



Pergamon

Children and Youth Services Review, Vol. 22, Nos. 11/12, pp. 813-837, 2000
Copyright © 2000 Elsevier Science Ltd
Printed in the USA. All rights reserved
0190-7409/00/\$-see front matter

PII: S0190-7409(00)00123-7

Risk Assessment in Context

Eileen Gambrill

Aron Shlonsky

University of California, Berkeley

This article provides an overview of the context in which decisions about risk are made in child welfare including personal, task, and environmental factors that may contribute to uncertainty and less-than-optimal decision making, as well as some of the methodological challenges posed by the use of current risk assessment instruments. Actuarial, consensus-based, and clinical instruments are discussed and the more successful track record of actuarial decision-making in child welfare and related fields is highlighted. Methodological challenges to assessing risk are also presented including lack of reliability and validity of measures, definitional dilemmas, temporal issues such as changes in risk over time, absence of base rate data, predicting for individuals and sensitivity and specificity of measures. Implications for the design and implementation of risk assessment tools are considered in light of contextual influences and methodological limitations. Lastly, an overview of the contents of Part One of this special issue on risk assessment is provided.

Child welfare staff make many different kinds of decisions. One concerns risk assessment: will this parent reabuse her/his child in the near future? What is the probability that she/he will do so? Risk assessment requires the integration of various kinds of data (e.g., self-report, observation, agency protocol) that differ in their accuracy, complexity, and subsequent value when making key decisions. Risk assessment is subject to a host of errors. Two key errors are overestimating the true probability of risk to a child and underestimating this risk. Such errors may result in failing to protect children from harm or imposing unneeded services that increase rather than decrease risk (e.g., unwarranted placement of children in foster care).

As of 1996, at least 76 percent of U.S. states used some type of risk assessment measure as a decision aid in child welfare cases (Tatara, 1996), however most have questionable reliability and/or validity (Camasso &

Requests for reprints should be to Eileen Gambrill, School of Social Welfare, 120 Haviland Hall, University of California, Berkeley, California, 94720.

Jagannathan, 2000; Lyons, Doueck, & Wodarski, 1996; Wald & Woolverton, 1990). More recent actuarial instruments developed by the Children's Research Center have the potential to lessen some of these concerns (Baird, Wagner, Healy & Johnson, 1999; Baird & Wagner, 2000), yet other issues remain.

Our purpose in preparing these special issues of *Children and Youth Services Review* on risk assessment in child welfare is twofold. A primary aim is to broaden the concept of risk, attending to behaviors and circumstances of parents in relation to their children, as well as considering staff behaviors that influence risk and the factors that affect them, such as child welfare policies, training programs, laws, organizational culture, and funding patterns (see Gambrill & Shlonsky, in press). A second purpose is to draw readers' attention to critiques, research, and theory that we hope will forward risk assessment. We have encouraged the authors to present material in a manner accessible to all readers, including practitioners and administrators.

The Decision Making Context

Decisions are made in a context of uncertainty. Caseworkers must distinguish between child neglect, bad parenting, and the effect of poverty, and they must do this without the aid of accurate assessment tools. Both personal and environmental factors influence decisions. Barriers to accurate decisions include: (1) limited knowledge; (2) limited information processing capacities; (3) personal obstacles such as lack of perseverance, reliance on ineffective problem-solving strategies and lack of familiarity with problem-related knowledge; and (4) the task environment. Problems that confront clients are often difficult, challenging even the most skilled staff. Predictions must be made under considerable uncertainty in terms of the relationship between the information at hand (predictor variables) and service outcome. Rarely is all relevant material available, hampering problem-solving efforts. Even when a great deal is known, this knowledge is usually in the form of statistical associations that cannot readily be calculated without assistance (Dawes, 1988). Although we know more about behavior today than we did years ago, we know little compared to what is unknown. We often do not know the true prevalence of a behavior or its natural history. The probabilities of different outcomes given certain interventions may be unknown. Empirical knowledge related to practice is

fragmentary and theory must be used to fill in the gaps. Every source of information has an unknown margin of error that may be small or large, and errors are compounded when decisions are made based on this inexact information. Competing values may also influence error. For example, steps must be taken to protect children from abuse while maximizing the decision-making freedom of parents.

Decisions are influenced by the personal characteristics of the decision makers (see, e.g., reviews in Gambrill, 1990; Gibbs & Gambrill, 1999; Nisbett & Ross, 1980). We can only consider a limited number of possibilities at one time, thus we tend to use certain strategies to categorize and interpret the flow of information including: (1) selective perception (we do not necessarily see what is there), (2) sequential (rather than contextual) processing of information, and (3) reliance on "heuristics" (strategies) to reduce effort (e.g., availability). Although the heuristic strategies we use to simplify judgmental tasks and decrease effort may often help us to make accurate judgments, at other times they may result in errors. Our memory is faulty. Avoidable mistakes may result from lack of knowledge on the part of staff. Knowledge may be available but not used. Preconceptions may get in the way as well as day-to-day mood changes that influence judgment. We attend to events that are vivid and often ignore data that are less vivid (but perhaps far more informative). We are influenced by primacy, or anchoring effects in which we are unduly influenced by what we first hear or consider. Not only are initial beliefs resistant to new evidence, they also are remarkably resistant to challenges of the evidence that led to them. We are subject to wishful thinking (i.e., our preferences for an outcome increase our belief that it will occur) and to the illusion of control (simply making a prediction may increase our certainty that it may come true). Lack of interest in having a carefully thought out position or a wish to appear decisive may compromise the quality of reasoning, as may a preference for mystery over mastery.

We are subject to a number of confirmation biases. For example, we disregard data that do not support our preferred beliefs and assign exaggerated importance to data that do support our beliefs (for a review, see for example Klayman, 1995). We often use different standards to criticize opposing evidence than to criticize supporting evidence. Some helpers have a pathological set; they search for deficiencies and neglect client assets. This will limit opportunities to solve problems. The fundamental attribution error is common in which causes are mistakenly attributed to the dispositional characteristics of the person (e.g., impulsivity) and environmental

variables (e.g., poor quality housing) are overlooked. Both of the latter tendencies encourage overestimates of pathology. Our assumptions about covariations (what events go together and how strongly they are associated) influence our causal assumptions. These cognitive biases highlight the importance of developing risk assessment measures that minimize their influence.

