implementation of the systems, the researchers had some difficulty assessing precisely when the systems were fully implemented and when, or if, workers began to use them appropriately.

The results of these studies highlight the need to determine clearly the degree to which the risk assessment systems under examination have been adequately implemented by the users of the systems before an outcome evaluation is undertaken. Despite problems with implementation found in the CARF and Philadelphia studies, however, the use of risk assessment tended to be associated with consistency of caseworker decisions and the use of a common framework for organizing information and communicating about cases.

DISCUSSION

The use of risk assessment models by child protective services agencies is a relatively new phenomenon, occurring in numbers only during the past 10 years. Given this rush to implementation, it is not surprising that research on model development and independent evaluations of the models have lagged behind. Despite the relative newness of model development and independent evaluation in the field of risk assessment, there are grounds for some optimism. For some risk assessment approaches, such as the Philadelphia model, WARM, the CARF system, and the CWBS, psychometric and process information is available. In addition, many of the models examined appear to have acceptable psychometric properties, especially internal consistency, interrater reliability, and concurrent validity. Furthermore, although the current level of predictive validity for the models evaluated would not allow for major dependence on them for case decision making, the research on actuarial models, such as those used in Alaska, Alameda County, and Vermont, is somewhat encouraging.

Further evaluation of risk assessment models in the field should be targeted at an examination of caseworker decision-making processes and the extent to which a given risk assessment model is congruent with those processes. Minimizing implementation problems may be as important for model development as concern about sensitivity and specificity currently is. To date the level of process evaluation or implementation research appearing in the published literature or presented at conferences has been less than adequate. Although we agree with English and Pecora's (1994) suggestion that the level of implementation research has been substantial, much of the research they refer to is unpublished. The published and presented research in this area is not particularly promising because it indicates that risk assessment models are imperfectly implemented, are at times used to document decisions rather than to assess risk or guide worker decisions, and are perceived by some as irrelevant to their work with troubled families. Findings from even the most well-designed outcome evaluations will be insufficient unless it is determined that the model being evaluated has been adequately (that is, reliably and consistently) implemented (Doueck, Bronson, & Levine, 1992; English & Pecora, 1994).

There also needs to be some general agreement about the major concepts in the field, specifically what is meant by "risk assessment." For example, is the term appropriately applied to "family assessment" models, like the CWBS or the CARF system? Does the term include assessing the risk of substantiation, recurrence of maltreatment, or risk for maltreatment? If it includes risk for maltreatment, does the term include an assessment of severity of maltreatment, frequency of maltreatment, or both? Without some consistency in definitions, comparison across models is difficult.

Further research will benefit from increased sophistication in research design and analysis and from larger sample size with more diverse populations. The more recent studies (for example, Camasso & Jagannathan, 1995; Fanshel et al., 1994) have begun to address some of the issues that have continually plagued risk assessment—the lack of information on the cultural sensitivity of instruments and the difficulties inherent in comparing predictive validity models. The introduction of ROC curve analysis may be a step forward for risk assessment research and offers the opportunity to revisit existing models in new studies.

Greater effort is needed in the field to use empirically based methods to enhance model development. Most models tend to have some degree of face validity, because they were originally developed by consensus or after a thorough review of the literature. However, to date few models have used empirical methods to further model development. In addition, there needs to be a greater focus on independent evaluations of risk assessment models using sound research methods.

Finally, concern has been expressed about the possibility that risk assessment may mistakenly be adopted as an alternative to education, supervision, and training (Wald & Woolverton, 1990). However, results from the implementation studies suggest that, on the contrary, risk assessment demands quality in education, training, and supervision, as well as vigilance and persistence on the part of administrators hoping to use it. ■

REFERENCES

Baird, C. (1988). Development of risk assessment indices for the Alaska Department of Health and Social Services. In T. Tatara (Ed.), *Validation research in CPS risk assessment: Three recent studies* (pp. 85–121). Washington, DC: American Public Welfare Association.

Beck, D. F., & Jones, M. A. (1973). *Progress on family problems: A nationwide study of clients' and counselors' views on family agency services.* New York: Family Service Association of America.

Berkowitz, S. (1991). *Key findings from the state survey component of the study of high-risk child abuse and neglect groups.* Washington, DC: National Center on Child Abuse and Neglect.

Brown, B. W., Jr., & Hollander, M. (1977). *Statistics: A biomedical introduction.* New York: John Wiley & Sons.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: CBS College Publishing.

Caldwell, R., Bogat, A., & Davidson, W. (1988). The assessment of child abuse potential and the prevention of child abuse and neglect: A policy analysis. *American Journal of Community Psychology, 16,* 609–625.

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

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage Publications.

Chmura-Kraemer, H. (1992). *Evaluating medical tests.* Newbury Park, CA: Sage Publications.

Coler, G. L. (1982). *Child abuse and neglect investigation decisions handbook.* Springfield: Illinois Department of Children and Family Services.

Cooley, W. W., & Lohnes, P. R. (1971). *Multivariate data analysis.* New York: John Wiley & Sons.

DePanfilis, D. (1993). *A proximate test of the construct and predictive validity of the elements and influences represented in the ACTION for Child Protection Child At Risk Field risk assessment and decision making system.* Baltimore: University of Maryland.

