Journal of Family Violence, Vol. 19, No. 2, April 2004 (© 2004)

Severity of Partner and Child Maltreatment: Reliability of Scales Used in America's Largest Child and Family Protection Agency

Amy M. Smith Slep[1,2,3] **and Richard E. Heyman**[1,2]

This paper describes two studies investigating the interrater agreement of severity scales for family maltreatment used in America's largest child and family maltreatment agency: the U.S. military's Family Advocacy Program (FAP). The USAF-FAP Severity Index is a multidimensional rating system for clinicians' evaluations of the severity of seven forms of family maltreatment: partner physical, emotional, and sexual abuse; child physical, emotional, and sexual abuse; and child neglect. The first study evaluated the reliability of the scale as it is used in the field. The second study compared a generalizable sample of clinicians' ratings to an established "gold standard" of what the ratings should have been. The Severity Index demonstrated fair-to-good levels of reliability, suggesting that with minimal cost, investigating caseworkers can routinely assess, and make fairly reliable ratings of, the severity of seven forms of family maltreatment for each case they investigate.

KEY WORDS: partner abuse; child abuse; child neglect; child maltreatment; family violence; severity; reliability.

Over 900,000 children, or 13.9 per 1,000, are estimated to be maltreated each year in the United States (U.S. Department of Health and Human Services, 2000), with neglect (53.5%), physical abuse (22.7%), and sexual abuse (11.5%) being the most prevalent. The effects of child maltreatment are well documented: maltreated children suffer from medical problems, cognitive deficits, behavioral problems, and socioemotional deficits (see review by the American Psychological Association working group on Child Abuse and Neglect, Becker *et al*., 1995). Partner abuse is also prevalent. In nationally representative surveys, approximately 12% of women report being physically victimized by their partners during the past year, including 5% (or about 2 million women per year) who report severe victimization (Straus & Gelles, 1990) and 1.2% who reported being sexually abused by their partners during the past year. The adverse mental health consequences of partner abuse are myriad, including heightened risk for Major Depressive Disorder (Cascardi *et al.*, 1995), elevated depressive symptomatology (e.g., Stets & Straus, 1990; Vivian & Langhinrichsen-Rohling, 1994), and Posttraumatic Stress Disorder (PTSD; e.g., Cascardi *et al*., 1995).

Much of what we know about both the risk factors for child and partner maltreatment (for recent reviews of the risk factor literatures see Black *et al*., 2001a,b,c,d; Schumacher *et al.*, 2001a,b,c) and the consequences of child and partner maltreatment comes from research contrasting groups with and without maltreatment. In these contrasted group studies, the operationalization of maltreatment is almost universally limited to a binary, "yes/no" decision. Knowing that a child or woman has been classified as "maltreated" does tell us that he or she is at elevated risk for a number of negative outcomes. However, such a classification does not capture the severity of the maltreatment, which, in theory, may be more highly linked to risk outcomes and indicative of different risk factors than would presence/absence (e.g., O'Leary, 1993). For

[1]Department of Psychology, State University of New York at Stony Brook, Stony Brook, New York.
[2]Both authors contributed equally to this article.
[3]To whom correspondence should be addressed at Department of Psychology, State University of New York at Stony Brook, Stony Brook, New York 11794-2500; e-mail: amy.slep@sunysb.edu.

0885-7482/04/0400-0095/0 © 2004 Plenum Publishing Corporation

example, the severity, not just the occurrence, of child maltreatment may factor greatly into caseworkers' decisions about whether to remove the child from the home and into therapists' decisions about appropriate treatment modalities for the parent(s). Similarly, severity of partner abuse may factor into arrest and treatment decisions. For example, unhappily married couples with infrequent, mild partner abuse may be unlikely to face arrest and may be appropriate candidates for partner abuse programs conducted either individually or conjointly (e.g., Heyman & Neidig, 1997; O'Leary *et al.*, 1999), whereas couples with severely abusive husbands may be better candidates for incarceration and shelter. Empirical support for dimensional, rather than categorical, distinctions can be gleaned from differences in risk factors for seven forms of family violence depending on whether the maltreatment was mild or severe (Black *et al.*, 2001a,b,c,d; Schumacher *et al.*, 2001a,b,c).

Researchers have called for the use of continuous indices of maltreatment and for methods that identify the multiple types of maltreatment that may have occurred (e.g., McGee & Wolfe, 1991). In response to this acknowledged need, systems for rating the severity of different types of child maltreatment have been developed and published. However, these measures have been targeted at researchers, involving ratings based on a structured interview (Chaffin *et al.*, 1997) or information collected from a number of sources (e.g., case records and observations: Kaufman *et al.*, 1994; McGee *et al.*, 1995).[4] Clearly, scales that reliably yield information regarding types and severity of maltreatment in families are important additions to our research assessment resources.