Misunderstandings regarding probabilities can result in faulty problem solving. CPS workers are not generally taught how to think statistically, instead relying on a combination of experience, intuition, and individual heuristics. As Wells (1988) suggests, "Workers are seldom educated in the use of probabilities, base rates, and knowledge about regression principles as they pertain to child welfare judgments" (p. 239). In the conjunctive fallacy, we act as if seemingly related multiple events have a greater probability of occurring than the single probabilities of which multiple events are comprised. In fact, the probability of both events occurring, by definition, can only be as great as the event with the smallest probability. Common errors in assessing how closely two or more events are related include ignoring non-occurrences, preconceptions about which events are related, and attempted proof by selected instances (attending to observed rather than relative frequency). That is, rather than examining all four cells of a contingency table, we focus on the present-present cell that represents having both a presumed characteristic and the problem, ignoring data in other cells. Additionally, the way in which we frame problems influences our judgments; we tend to have more extreme reactions to problems posed in terms of possible losses rather than possible gains, even though these probabilities may be equal.

Environmental characteristics also influence decisions. Decisions made in child welfare are affected by the values and policies of agencies and the broader community (Costin, Karger, & Stoesz, 1996; Margolin, 1997; Pelton, 1989). Time pressures and distractions may encourage a mindless, mechanical approach in which decisions are made without due consideration. Pressure to conform may result in poor decisions as illustrated by the play of "group think" in case conferences (Dingwall, Eekelaar, & Murray, 1983; Janis, 1982). Group think refers to neglect alternative views in a group focused on attaining agreement with one particular view (see, for example, Janis, 1982). Tolerating feeble inferences, rewarding gold and garbage alike, and the buddy-buddy syndrome (a reluctance to criticize friends) may dilute the quality of decisions in case con-

ferences (Meehl, 1973). Thus, the agency culture influences the quality of decisions and subsequent risk to all involved parties.

Actuarial vs. Clinical Decision-Making

The many sources of bias discussed above suggest the need for procedures that minimize them. Actuarial models are designed to address some of these biases. They are based on empirical relationships between certain predicted variables and outcomes. Actuarial models can be contrasted with consensus-based systems in which practitioners assess selected characteristics identified by agreement among experts, and then make their own judgment about an outcome such as risk. Both can be contrasted to clinical intuition, which is not informed by expert-consensus. To date, more than 100 studies have found actuarial models to be superior to clinical prediction in a variety of complex circumstances (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996). Even when the predictions of "experts" are compared with statistical models, they fall short of the mark (Dawes, 1994). Part of the reason for the success of the actuarial model lies in its reliability. By simply improving the reliability of decisions (consistently making the same decisions using the same criteria), the "hit rate" or number of accurate predictions is likely to increase (Dawes et al., 1989; Ruscio, 1998). The reliability of consensus-based tools should be higher than unassisted clinical intuition if users rely on a common set of criteria to evaluate risk. Indeed, consensus-based tools are better predictors than the "experts" who collected the data for the instrument as a result of the increased reliability (Dawes, 1994). However, if these models do not test how well the instrument predicts and the tool is not revised accordingly (in effect, transforming it into an actuarial model), they cannot compete with actuarial models. Despite the superiority of actuarial models in many areas, clinical prediction (based on intuition or impressions) continues to thrive in human services.

Actuarial vs. Clinical Decision Making in Child Welfare

Risk assessment models in child welfare are, essentially, lists of variables (e.g., caregiver and child characteristics or attributes, abuse circumstances, or environmental circumstances) that have been found to predict an outcome of interest (e.g., the initial occurrence or the recurrence of

abuse). The availability of more sophisticated and timely client tracking techniques, developments in computer technology, and more advanced statistical methods encourage the use of actuarial models for assessing risk. The hope is that such methods will provide greater consistency and enhanced predictive validity of decisions through the optimal weighting of risk indicators. Traditionally, lists of predictors have been based on consensus (expert opinion and accepted practice wisdom) and/or simple correlations found in exploratory research (Baird, Wagner, Healy, & Johnson, 1999; Pecora, 1991; Wald & Woolverton, 1990). Actuarial models in child welfare are generally developed by taking a sample of children and families involved in the child welfare system, analyzing their paths while in the system, relating those paths to a set of characteristics or events related to each family, and identifying events which are highly associated with an outcome of interest (usually recurrence of abuse/neglect). An event/characteristic becomes "predictive" if it remains highly associated with the outcome when the independent effects of other highly associated variables are considered. In (1990), Wald and Woolverton wrote a key article identifying many concerns about the use of actuarial models in child protective agencies. Although they maintained that actuarial processes have the greatest potential for creating reliable and valid risk assessment instruments, they cautioned that these models are far from perfect, that implementation errors may invalidate the models, and that development of accurate actuarial models will take time. Concerns raised by Wald and Woolverton nearly a decade ago, remain (see later section on methodological challenges).

Decision-making in child welfare has been consistently characterized by low reliability (Lindsey, 1992). Reliability refers to the degree to which different workers make the same placement decisions when presented with the same data. Rusio (1998) builds a compelling case for making improvements in this area. He cites a review of the literature by Lindsey (1992), who generously estimates the reliability of placement decisions to be .25 and, using figures derived from a nationwide survey, (Lindsey, 1991) estimates that even with perfect validity (the degree to which a measure actually predicts an event), the actual "hit rate" (number of correctly classified cases) with this reliability can only reach 72%. According to Lindsey, this finding means that, overall, 48% of placements were unnecessary and 45% of the children needing placement remained at home. Following up on Lindsey's work, Ruscio showed that the hit rate assuming zero validity (completely random) is 58%. Thus the actual hit rate proba-

bly lies between 58% and 78%, indicating that there is a high error rate in child protective services placement decisions.