Doueck, H. J. (1995). Screening for child abuse: Problems and possibilities. *Applied Nursing Research, 8,* 191–198.

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

Doueck, H. J., English, D. J., DePanfilis, D., & Moote, G. T., Jr. (1993). Decision making in child protective services: A comparison of selected risk assessment systems. *Child Welfare, 72,* 441–453.

Doueck, H. J., Levine, M., & Bronson, D. E. (1993). Risk assessment in child protective services: An evaluation of the Child At Risk Field system. *Journal of Interpersonal Violence, 8,* 446–467.

English, D. J., Aubin, S. W., Fine, D., & Pecora, P. J. (1993). *Improving the accuracy and cultural sensitivity of risk assessment in child abuse and neglect cases.* Seattle: Washington State Department of Social and Health Services.

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

Fanshel, D., Finch, S. J., & Grundy, J. F. (1994). Testing the measurement properties of risk assessment instruments in child protective services. *Child Abuse & Neglect, 18,* 1073–1084.

Fluke, J. D., Wells, S., England, P., Walsh, W., English, D. J., Johnson, W., Gamble, T., & Woods, L. (1993). Evaluation of the Pennsylvania approach to risk assessment: Summary of the results for project objectives 1, 2, and 4. In T. Tatara (Ed.), *Seventh National Roundtable on Child Protective Services Risk Assessment: Summary of highlights* (pp. 113–172). Washington, DC: American Public Welfare Association.

Gaudin, J. M., Jr., Polansky, N., & Kilpatrick, A. C. (1992). The Child Well-Being Scales: A field trial. *Child Welfare, 17,* 319–328.

Holder, W., & Corey, M. (1987). *Child protective services risk management: A decision-making handbook.* Charlotte, NC: ACTION for Child Protection.

Holder, W., & Corey, M. (1989). *The Child At Risk Field system: A family preservation approach to decision-making in child protective services.* Charlotte, NC: ACTION for Child Protection.

Hutchison, E. D. (1990). Child maltreatment: Can it be defined? *Social Services Review, 64,* 60–78.

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

Keller, R. A., Cicchinelli, L. F., & Gardner, D. M. (1988). *A comparative analysis of risk assessment models: Phase 1 report.* Denver: Applied Research Associates.

Klecka, W. R. (1985). *Quantitative applications in the social sciences 19: Discriminant analysis.* Beverly Hills, CA: Sage Publications.

Magura, S., & Moses, B. S. (1986). *Outcome measures for child welfare services: Theory and applications.* Washington, DC: Child Welfare League of America.

Magura, S., Moses, B. S., & Jones, M. A. (1987). *Assessing risk and measuring change in families: The Family Risk Scales.* Washington, DC: Child Welfare League of America.

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

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

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

National Center on Child Abuse and Neglect. (1988). *Study findings: Study of national incidence and prevalence of child abuse and neglect.* Washington, DC: U.S. Government Printing Office.

Palmer, M. (1988). *Risk assessment in child protective services: Part 1. Risk assessment models: A comparative analysis.* Portland: University of Southern Maine, National Welfare Resource Center for Management and Administration.

Pecora, P. J. (1989). Evaluating risk assessment systems: Methodological issues and selected research findings. In P. Schene & K. Bond (Eds.), *Research issues in risk assessment for child protection* (pp. 47–59). Denver: American Association for Protecting Children.

Pecora, P. J. (1991). Investigating allegations of child maltreatment: The strengths and limitations of current risk assessment systems. *Child and Youth Services, 15*(2), 73–92.

Pecora, P. J., & English, D. J. (Eds.). (1993). *Multicultural guidelines for assessing family strengths and risk factors in child protective services.* Olympia: Washington Department of Social and Health Services.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart & Winston.

Schuerman, J. R., Mullen, E., Stafner, M., & Johnson, P. (1989). First generation expert systems in social welfare. *Computers in Human Services, 4*(1/2), 111–122.

Seaberg, J. R. (1988). Child Well-Being Scales: A critique. *Social Work Research & Abstracts, 24*(3), 9–15.

U.S. Advisory Board on Child Abuse and Neglect. (1990). *Child abuse and neglect: Critical first steps in response to a national emergency.* Washington, DC: U.S. Department of Health and Human Services, Office of Human Development Services.

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

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

Zuravin, S. J. (1991). Research definitions of physical child abuse and neglect: Current problems. In R. H. Starr & D. A. Wolfe (Eds.), *The effects of child abuse and neglect* (pp. 100–128). New York: Guilford Press.

Original manuscript received November 28, 1994
Final revision received January 23, 1996
Accepted February 23, 1996

An earlier version of this article was presented at the Evaluation of Social Work Practice Conference, September 1995, West Yorkshire, England. Preparation of this manuscript was facilitated in part by funds from the Janet B. Wattles endowment and the Buffalo State College Center for the Development of Human Services.

Correction

In the article "The scientific practitioner revisited: Some obstacles and prerequisites for fuller implementation in practice" by Aaron Rosen (June 1996, Vol. 20, No. 2, pp. 105–111), a paragraph was misplaced. The first paragraph in the right hand column, beginning "The similarities in problem situations," should have appeared on page 108 immediately before the heading "Improvement in the Professional Curriculum."