Yet the published measures described above were designed for research purposes and are therefore perhaps better suited for clinical research than for everyday clinical use. For a measure to be truly useful to caseworkers and clinicians, it must be straightforward and should not require information that is difficult to obtain or that requires substantial time to collect. This is not to suggest, however, that such a measure should be simplistic. To be clinically informative, a maltreatment severity measure must capture the multifaceted nature of a form of maltreatment and the myriad ways that facets can combine to produce a mild, moderate, or severe maltreatment incident. Such a measure (simple to use but complex in content) of child and partner maltreatment has been used for several years by caseworkers in America's largest child and family maltreatment agency, the U.S. military's Family Advocacy Program (FAP).

One of FAP's missions is to function as the child protective services agency within military installations. Child maltreatment allegations are brought to FAP's attention, and FAP investigates to determine whether the allegation is substantiated. When the allegation is against a civilian family member, investigation, substantiation decisions, and protective custody decisions are the jurisdiction of the civilian child protective service (CPS) agency.[5] Unlike state CPS agencies, FAP is also charged with investigating spouse maltreatment allegations involving a service member. A second important difference is that FAP is explicitly charged with tracking the prevalence of maltreatment in military communities, developing and offering family maltreatment primary and secondary prevention programs, and providing treatment to both victims and perpetrators of maltreatment. Thus, for both child and partner maltreatment, FAP investigates maltreatment allegations, recommends case dispositions, and treats offenders and victims. (For a detailed discussion of FAP's response to child maltreatment, including detailed statistics on prevalences, see Mollerstrom *et al.*, 1995.)

That FAP handles both investigation and treatment—for child and/or spouse maltreatment—necessitates a broad perspective. This perspective is what precipitated the development and widespread implementation of a system for rating pan-maltreatment severity. The U.S. Air Force (USAF) FAP's Family Violence Severity Index is a scale used to quantify the severity of each type of maltreatment that they investigate: partner physical, emotional, and sexual abuse; child physical, emotional, and sexual abuse; and child neglect. The Severity Index is a grid (see Appendix), with rows for each of the seven forms of family maltreatment and columns for the five levels of possible severity. Severity is rated as "none," "mild," "moderate," "severe," and "death." Operational definitions for each severity level of each form of maltreatment are contained within the cells of the Severity Index's grid. FAP clinicians make a rating for each form of family violence that was substantiated (i.e., the severity of all forms of maltreatment that were not substantiated is assumed

[4]For spouse abuse, no comparable scales exist. The Conflict Tactics Scale (Straus, 1979; Straus *et al.*, 1996), the most widely used measure, has *a priori* defined mild and severe behaviors. Although providing some degree of dimensionality, the mild/severe distinction is too limited to give a true sense of the continuous phenomenon of abuse severity.

[5]FAP usually works in conjunction with the state CPS agency in child maltreatment cases, as only CPS has the authority to remove children from the home and proceed with legal action to terminate parental rights.

to be "none"). For example, if only physical child abuse was substantiated, then the only severity rating would be for child physical. If, however, both physical and sexual child abuse were substantiated, then both of these forms of maltreatment would receive a severity rating. In a sense, severity ratings are dependent on decisions to substantiate, because if a particular form of maltreatment is not formally substantiated, its severity will not be rated (i.e., it is presumed to be "none"). The Severity Index was originally developed over a decade ago, and has been reviewed by a number of committees including FAP clinicians, treatment managers, and program managers to refine the operationalizations to capture clinical decision-making as clearly as possible. The current version of the Severity Index has been in use worldwide for over 4 years. Other than providing clinicians with a copy of the measure, and the review of severity ratings as part of routine case review meetings, no special training is given in the application of the Severity Index to cases.

The operationalized definitions contained in the Severity Index combine several clinically relevant domains, which are summarized in Table I. The multiple foci of the Severity Index provide complex, multidimensional operationalizations of each level of severity for each form of maltreatment. The complexity and variability of the operationalizations make investigating the reliability with which these operationalizations can be applied particularly important. Although clinical presentations of family maltreatment tend to be complex, it is typically extraordinarily difficult to establish interrater agreement for complex, multidimensional operationalizations. This is usually true even when such a measure is used under tightly controlled conditions (e.g., a university laboratory). It is especially true for a scale that will be used under "real world" conditions (i.e., across many sites and raters) and without formal training.