Rossi, Schuerman, and Budde (1996) developed a series of case vignettes and asked CPS "experts" from four states to determine whether the child(ren) should be placed in out-of-home care. Responses varied by individual and by county, consistent with findings of poor inter-rater reliability among unassisted human judges in other fields (Dawes, 1994). The study also found that, given the additional choice of family preservation services, case workers often chose to provide child welfare services to children who would otherwise be left in the home without further intervention, suggesting that decisions are influenced by the availability of options. Baird et al. (1999) compared the reliability of two consensus-based risk assessment tools (The California Risk Assessment System and the Washington Risk Assessment Matrix) and one actuarial model (The Michigan Family Risk Assessment Model) using CPS cases ($n=80$) with known recurrence outcomes. The level of agreement among the readers using the actuarial model was significantly higher.

Since 1990, many studies have described risk assessment tools with statistically significant predictive capacity (Camasso & Jagannathan, 1995; Lyons, Douek, & Wodarski, 1996; McDonald & Marks, 1991; Zuravin, Orme, & Hegar, 1994). Actuarial models have been found to outperform "expert-driven" models in the limited research conducted to date. Johnson (1992) conducted an exploratory study ($n=54$) comparing individual clinical predictions of substantiation of child physical abuse with a statistical model. The actuarial model had greater predictive efficiency (72.2% vs. 63% predictions correctly classified) and, unlike clinical judgment, was highly correlated with substantiation. Although the actuarial model was better at predicting abuse, roughly 1 in 4 reports (27.8%) were not predicted accurately.

Falco and Salovitz (1997) compared New York's clinical risk assessment instrument with an actuarial model derived from a cohort of CPS cases that were followed from inception for 21 months ($N=740$). The actuarial model was more likely than the clinical model to correctly label a case that did not have a subsequent reabuse incident as "low risk" and was more likely to correctly label a case that had a subsequent reabuse incident as "high or very high risk." The actuarial model also had a more uniform distribution (meaning the number of cases classified in each risk category was more evenly distributed using the actuarial model, thereby increasing the predictive power of the instrument) and showed superior predictive

validity across counties and ethnic groups (potentially minimizing the biases associated with these variations). More recently, Baird and Wagner (this issue) found that an actuarial model (The Michigan Family Risk Model) had far greater predictive validity compared to two consensual models (The Washington and California Models).

Methodological Challenges

Actuarial risk assessment tools have the potential to provide the most objective, consistent treatment of children and families affected by the child welfare system (Grove & Meehl, 1996; Wald & Woolverton, 1990). However, while actuarial models out-predict clinical decisions, they are limited in their predictive capacity by many factors. Other tools may be useful for different types of case decisions, but suffer from many of the same factors.

Definitional Dilemmas

Predicting child abuse and neglect is made more difficult as a result of vague definitions of outcome measures. In particular, abuse and neglect are not sufficiently well defined to build accurate prediction models. The criteria defining maltreatment are diffuse across studies, making meta-analysis impossible (Wald & Woolverton, 1990). Not only must we be concerned about individual variability, we must be concerned about the variability of the outcome being measured. For example, combining studies with different cut points can expand the definition of abuse beyond any individual study's cut points. Or, a broad view of abuse may be combined with narrow definitions, creating a sample with a wide variability of abuse events. Although neglect is the most common form of child maltreatment with occurrence estimates ranging from 52 percent in the U. S. (U.S. Department of Health and Human Services, 1997) to over 70 percent in California (Needell, Webster, Cuccaro-Alamin, Armijo, Lee & Brookhart, 2000), its definition is characterized by subjectivity (Rose & Meezan, 1996), decreasing the likelihood of accurate assessment.

Zuravin (1999) reviewed all empirical studies with findings regarding child neglect in a major child maltreatment journal between 1992 and 1996. Out of 25 articles, only two used the same operational definition. She also noted that arbitrary cut points for inclusion criteria may influence

rates of false positives and false negatives (see, e.g., Camasso & Jagannathan, 1995; Ruscio, 1998). Further, reabuse rates for abuse and neglect are different indicating that they should be considered separately. Most studies define abuse as a categorical variable and use statistical techniques that are based on this assumption when, in fact, the outcome may resemble a continuum based on severity and on the variability of definition across studies and jurisdictions (Macdonald & Macdonald, 1999). Thus definitional variation blurs the line between categories (cut points), expanding the range of categories to more closely approximate a continuum of risk.

To date, actuarial tools in child welfare use substantiation (social work finding that maltreatment has occurred) as their outcome variable measuring the recurrence of abuse, and this is usually done retrospectively. However, substantiation may not be the most valid measure of recurrence (Drake, 1996; Wolock, Sherman, & Feldman, in press) and some have argued that this estimate is too conservative (see, e.g., English, Marshall & Orme, 1999). Although this may affect the true predictive power of an actuarial model, it does not invalidate the actuarial approach. It simply means that better ways of measuring recurrence must be developed and incorporated into the instrument. The use of the instrument itself must also be considered. Is the tool being used to initially assess risk, to assess whether a child should be returned to their biological parents, or at some other point in the life of the case? Although these distinctions have varying types of risk associated with varying sets of predictors, the same risk assessment instrument may be used by agencies at different points in time, resulting in varying degrees of reliability and validity (see Camasso & Jagannathan, 2000). Further, if the object is not to assess risk, but to assess family functioning or some other construct, perhaps an actuarial risk assessment instrument is not the best tool.

Lack of Reliability and Validity Measures

Research concerning risk assessment often makes use of previously constructed measures (see English & Graham, 2000), or some portion of them—essentially a new measure (see Wolock et al., in press), or creates a new measure. Examples include measures of social isolation, family conflict, parenting skills, and depression. There is a striking lack of attention to and concern with reliability and validity of measures used (e.g., of parenting skills, child behavior, and so on). Too often, research reports do not clearly describe the reliability and validity of these measures (or lack

thereof), so readers are not accurately appraised regarding the status of each measure in relation to its reliability or validity. The specific kind of reliability (e.g., homogeneity, stability) or validity (e.g., construct, concurrent, predictive) explored is often not described. Too often, we find purely rhetorical statements "that praise the sample [or the reliability and validity of a measure] vaguely and without evidence, or that criticize the sample [or some other aspect of methodology] vaguely and without evidence" (Lykken, 1991, p.4). Typically in these studies, the aim is to determine the relationship between certain concepts allegedly measured by such scales and other events (e.g., reabuse). Low correlations found between measures and events may be due to the fact that measures used are not valid indicators of concepts of interest (e.g., family conflict). A researcher may decide prematurely that risk factors identified empirically are not valid because they do not correlate with measures of social isolation or parenting skills when, in fact, this finding results because the measures used are neither reliable or valid. Of course, the opposite can also be true—that high correlations on certain scales or items may reflect the repeated poor measurement of a construct rather than a real association.