Table I. Domains Used in Rating Severity of Maltreatment

Type of Abusive Behavior
Verbal threats
Physical contact that does not involve oral, vaginal, or anal penetration
Physical contact involving oral, vaginal, or anal penetration
Severity of injury (minor[a] or major[b])
Medical treatment
Sought
Indicated
Type
Short-term
Long-term
Inpatient
Mental health treatment
Short-term
Long-term
Repetitiveness of alleged abusive behavior
Potential harm of the alleged abusive behavior
Alternate placement of child or spouse

[a] DoD Instruction 6400.2 defines minor injury as "twisting, shaking, minor cut, bruise, welt, or any combination thereof, which do not constitute a substantial risk to the life or well-being of the victim."

[b] DoD Instruction 6400.2 defines major injury as "brain damage, skull fracture, subdural hemorrhage or hematoma, bone fracture, dislocations, sprain, internal injury, poisoning, burn, scald, severe cut, laceration, bruise, welt, or any combination thereof, which constitutes a substantial risk to the life or well-being of the victim."

Interrater agreement refers to the consistency (i.e., reliability) of ratings across at least two independent judges. It describes the extent to which different raters consistently interpret and apply the scale definitions. It quantifies to what degree one rater's interpretation of the criteria of "mild," for example, corresponds to a second rater's interpretation of the same criteria. High levels of interrater agreement suggest both that (a) a scale is well-operationalized, and (b) raters have been trained to apply the criteria in the same way.

Although the Severity Index had been used to rate thousands of cases of maltreatment worldwide in the 1990s, the reliability of the scale had not been investigated.[6] In preliminary investigations of the psychometrics of a measure, reliability should be the first focus because reliability constrains the possible validity of a measure. That is, because unreliable measures contain large amounts of error variance, unreliable measures cannot be valid ones (e.g., Wiggins, 1973).

To determine the interrater agreement of the FAP Severity Index, we conducted two studies. The first study involved a select group of FAP clinicians reviewing and making severity ratings on archival FAP case records. This study evaluated the reliability of the scale as it is used in the field. The second study involved having a representative sample of FAP clinicians rate vignettes based on USAF-FAP cases. This allowed us to compare a generalizable sample of clinicians' ratings to an established "gold standard" of what the ratings should have been.

[6] The co-occurring and interactive nature of forms of family maltreatment would suggest that reliability also be established for both (a) ratings of a single form of maltreatment; and (b) simultaneous ratings of multiple forms of maltreatment. Given that the reliability of the Severity Index has not yet been investigated, we will focus on the former.

STUDY 1: RATINGS OF EXISTING CASE RECORDS

Rationale

The purpose of the first study, which compared independent judges' ratings to the original ratings of the severity of maltreatment for actual FAP maltreatment cases, was to estimate the reliability of the severity ratings made in the field. The optimal way to establish the actual "in the field" reliability of ratings would be to have multiple raters assess each case as it presented, and examine the reliability of final ratings. This method was not feasible logistically, however. The approach that came closest to the multiple original ratings method was to assess the reliability of the Severity Index by having independent judges rate the severity of abuse from actual FAP case records. Actual FAP case records (a) reflect actual cases of detected abuse, with all their attendant richness and complexity and (b) include the original severity rating. The clinicians who made the original ratings did not know that their ratings would be evaluated for reliability, thus eliminating a potential source of bias. One potential limitation to this approach is that, because not all levels of severity occur with equal frequency (e.g., mild is more common than severe), the optimal-level reliability of the measure will not be fully assessed.

The reliability of the original ratings is established if the original ratings and independent judges' ratings agree. However, it is also possible that the original clinician had access to additional information that was not included in the formal case record. If this were the case, the original rater and the independent judge might disagree, but only because of access to different information, not because of unreliable implementation of the Severity Index. If this were the case, then ratings made by two independent judges would demonstrate more agreement than ratings made by the original rater and an independent judge. Thus, by using both approaches (i.e., archival ratings versus independent judges, independent judges versus each other) to assess reliability on the same cases, we can obtain a good estimate of the reliability of actual severity ratings as they are made in the field.[7]

[7] Assessing the concordance of scores given by two independent judges separately rating the case record is likely to capture the upper range of ratings as they are made in the field. This is because the independent judges (a) were not blind to the purpose of the study, and thus may have been more attentive to the grid than under normal circumstances and (b) were identified as experienced FAP clinicians who may have more expertise than the average FAP clinician.

Method

Twelve experienced USAF-FAP clinicians in the San Antonio, TX, area were identified and invited to participate in 2 days of case ratings at Brooks Air Force Base. Descriptive information about these raters is detailed as part of Study 2 (see below).

We were provided with the case numbers for all the FAP cases of substantiated child and partner maltreatment from the preceding 2 years from the four bases in the San Antonio area: Brooks, Randolph, Lackland, and Kelly.[8] We sorted these case identifiers by the type(s) of maltreatment that had been substantiated. For forms of maltreatment with more than 25 substantiated cases (e.g., child physical abuse), we randomly selected 25 cases to be rated. For forms of maltreatment with fewer than 25 substantiated cases (e.g., partner sexual abuse), all of the cases were selected to be rated. This resulted in the selection of a total of 184 FAP case records. These records were brought to the case rating site and the original ratings were masked.

All cases were rated by two (of the 12) independent raters at one of two full day rating sessions. Raters were provided with copies of (a) the USAF-FAP Severity Index, and (b) the definitions of maltreatment (Department of Defense, 1987, Instruction 6400.2). Raters who were currently working as clinicians were instructed not to rate any cases that had originated from their bases or with which they were otherwise familiar. All records were reviewed and rated independently. Raters were told that they were participating in an evaluation of the reliability of the Severity Index. Observers were present to ensure that participants did not confer with each other. For cases comprising more than one form of abuse, each form was rated for severity. After two raters independently completed their ratings, the case record was passed to an administrative staff member, who (a) ensured that the case had in fact been rated by two raters, (b) unmasked the original severity rating, (c) recorded it, and (d) prepared the data to send to the authors.[9]

[8] The total number of cases across the bases were as follows: child physical abuse ($n = 49$); child sexual abuse ($n = 18$); child neglect ($n = 23$); child emotional abuse ($n = 17$); child physical/emotional abuse ($n = 4$); spouse physical abuse ($n = 227$); spouse sexual abuse ($n = 18$); spouse emotional abuse ($n = 17$); spouse physical/emotional abuse ($n = 14$); spouse sexual/physical abuse ($n = 1$).

[9] After being received by the authors, data were again checked for completeness. We found approximately eight case records that had data for only one rater. These case numbers were sent to the field coordinator, who arranged to have the cases rated again (by a rater other than the sole prior rater) and the ratings sent the authors.

Table II. Descriptive Statistics for USAF-FAP Case Record Severity Ratings

	Form of partner abuse			Form of child maltreatment			
	Emotional	Physical	Sexual	Emotional	Physical	Sexual	Neglect
Mean	2.61	2.33	0.00	2.40	2.31	3.00	2.52
Standard deviation	0.50	0.57	0.00	0.70	0.67	0.63	0.79
n	18	83	0	10	36	11	23

As stated above, all cases rated had at least one of the seven forms of maltreatment (i.e., partner emotional, physical, sexual abuse; child emotional, physical, sexual abuse; child neglect). Each case record was rated for all forms of family violence, even though most case records suggested that only one or two forms of maltreatment had occurred. In order to avoid inflating our reliability statistics with universal agreement among raters that a particular form of maltreatment did not occur in a particular case (e.g., all scored child sexual abuse as "none"), we removed all instances when both independent judges and the original rater scored a particular form of maltreatment as "none."

RESULTS AND DISCUSSION

Preliminary examination of the data for each form of maltreatment revealed that the distributions were clearly not normal. "Mild" was the most common response, with "moderate" being less common, and "severe" being quite rare. There was only one "*death*" rating. Mean original severity ratings, standard deviations, and the number of records indicating a substantiated case of each form of maltreatment are presented in Table II.

Interrater reliabilities were assessed separately for each form of maltreatment. Intraclass correlations are appropriate for assessing the reliability of rating scale data where the issue is not point-by-point agreement by raters, but rather agreement by raters in rank ordering of cases. In other words, with rating scale data, a "2" is closer to a "3" than to a "5." Statistics that assess point-by-point agreement (e.g., unweighted Cohen's kappa) would consider both of these as disagreements, without indexing the degree of disagreement. Finn's r, a variant of the intraclass correlation, was selected for these analyses because it is not as sensitive to deviations of response distributions from normality as the intraclass correlation (Whitehurst, 1984, 1985; see also Cicchetti, 1985). When both the underlying construct and observers' ratings are evenly distributed, Finn's r and typical intraclass correlations yield equal results. However, if the underlying distribution, and thus the observers' ratings, are significantly skewed, Finn's r accurately indexes reliability whereas intraclass correlations underestimate reliability.[10] Because the severity of maltreatment should be heavily skewed, with many more mild cases than severe cases, we used Finn's r to assess the interrater reliability of severity ratings.

Interrater reliability is presented in Table III in two ways. First, the data from the independent judges rating the archival case records were compared to each other and to the original rating.[11] As shown in Table III, interrater reliability is good for rating the severity of partner physical abuse (Finn's $r = .85$–$.87$) but is poor for partner emotional abuse (Finn's $r = .55$–$.64$). Too few case records of partner sexual abuse were available (i.e., 1 case) to calculate meaningful interrater reliability statistics. The interrater reliabilities reflected are fair to good (Finn's rs between .67 and .85) for all forms of child maltreatment.