Changes in Risk Over Time/Stages of Development

Risk may change over time (DePanfilis & Zuravin, 1998) and we may be unaware of the point in the cycle at which we are intervening. This has implications for prediction. If escalation is always assumed at the point of risk assessment, the false positive rate might be very high. If escalation is not assumed, the number of false negatives might be high. We may get some of the more obvious serious cases at a certain stage in the process (high escalation) and encounter a certain number of low risk cases with the potential to escalate, only we do not know how to identify them.

The Absence of Baseline Data

The study of risk factors associated with abuse/neglect is limited by the context in which it occurs. The actual risk of recurrence cannot be explored in the absence of intervention by child protective service agencies (Wald & Woolverton, 1990). For instance, we cannot know the real rates of recurrence of the most severe maltreatment since most children experiencing such abuse are most likely removed from the homes of their abusers or major steps are taken to ensure their safety. Therefore, we are

mostly limited to knowing the rates of recurrence among those who committed less serious offenses, discovery of the extent of maltreatment was limited, or the maltreatment is in the beginning stages of a more severe progression. Given these limitations, obtaining an accurate base rate of maltreatment is probably impossible. We can only measure the effectiveness of one predictive tool against another. Further, discovering the false positive rate is almost impossible. Once the risk has been responded to (i.e., child welfare services are provided) the likelihood of recurrence of abuse in the absence of intervention cannot be determined. Child abuse investigation is, by definition, a reactive process. Families are reported to child protective services by mandated reporters, or other involved parties, after the alleged abuse/neglect has occurred. Therefore, most predication is an estimate of whether the abuse will occur again. This is problematic when predictive models are based on the shared characteristics of abusive/neglectful families (Wald & Woolverton, 1990). In other words, the caregivers being modeled in risk assessment studies and upon whom risk assessment instruments are being used are already abusive/neglectful. These shared characteristics should have already occurred and do not necessarily portend the future.

Predicting For Individuals

While an instrument may have high overall predictive validity, the predictive capacity for an individual is lower due to the wide variation among individuals. For instance, a model might predict that a substance-abusing parent is four times as likely to reabuse (relative risk = 4.0) than a non-abusing parent, but the confidence interval (minimum and maximum values between which the true parameter lies) has its minimum value at 1.15 and its maximum value at 15.0. These numbers suggest that our best estimate of risk (maximum likelihood) is that substance-abusing parents are four times as likely to reabuse. However, any parent who is a substance abuser may be just slightly more likely to reabuse (Odds Ratio = 1.15) or be up to 15.0 times as likely to reabuse (Odds Ratio = 15.0). In addition, effect size must be considered, but is often not (Dar, Serlin, & Omer, 1994). When a statistically significant result on an item in a risk assessment instrument is found, the next question should be, "Does it make a real world difference in predictive capacity?" For instance, in our example above, the risk of reabuse is significant and shows a strong trend. However, if the baseline group against which substance abusers are being com-

pared are "not at all likely" to reabuse, then four times "not at all likely" may remain "not at all likely" in real world terms.

Severity As a Problem

Yet another criticism of CPS risk assessment instruments is inclusion of a criterion for severity of abuse, even though it may not have any predictive validity. Although severity of abuse is listed in almost every consensus-based and actuarial model, there is little or no indication that it is related to recurrence of abuse (Camasso & Jagannathan, 2000; Wald & Woolverton, 1990). This may be due, in part, to the high likelihood that the most severe cases result in the most severe interventions (placement), thereby eliminating the possibility of future harm (again, we come back to the intervention effect). If this is the case, severity of abuse should be left out of the model once intervention has occurred.

An examination of the severity of reabuse (as opposed to severity of initial abuse) may be more fruitful. When examining recurrence, however, severity of abuse is not usually addressed. Most studies use rate of re-report, rate of re-substantiation, and rate of reentry to foster care as indicators of reabuse. These measures do not address severity. The only prevalent post-intervention severity measure in the literature appears to be child death (Gober, Baumann, Graham, & Kern, 1998; Lindsey, 1994), which is such a rare occurrence that it may be unfeasible to build a risk assessment model using it as an outcome criterion (Macdonald & Macdonald, 1999). Also, using child death as the only measure of severity ignores a larger subset of children who are injured but not killed. It can be argued that any type of maltreatment is unacceptable and should be treated in the same fashion regardless of severity. This argument, however, seems illogical in the face of poorly defined maltreatment criteria, the negative effects children may experience as a result of intervention, and the poor outcomes faced by children who enter the foster care system. In the absence of an actuarial risk assessment tool that successfully incorporates severity, responses to severity of abuse must be based on moral or administrative criteria (Wald & Woolverton, 1990).

Sensitivity and Specificity

Despite being an improvement over clinical and consensus-based models, actuarial models are rarely able to predict reabuse at acceptable levels of sensitivity (correctly classifying those children who will be reabused). There is an inverse relationship between sensitivity and specificity (correctly classifying those children who will not be reabused). That is, as the sensitivity of an instrument is increased, its specificity decreases. High sensitivity results in a high percentage of True Positives (children identified as high risk who subsequently experience reabuse) as well as a high percentage of False Positives (children identified as high risk who do not subsequently experience reabuse). High specificity results in a high percentage of True Negatives (children identified as low risk who do not subsequently experience reabuse) as well as a high percentage of False Negatives (children identified as low risk who subsequently experience reabuse).

Assigning risk involves establishing cut points—a value at which a different decision is made above and below that value. These cut points can be purposely or inadvertently changed by administrative policy. For example, the media coverage surrounding a child abuse fatality might prompt a shift to “conservative” policies (lowering of cut points), resulting in an increase in the number of false positives. These administrative shifts are of greatest concern when heightened response is fueled by the occurrence of a rare event (such as child fatality) as the number of false positives will likely increase to the point of being untenable (for a more detailed discussion of sensitivity and specificity, see Gray, 1997). However, attending to false positives and false negatives may not be the best way to assess the efficacy of risk assessment tools (see Baird & Wagner, this issue).

Implications for the Design and Use of Risk Assessment Tools

Concerted and collaborative efforts on the part of researchers, child welfare administrators, and line staff will be required in order to address challenges to the development of valid risk assessment tools.

Increasing Reliability

Researchers, while leaving room for innovation, should standardize the operational definitions of key variables included in studies of risk to

provide a more easily analyzable body of research. Although actuarial models tend to be the best predictors of future maltreatment, they are far from perfect. The role of staff is to gather and input reliable, valid information. As with all statistical tools, the result is only as good as the accuracy of the data included. Many individual risk items are open to interpretation and are based upon clinical decisions. Each interpretation and action is subject to error. Individual risk measures such as "level of social or familial support," "seriousness of injury," and "severity of abuse," require judgments that may be influenced by the availability and representative heuristics.

By knowing the overall consequence of each individual risk measure, a user may either underrate or exaggerate individual items in order to have the risk instrument arrive at the conclusion he/she has made a priori about the maltreatment under investigation. In addition, protective service workers make decisions about the authenticity of claims made by parents, family members, abuse reporters, and others as they input items into the risk assessment instrument. They must not only judge whether the abuse/neglect occurred, but must also judge the accuracy of answers to items on the risk assessment instrument. Once risk has been estimated, clinical skills are used to select and carry out service plans, identify other needs, and evaluate progress. These tasks create other sources of risk.

Clearly Describe the Reliability and Validity of All Measures

Authors should clearly describe the reliability and validity of measures used and clearly inform readers when measures are of unknown reliability and validity. They should describe the kind of reliability and validity investigated and give facts and figures regarding findings (e.g., sizes of correlations, samples sizes, and sample descriptions). If stability is reported, time between administrations of the measures should be given. If criterion validity is explored (either concurrent or predictive) authors should describe the "criterion measures" used, size of correlations, and source and size of samples. The aim is to provide sufficient information so that readers can review the reliability and validity of scales used for themselves.

Considering Baserates

Consideration of the baserates of relevant events in the general population should be incorporated into risk assessment models. For instance,

child injuries occur in non-CPS involved families and the rate of occurrence in the child welfare population is misleading if a comparable rate of occurrence in the non-involved population is not considered in statistical calculations.

Temporal Considerations

Relating factors to outcome is insufficient without measuring length of exposure to risk (Breslow & Day, 1980; Zuravin, 1999). Simply taking a cross-section of child welfare cases, random or otherwise, counting the numbers of children who are reabused, and comparing risk factors can lead to inaccurate results as a consequence of sampling bias. Families involved in the child welfare system for longer periods of time would have a greater probability of study inclusion than would families entering and exiting care more quickly (see Figure 1 depicting a cross-sectional study), yet these two types of families may be very different. The characteristics of a family involved in the child welfare system for 30 days without an incidence of reabuse would be considered in the same manner as the characteristics of a family involved in the child welfare system for 5 years without an incident of reabuse. Retrospective cohort studies represent a viable method for temporal measurement, and the growing number of large administrative data sets seems to be the best way to take time into account through statistical techniques such as event history and time series analysis (see Camasso & Jagannathan, 2000; Simpson, Imrey, Geling, & Butkus, 2000). Weighing length of exposure may also reveal some of the unknown dynamics of reabuse, which may have implications for prevention and treatment (see, for example, DePanfilis & Zuravin, 1998).

Attending to Strengths

Although there is a great deal of emphasis on family strengths in the child welfare literature, this emphasis is often lost in risk assessment models. Protective influences may interact with identified risk factors to minimize the likelihood of negative events (English, 1998; Macdonald & Macdonald, 1999). Certain variables may predict differently in the presence of other variables. For example, the presence of a young child in the home, when considered individually, may not translate to a high risk of reabuse. If the child also has a serious medical condition, the combination of these factors may create a higher risk level than each factor taken individually.

However, if the caregiver has a strong, positive attachment to the child or a strong family support system, risk may be substantially diminished.

Further, risk may not be additive (i.e., adding deficits and subtracting strengths), but may be multiplicative (i.e., a specific combination of risk factors modifies their individual effect, increasing or decreasing risk in different ways) or have some other non-linear function (for a good discussion, see Selvin, 1996). The possibility of this type of interaction among predictor variables highlights the strength of actuarial models as it is highly unlikely that an unassisted individual could accurately carry out these types of calculations, especially given time constraints. Interaction should be explored in greater detail to take advantage of the benefits of statistical models for estimating risk. The availability of larger data sets should enable more accurate assessment of interactive terms.

Establishing Clinical Overrides

Use of risk assessment measures may introduce errors, and these should be addressed. For example, workers may rely solely on the instrument and ignore other relevant evidence that the instrument does not address. Family strengths may be ignored if the tool is deficit-based. A tool cannot cover everything and a worker and/or supervisor should be able to override it for good cause.

“Increased automation does not cure the human factors problem, it simply changes its nature. Systems become more opaque to their operators. Instead of causing harm by slips, lapses, trips, and fumbles, people are now more prone to make mistaken judgments about the state of the system” [i.e., that the system is infallible] (Reason, 1995, p.44).

Overrides should be tracked to establish their reliability and validity in comparison with the model as part of the model improvement process. However, we should only override actuarial models when we know the probability of the outcome is close to zero. Otherwise, we will tend to make more errors in the opposite direction and the two will not balance out (Grove & Meehl, 1996). For instance, if a risk assessment instrument places a child at low risk and there is an overriding safety concern not addressed by the tool, high-risk action should be taken. This decision, how-

ever, should only be made if subsequent abuse is almost sure to occur, as the probability of error is extremely high.