To summarize, the reliabilities of severity ratings of partner physical abuse were reflective of relatively good interrater agreement. The reliabilities for child emotional and physical abuse ratings were also adequate, with both agreement between raters and between a rater and the original case rating falling above the standard cutoffs. For child sexual abuse and child neglect, the agreement between the original case rating and the independent rater was adequate, but the agreement between the raters fell

[10] In cases where the underlying distribution of a variable is thought to be even but the distributions of observers' ratings are heavily and similarly skewed, such that this skew reflects a shared rating bias on the part of the observers, intraclass correlations would be preferred over Finn's r. In these circumstances, Finn's r would overestimate unbiased observer agreement (Cicchetti, 1985; Whitehurst, 1985).

[11] Because the original rating was based both on information documented in the case record and on other information (e.g., nonspecific impressions), it might be expected that this rating could differ slightly from ratings based solely on the archival records. Thus, we have decided to present both the original rating-independent judge reliabilities and the rater-rater reliabilities.

Table III. Interrater Reliability (Finn's *r*) for USAF-FAP Case Record Ratings

	Form of partner abuse						Form of child maltreatment							
	Emotional		Physical		Sexual		Emotional		Physical		Sexual		Neglect	
Rater	Finn's *r*	*n*	Finn's *r*	*n*	Finn's *r*	*n*	Finn's *r*	*n*	Finn's *r*	*n*	Finn's *r*	*n*	Finn's *r*	*n*
Rater 1 vs. Rater 2	.64	24	.87	86	.75	1	.78	18	.78	42	.69	12	.67	23
Raters vs. case record rating	.55	51	.85	177	.75	1	.76	30	.79	80	.85	23	.76	46

Note. Cases were included in analysis of a particular form of family violence if either rater or the case management team rated higher than "none." Finn's *r*s $\geq$ .90 are excellent, .80–.89 are good, and .70–.79 are fair.

below this threshold. Finally, severity ratings for partner emotional abuse were not reliable for either method.

Overall, these levels of agreement are encouraging. Moderate, rather than excellent, reliabilities between original ratings and those of independent raters could have been due to a number of factors, including the possibility that the original ratings were based substantially on information about the case that was not reflected in formal case record. If this were the case, then the reliabilities of ratings from two independent judges, both using only the information in the archival record to determine their ratings, would be better than the reliabilities of each judge and the original rating. However, this was not the case. Where the two methods indicated somewhat different levels of agreement, it was as likely that the comparison with original ratings resulted in better agreement. It could be that this moderate level of reliability is inherent in the Severity Index. However, as it is also possible that some of the reliabilities obtained were attenuated by the relative lack of variability in severity (i.e., a preponderance of mild cases) in the case records used in Study 1. For example, the original case ratings of partner emotional abuse have the smallest standard deviation of any of the seven forms of maltreatment, and also had the lowest levels of reliability among the seven forms.

The advantages of using archival case records to assess the interrater reliability of ratings made using the Severity Index (i.e., real world realism) were discussed above. There are four disadvantages of this strategy. First, as just noted, the restricted range of actual maltreatment severity reflected in the records may have attenuated reliability statistics. In other words, lower reliabilities for forms of maltreatment with little variance may be due to statistical, not reliability, difficulties. Second, the small number of cases of some forms of maltreatment may have hindered accurate assessment of their reliabilities, and made the assessment of the reliability of the severity of partner sexual abuse impossible. Third, there was no "gold standard" for what the ratings *should have been*. Thus, a judge who makes an accurate rating (if compared to an idealized "true" rating) may be penalized for not agreeing with the original clinician (who may not have followed the scale as intended). Finally, the sample of independent judges used in Study 1 was not representative of FAP clinicians as a whole, and thus the reliabilities obtained, although predominantly encouraging, cannot be generalized to USAF-FAP clinicians more broadly. This is an inherent limitation of the case record approach, as actual, confidential case records cannot be easily widely distributed, thereby requiring judges to be physically available to make ratings. The purpose of the second study, assessing interrater reliability with standardized vignettes, was (a) to create a standardized, disseminateable measure that would more evenly distribute cases across the severity spectrum, thus reducing statistical problems due to attenuated ranges; (b) to establish the "gold standard" ratings for the measure; (c) to collect severity ratings from a generalizable sample of FAP clinicians; and (d) to compare the clinicians' ratings to gold standard ratings.