Statistical Concerns

Although certain variables may be significant when considered individually, they may also be highly correlated and significance may fade when they are included in a multivariate analysis. They are, essentially, explaining the same thing twice. Even so, an argument might be made that a greater number of predictor variables results in a model with greater overall predictive accuracy. However, the real world need for a simple, parsimonious instrument necessitates the inclusion of only variables that significantly contribute to a tool's predictive capacity. Each variable must maintain its significant contribution in the presence of other variables in order for the risk assessment instrument to be efficient.

While it is appropriate to use logistic regression, caution in interpretation is warranted when results are reported in terms of relative odds ratios. A relative odds ratio only compares one group of people with a certain set of characteristics to another group of people with a different set of characteristics within the same data set. It is not a measure of the prevalence of reabuse but is, rather, a comparison of the likelihood that one group will be reabused as opposed to another (or multiple others). For example, an odds ratio of 1.4 would not tell us that there is a 40 percent chance that a certain child will be reabused within a population. Instead, it tells us that a child with x, y, and z characteristics in the sample was 40 percent more likely to be reabused than a child without these same characteristics. The overall probability that a child will be reabused within a population is not found in this number. Thus, when the research question and study design call for it, the extra step should be taken to translate findings into marginal probabilities.

Statistical bias is also a concern. Although relative risk between groups can be estimated using logistic regression, results are reported in terms of relative odds ratios (e.g., the odds are 4-1 that a child will be reabused given this set of characteristics). Relative odds ratios, however, always overestimate relative risk if the true relative risk is greater than 1 (more likely) and underestimate the relative risk if the true relative risk is less than 1 (less likely). These terms are only equivalent when there is no association (true relative risk = 1). This estimation bias becomes more pronounced as the risk of occurrence in the population increases and is

substantially biased when the risk of occurrence in the exposed and unexposed groups is greater than 0.1 (Selvin, 1996). Significant findings of elevated risk are also suspect if the resulting ratios are not of a certain magnitude. Some leading epidemiologists have expressed skepticism about relative risk or odds ratio findings of less than 3.0 or 4.0 without replication, even if the study and analysis are conducted correctly (Taubes, 1995). Thus, the magnitude of the findings must be weighed along with effect size and implications for risk assessment.

Implementation Concerns

Developing a valid risk assessment tool may be easier than implementing it effectively. As with any new technology, maximizing hoped-for outcomes (including appropriate implementation) is fraught with difficulty. A clumsy implementation process unattuned to the culture of the agency may undermine a tool's use, despite its ability to predict well. The development of risk assessment tools should include key stakeholders (including line staff and middle management) to increase staff buy-in and to gather site-specific information that may augment the effective use of the tool. Risk assessment tools should be implemented in a context of careful evaluation, including review of the fidelity of the implementation process. Doueck, Bronson, and Levine (1992) suggest a multi-faceted approach that attends to the impact on the organization and individual practice, as well as outcomes for families. Accurate records should be kept and translated into information of value in maximizing the appropriate use of risk assessment tools. User-friendly computer systems should be in place that permit ease of data entry and review. All staff should be trained in the proper use of the tool and proficiency tests should be given to check the effectiveness of training (e.g., in enhancing reliability). Instruction in critical thinking may also be of value in successful implementation (e.g., teaching staff how to avoid common thinking errors and how to critically appraise practice-related claims). Risk assessment tools should be developed and implemented in an environment that supports consistency, constructive criticism, and accountability at all staff levels, as well as with contracted service providers, to increase safety and reduce risk to children, families, child welfare staff, and involved agencies. This will require a comprehensive risk management program designed to minimize risks and maximize benefits (see Gambrill & Shlonsky, forthcoming). In sum, there is much we do not yet understand.

Overview of Contents—Part I

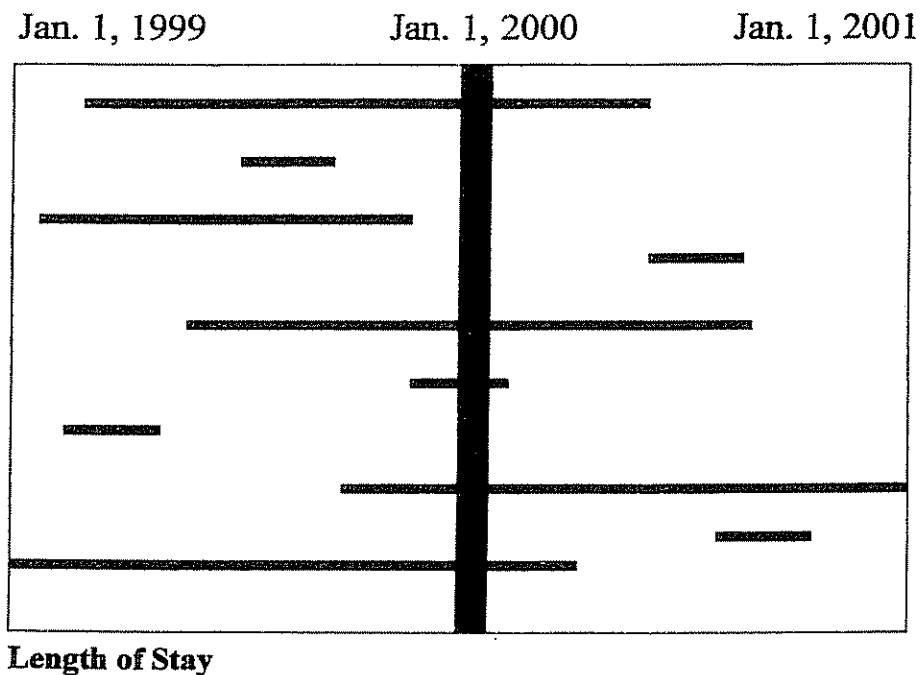
In its current form, the principal function of Child Protective Services is to protect children from harm. Thus, identifying risk of maltreatment is of paramount import. However, the issues raised in this article illustrate that, despite the dramatic growth in the number of children in out of home placements during the 1980's and early 1990's (American Public Human Services Association, 1998; Tatar, 1994), this component of the field is still in its infancy. The articles contained in this issue address some of the concerns discussed earlier.