STUDY 2: VIGNETTE RATINGS

Rationale

Study 2 aimed to use experimental procedures to build on the advantages and to minimize the disadvantages of the archival case record approach (Study 1). To accomplish this, we developed a series of vignettes based on the case descriptions in the minutes of family maltreatment case management teams from 12 USAF installations. We modified the case descriptions as necessary to ensure that the vignettes had a greater range of maltreatment severity than would a random selection of cases and represented all forms of maltreatment. We then obtained master ratings to use as the gold standard.

Whereas Study 1's case records approach used expert clinicians (because actual case records needed to be

physically available), Study 2's standardized vignette approach allowed us to collect a representative sample of all FAP clinicians as raters. Using vignettes and a larger, more representative sample of FAP clinicians offered several advantages over Study 1. First, a better estimate of the full reliability of the Severity Index (i.e., all levels of severity for all forms of maltreatment) could be obtained because realistic vignettes could be constructed that were based on actual cases, used the full range of non-fatal abuse severity, and could have the "gold standard" answer decided. Second, rater characteristics (e.g., clinical experience, frequency of referencing the severity grid when making ratings) could be compared to raters' deviations from the gold standard ratings. Third, because the raters constitute a representative sample of all FAP clinicians, the reliability estimates could be generalized to the population of FAP clinicians.

Method

Construction of Vignettes

Family maltreatment case management team minutes (*with no identifying information*), from 12 USAF installations were sent to the authors. Initially, all complete case descriptions were considered as potential vignette material. These case descriptions were then reviewed to select relatively representative examples of each form of maltreatment with differing levels of severity. Case descriptions were then edited to ensure that (a) sufficient detail was provided to allow a substantiation decision and severity rating and (b) vignettes adequately reflected the full range of severity of maltreatment from mild to severe. Because "death" is an incontrovertible state that should result in perfectly reliable ratings,[12] we did not create any vignettes with fatal outcomes. Most of the case descriptions required relatively few changes, but a few case descriptions required extensive elaboration or modification. To ensure face validity and content validity,[13] senior FAP headquarters staff reviewed the 21 vignettes and offered suggestions and edits for clarity. We attempted to make the vignettes clear while retaining the complexity of the actual cases on which the vignettes were based.

Master Ratings

Three experienced FAP staff members were selected by FAP headquarters staff to decide on the gold standard ratings. These "master raters" completed their ratings independently. If only two master raters agreed, the majority opinion was retained. In cases where all three master raters disagreed, a committee of four senior USAF expert clinicians determined the final rating.

As shown in Table IV, all three master raters agreed on 104 (70.74%) of the 147 ratings. Ninety-one (61.90%) were agreements that that form of family violence did *not* occur (e.g., there was no child sexual abuse in a vignette that did not mention children); 13 (8.84%) were agreements that (a) a form of family violence was substantiatable *and* (b) a certain level on the Severity Index should be rated. Table IV also breaks out the data by focusing only on individual ratings for which at least one master rater scored a form of maltreatment as substantiatable (i.e., the condition under which a severity rating would be made in the field). Fifty-six items met that criterion, although all three raters agreed on only 13 (23.21%) of them.

Sample Characteristics

Sample 1 (Study One Case Raters). The first sample of raters included in the vignette study comprised eight of the twelve[14] FAP clinicians who participated in Study 1. This sample was selected for two reasons. First, these raters were experienced FAP clinicians who were likely to provide ratings with maximal interrater reliability. Second, the interrater reliabilities from this sample can be compared to those obtained in Study 1. If this sample's reliabilities were to differ between Study 1 and Study 2, the differences could be attributed to differences in the methodology of the studies. If, however, this sample produced similar reliabilities in Study 1 and Study 2, yet the reliabilities differed between Study 1 and the Study 2's Sample 2 (the randomly selected FAP sample, described below), then the differences could be attributed to differences in samples.

[12] It is not always so clear, however, that a child's death was the result of maltreatment. Because severity ratings are only made *after* the case has been substantiated as maltreatment, the severity index rating of death should result in perfect reliability. Substantiation decisions about whether the death was due to maltreatment may be more difficult.

[13] Face validity is " the simplest . . ., which tells whether the measure appears (on the face of it) to [overtly] measure what it is supposed to measure" (Cozby, 1981, p. 59). Content validity is " the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct" (Haynes *et al.*, 1995).

[14] Two study 1 participants helped edit draft vignettes and were therefore excluded from participating in Study 2. Two other raters did not return their packets.

Table IV. Agreement of Master Raters

Agreement	*n*	% of total	% on occurrence
Agreement on non-occurrence	91	61.90	
Agreement on occurrence	13	8.84	23.21
Disagreement = 1	31	21.09	55.36
Disagreement > 1	12	8.16	21.43
Total	147	100.00	100.00

Note. Disagreement = 1 indicates the proportion of ratings where the master raters differed by 1 point in their severity ratings for a particular form of maltreatment for a vignette. Disagreement > 1 indicates the proportion of ratings where the master raters differed by more than 1 point in their severity ratings for a particular form of maltreatment for a vignette.

Sample 1 participants ($n = 8$) had an average of 12.0 years ($SD = 7.44$, range = 1–20) of experience with FAP and 8.28 years ($SD = 4.82$, range = 3–16) of clinical experience and were familiar with maltreatment standards, $M = 3.75$ on a 1 (*somewhat familiar*) to 5 (*extremely familiar*) scale ($SD = 1.28$, range = 1–5). When asked, "When making severity ratings, what percentage of the time do you refer to the abuse and substantiation definitions in the standards manual?" using the following scale (1 = 0–25%; 2 = 26–50%; 3 = 51–75%; 4 = 76–100%), participants responded with $M = 2.43$ ($SD = 0.98$, range of responses given = 1–4). When asked, "When making severity ratings, what percentage of the time do you refer to the severity grid?" participants responded with $M = 3.43$ ($SD = 0.79$, range of responses given = 2–4; scale: 1 = 0–25%; 2 = 26–50%; 3 = 51–75%; 4 = 76–100%). Participants held the following degrees: MSW (20%), MA (40%), DSW (10%), PhD (10%), EdD (10%), and missing (10%). When asked, "At your base, who makes the severity ratings" 71.4% reported that the clinician makes the severity rating and 28.6% reported that the case management team makes the ratings.

Sample 2 (Random FAP). Seventy-five FAP clinicians were randomly selected to participate from a roster of worldwide FAP staff. This sample was chosen to be representative of FAP clinicians who actually make severity ratings on maltreatment cases. Results from this sample provide the best indication of how reliably the Severity Index is used by FAP clinicians. Of the 75 individuals initially contacted to participate, 15 individuals did not participate despite follow-up contacts from the FAP headquarters staff, two formally declined to participate, two moved and became ineligible, and two became master raters. Fifty-four (72%) sent returned surveys within the data collection period. Drew *et al.* (1996, p. 147) wrote that "a 70% response rate can be considered adequate" for a random sample to be considered representative of the population from which it is drawn. Participants had an average of 6.22 years ($SD = 6.22$, range = .83–30) of experience with FAP and 9.65 years ($SD = 7.52$, range = 0–32) of clinical experience; and were extremely familiar with maltreatment standards, $M = 4.02$ on a 1 (*somewhat familiar*) to 5 (*extremely familiar*) scale ($SD = 0.74$, range = 3–5). When asked, "When making severity ratings, what percentage of the time do you refer to the abuse and substantiation definitions in the standards manual?" participants responded with $M = 2.69$ ($SD = 1.21$, range = 1–4) on the following scale (1 = 0–25%; 2 = 26–50%; 3 = 51–75%; 4 = 76–100%). When asked, "When making severity ratings, what percentage of the time do you refer to the severity grid?" participants responded with $M = 3.43$ ($SD = 0.82$, range = 1–4; scale: 1 = 0–25%; 2 = 26–50%; 3 = 51–75%; 4 = 76–100%). Participants held the following degrees: MSW (including CSW and LCSW, 85.3%), MA (11.1%), and DSW (1.9%). When asked, "At your base, who makes the severity ratings" 92.6% reported that the clinician makes the severity rating, 3.7% reported that the case management team makes the ratings; and 3.7% reported "other."

Procedures

Vignette packets, response packets, and cover letters from the Air Force FAP commanding officer and from the university research team were mailed to all the selected participants, along with a return envelope addressed to the university research team. Respondents were informed that they had been selected to participate in a study of "how well the Severity Index works" and were requested to make ratings for each of the 21 vignettes, being certain to rate severity for all forms of family maltreatment applicable to each vignette. Further, they were instructed to refer to any materials they might usually use when assigning severity ratings to actual cases (e.g., maltreatment definitions and severity grid). Finally, respondents were asked not to discuss the materials with anyone until after they had

completed their ratings and to return the response packet, along with any comments, to the university research team.

Two weeks after the deadline, Air Force FAP headquarters staff began making reminder contacts with clinicians who had not returned packets (i.e., had either not completed packets or had not notified us of a desire not to participate). FAP headquarters staff continued their efforts until 50 completed response packets were received.