We are delighted to include as the first article in Risk Assessment part I "The relative validity of actuarial-and consensus-based risk assessment systems" by Christopher Baird and Dennis Wagner (2000). This paper describes the results of a comparison of three risk assessment instruments; two consensus-based and one actuarial-based. This is another finding among a large number of research reports showing the superiority of actuarial compared to clinical and/or consensus-based prediction. This trend is also supported by the work of Michael Camasso and Radha Jagannathan (2000), who investigated the reliability of a widely used consensus based risk assessment instrument, the Washington Risk Assessment Matrix (WRM). The authors found that the WRM exhibited high levels of measurement error and was insensitive to trends over time. This study raises concerns about the premature diffusion of risk assessment instruments characterized by high levels of measurement error.

Diana English and Christopher Graham (2000) compared CPS worker assessment of risk on nine risk factors with CPS clients' self-reports on measures of similar risk constructs. Her findings indicate that, although there is agreement in some areas, certain risk factors may be perceived differently by clients and may be measured differently depending on the instrument used. While there is concern about the validity of some of the measures used, these differences raise questions about the current interpretation and design of risk assessment instruments. Charles Lyle and Elliott Graham (2000), program evaluators in a human services department, describe the use of a risk assessment instrument to measure service outcomes using a pre-post design. They found that decreases in risk ratings may have been influenced by the practice context rather than by actual decreases in risk of maltreatment, highlighting the importance of considering such influences when evaluating the effectiveness of risk assessment instruments. The possibility of bias related to length of exposure is also

addressed in this issue. Children entering the child welfare system are often exposed to varying episodes, durations, and intensities of risk. These different time periods and intensities are often overlooked, resulting in incorrect estimates. Simpson, Imrey, Geling, and Butkis (2000) show the effect of such biases and provide a statistical method for considering them.

Figure 1
Variable Length of Exposure



In this chart, only those cases open for services on January 1, 2000 (represented by the vertical line) would be included in an analysis. Cases open for longer periods of time are much more likely to be included in any cross-section of cases taken at a single point in time because there is a greater opportunity for selection.

We have encouraged our contributors to clearly describe the reliability and validity of measures central to their reports. Such information is often missing. That is, a measure is simply declared to be reliable and valid (e.g.,

“well validated,” of “established reliability”) without the provision of related facts and figures. We have requested contributors to clearly describe the exact nature of the sample used and the population from which it was derived. This can be complex in child welfare, as many inter-related decisions are made (e.g., substantiated, unsubstantiated). We suggest that authors who use a sample that is complexly related to a population present a tree diagram describing exact derivation of their sample. (See Figure 1 of DePanfilis and Zuravin, in press, designed by Aron Shlonsky during the editorial process). We have also encouraged (and sometimes requested) our authors to include additional information clearly describing the statistical procedures they used (e.g., event history analysis). We are grateful to our authors for allowing us to review their work and thank them for their patience in increasing the clarity of their methodology.

References

- American Public Human Services Association (1998). Analysis of State Child Welfare Data: VCIS Survey Data From 1990 Through 1994. <http://www.act.dhhs.gov/programs/cb/stats/index.htm#VCIS>.
- Baird, C., & Wagner, D. (2000). The relative validity of actuarial and consensus-based risk assessment systems. *Children and Youth Services Review, 22*, 839-871.
- Baird, C., Wagner, D., Healy, T., & Johnson, K. (1999). Risk assessment in child protective services: Consensus and actuarial model reliability. *Child Welfare, 78*(6), 723-748.
- Berkowitz, S. (1991). *Key findings from the state survey component of the study of high risk child abuse and neglect groups*. DC, Washington: National Center on Child Abuse and Neglect.
- Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Vol 1. the analysis of case-control data*. Lyon International Agency for Research on Cancer.
- Camasso, M. J., & Jagannathan, R. (1995). Prediction accuracy of the Washington and Illinois risk assessment instruments: An application of receiver operating characteristic curve analysis. *Social Work Research, 19*(3), 174-183.
- Camasso, M.J., & Jagannathan, R. (2000). Modeling the reliability and predictive validity of risk assessment in child welfare. *Children and Youth Services Review, 22*, 873-896.
- Costin, L. B., Karger, H. J., & Stoesz, D. (1996). *The politics of child abuse in America*. New York: Oxford University Press.

- Dar, R., Serlin, R., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*(1), 75-82.
- Dawes, R. M. (1988). *Rational choice in an uncertain world*. Orlando: Harcourt, Brace, Jovanovich.
- Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: Free Press.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.
- DePanfilis, D., & Zuravin, S. J. (1998). Rates, patterns, and frequency of child maltreatment recurrences among families known to CPS. *Child Maltreatment, 3*(1), 27-42.
- DePanfilis, D., & Zuravin, S.J. (in press). Assessing risk to determine the need for services. *Children and Youth Services Review*.
- Dingwall, R., Eekelaar, J., & Murray, T. (1983). *The protection of children: State intervention and family life*. Oxford, England: B. Blackwell.
- Doueck, H. J., Bronson, D. E., & Levine, M. (1992). Evaluating risk assessment implementation in child protection: Issues for consideration. *Child Abuse and Neglect, 16*(5), 637-646.
- Drake, B. (1996). Predictors of preventive services provision among unsubstantiated cases. *Child Maltreatment: Journal of the American Professional Society on the Abuse of Children, 1*(2), 168-175.
- DePanfilis, D., & Zuravin, S.J. (in press). Assessing risk to determine need for services. *Children and Youth Services Review*.
- English, D. (1998). Evaluation and risk assessment of child neglect in public child protective services. In H. Dubowitz (Ed.), *Neglected children: Research, practice, and policy*. Thousand Oaks: Sage.
- English, D. J., Marshall, D. B., & Orme, M. (1999). Characteristics of repeated referrals to child protective services in Washington State. *Child Maltreatment: Journal of the American Professional Society on the Abuse of Children, 4*(4), 297-307.
- English, D.J. & Graham, J.C. (2000). An examination of relationships between children's protective services social worker assessment of risk and independent longscan measures of risk constructs. *Children and Youth Services Review, 22*, 897-934.
- Falco, G., & Salovitz, B. (1997). *Clinical versus actuarial risk assessment in child protective services: Results from recent research in New York*. Paper presented at the 11th Annual CPS Roundtable on Risk Assessment, San Francisco.
- Fuller, T.L. & Wells, S.J. (in press). Predictors of maltreatment occurrence at two milestones in the life of a case. *Children and Youth Services Review*.
- Gambrill, E. (1990). *Critical thinking in clinical practice: Improving the accuracy of judgments and decisions about clients*. San Francisco: Jossey-Bass.

Gambrill, E. & Shlonsky, A. (in press). The need for comprehensive risk management systems in child welfare. *Children and Youth Services Review*.

Gibbs, L., & Gambrill, E. (1999). *Critical thinking for social workers: Exercises for the helping professions*. (2nd edition). Thousand Oaks, CA: Pine Forge Press.

Gober, K. J., Baumann, D. J., Graham, J. C., & Kern, H. (1998). *The Texas child fatality study: A comparison of fatality and non-fatality cases* : Texas Department of Protective and Regulatory Services.

Gray, J. A. M. (1997). *Evidence-based healthcare: How to make health policy and management decisions*. New York: Churchill Livingstone.

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

Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie, & D.L. Medin, (Eds.), *Decision making from a cognitive perspective*. New York: Academic.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes*. (2nd ed.). Boston: Houghton Mifflin.

Johnson, W. (1992). *Clinical versus actuarial risk assessment predictions in emergency response*. Paper presented at the Sixth National Roundtable on CPS Risk Assessment, San Francisco, CA.

Lindsey, D. (1991). Factors affecting the foster care placement decision: An analysis of national survey data. *American Journal of Orthopsychiatry*, 61, 272-281.

Lindsey, D. (1992). Reliability of the foster care placement decision: A review. *Research on Social Work Practice*, 2(1), 65-80.

Lindsey, D. (1994). Mandated reporting of child abuse fatalities: Requirements for a system to protect children. *Social Work Research*, 18(1), 41-54.

Lykken, D.T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W.M. Grove (Eds.), *Thinking clearly about psychology. Essays in honor of Paul Meehl. Vol. I: Matters of public interest* (pp. 3-39). Minneapolis, MN: University of Minnesota Press.

Lyle, G., & Graham, E. (2000). Looks can be deceiving: Using a risk assessment instrument to evaluate the outcomes of child protective services. *Children and Youth Services Review*, 22, 935-949.

Lyons, P., Doueck, H. J., & Wodarski, J. S. (1996). Risk assessment for child protective services: A review of the empirical literature on instrument performance. *Social Work Research*, 20(3), 143-155.

Macdonald, K.I., & Macdonald, G.M. (1999). Perceptions of risk. In P. Parsloe (Ed.), *Risk assessment in social care and social work*. Philadelphia Jessica Kingsley.

Margolin, L. (1997). *Under the cover of kindness: The invention of social work*. Charlottesville, VI: University of Virginia Press.

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

Meehl, P. E. (1973). Why I do not attend case conferences. In P. E. Meech (Ed.), *Psychodiagnosis: Selected papers* (pp.225-302). Minneapolis, MN: University of Minnesota Press.

Needell, B., Webster, D., Cuccaro-Alamin, S., Armijo, M., Lee, S., & Brookhart, A. (2000). *Performance indicators for child welfare services in California project*.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Edgewood Cliffs, NJ: Prentice-Hall.

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

Pelton, L. (1989). *For reasons of poverty: A critical analysis of the public child welfare system in the United States*. New York: Praeger.

Reason, J. (1995). Understanding adverse events. In C. Vincent (Ed.), *Clinical risk management* (pp. 31-54). London: BMJ.

Rose, S. J., & Meezan, W. (1996). Variations in perceptions of child neglect. *Child Welfare*, 75(2), 139-160.

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

Ruscio, J. (1998). Information integration in child welfare cases: An introduction to statistical decision making. *Child Maltreatment: Journal of the American Professional Society on the Abuse of Children*, 3(2), 143-156.

Selvin, S. (1996). *Statistical analysis of epidemiologic data*. (2nd ed.). New York: Oxford University Press.

Simpson, D.G, Imrey, P.B., Geling, O., & Butkus, S. (2000). Statistical estimation of child abuse rates from administrative databases. *Children and Youth Services Review*, 22, 951-971.

Tatara, T. (1994). Some additional explanations for the recent rise in the U.S. child substitute care population: an analysis of national child substitute care flow data and future research questions. In R. Barth, J. D. Berrick, & N. Gilbert (Eds.), *Child Welfare Research Review* (Vol. 1, pp. 126-145). New York: Columbia University Press.

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

Taubes, G. (1995, July 14). Epidemiology faces its limits. *Science*, 269, 164-169.

U.S. Department of Health and Human Services - National Center on Child Abuse and Neglect. (1997). *Child maltreatment in 1995: Reports from the states*

to the national child abuse and neglect data system. Washington, D.C.: U.S. Government Printing Office.

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

Wells, S. J. (1988). Factors influencing the response of child protective service workers to reports of abuse and neglect. In G. T. Hotaling, D. Finkelhor, J. T. Kirkpatrick, & M. A. Straus (Eds.), *Coping with family violence: Research and policy perspectives* (pp. 238-256). Newbury Park, CA: Sage Publications, Inc.

Wolock, I., Sherman, P., & Feldman (in press). Child Abuse and Neglect Referral Patterns: A longitudinal analysis. *Children and Youth Services Review*.

Zuravin, S. (1999). Child neglect: A review of definitions and measurement research. In H. Dubowitz (Ed.), *Neglected children: Research, practice, and policy* (pp. 24-46). Thousand Oaks: Sage.

Zuravin, S. J., Orme, J. G., & Hegar, R. L. (1994). Predicting severity of child abuse injury with ordinal probit regression. *Social Work Research*, 18(3), 131-138.