Results and Discussion

Interrater Reliability

Study 2 used standardized vignettes, a gold standard rating, and a representative sample of raters to estimate the reliability of the Severity Index (i.e., the upper range of reliability with which ratings can be made by FAP clinicians). Interrater reliability is presented in Table V. Comparable levels of reliability for all forms of maltreatment were found for Sample 1 (Study 1 participants) and Sample 2 (representative FAP clinician sample). Where there are even small discrepancies between the reliabilities from the two samples, the representative FAP clinician sample appears to have made more reliable ratings. Partner abuse severity ratings were fairly reliable for all three forms. Child physical and sexual abuse and neglect severity ratings were fairly reliable, whereas child emotional abuse severity ratings were unreliable.

Thus, the overall reliability of ratings based on the Severity Index appears roughly equivalent when evaluated with standardized vignettes (Study 2) and with case records (Study 1). However, the reliability of severity ratings for child emotional abuse do appear better in Study 2 than they were in Study 1 for both samples, perhaps suggesting that for this form of maltreatment, the restricted range of severity reflected in the case records affected the reliabilities obtained in Study 1. The reliabilities of severity ratings for two forms of maltreatment, partner physical abuse and child emotional abuse, were somewhat lower for both samples included in Study 2 than in Study 1. Given that the two samples do not differ, it may be that raters are more reliable in their use of the Severity Index when distinguishing between "none," "mild," and "moderate," than when distinguishing between "moderate" and "severe." The case records from Study 1 suggest that clinicians do have more practice using the Severity Index at the less severe end of the distribution. The severity ratings for the other four forms of maltreatment (i.e., partner sexual abuse, child physical abuse, child sexual abuse, and child neglect) possessed comparable levels of reliabilities across the two studies. Although all Study 2 reliabilities fall short of the standards typically required for a measure to be considered highly reliable, given the complexity of the phenomena and the scale and that no training to facilitate reliability takes place, these reliabilities are encouraging.

Predictors of Reliability

We assessed several descriptive variables to see if they predicted the reliability of ratings (i.e., correspondence between clinicians' ratings and master ratings). These analyses were conducted only on the representative FAP clinician sample (Sample 2) because Sample 1's n was too small. The descriptive variables evaluated as predictors of reliability included years of FAP experience, years of clinical experience, self-reports of perceived familiarity with the DoD standards, self-reports of the proportion of time the rater references the DoD abuse definitions, and self-reports of the proportion of the time raters referenced the severity grid when making severity ratings. The only significant associations were as follows: clinical experience was negatively associated with reliability of ratings of child emotional abuse ($r = -.35$, $p < .01$), familiarity with FAP standards was positively associated with reliability of ratings of partner physical abuse ($r = .28$, $p < .05$) and child emotional abuse ($r = .29$,

Table V. Interrater Reliability (Finn's r) for Severity Ratings of Maltreatment Vignettes

	Form of partner abuse						Form of child maltreatment							
	Emotional		Physical		Sexual		Emotional		Physical		Sexual		Neglect	
Rater	Finn's r	*SD*	Finn's r	*SD*	Finn's r	*SD*	Finn's r	*SD*	Finn's r	*SD*	Finn's r	*SD*	Finn's r	*SD*
Study 1 raters vs. master rating	.7	0.1	.73	0.1	.71	0.3	.57	0.2	.76	0.1	.76	0.2	.72	0.2
FAP raters vs. master rating	.74	0.1	.75	0.1	.83	0.1	.55	0.1	.79	0.1	.87	0.1	.79	0.2

Note. Cases were included in analysis of a particular form of family violence if either the rater or the master rating rated higher than "none."

$p < .05$), and the proportion of the time the rater reported referencing the severity grid was negatively related to the reliability of ratings of partner sexual abuse ($r = -.26$, $p < .05$).

Taken as a whole, these correlations do not suggest that the predictor variables we explored are consistently related to the reliability of ratings of severity. When clinical experience was related to reliability, it was negatively related, suggesting that more experienced raters may be less likely to apply the severity grid definitions as written when making severity ratings. Familiarity with standards and how often the rater referenced the grid were also related to reliability of ratings for at least one form of maltreatment. Although these correlations should all be interpreted with caution and considered exploratory, this finding lends some support to the notion that training and guidelines could assist in increasing the reliability with which ratings are made.

SUMMARY AND CONCLUSIONS

The USAF-FAP Severity Index is a multidimensional rating system for quantifying the severity of seven forms of family maltreatment: partner physical, emotional, and sexual abuse; child physical, emotional, and sexual abuse; and child neglect. All substantiated cases of any form of family maltreatment are rated by a caseworker for presence and severity of all forms of maltreatment.

In contrast to other severity measures that have appeared in the literature, the Severity Index is specifically designed for everyday clinical use. It requires no special training or data collection, and takes only moments to complete. Despite its ease of use, the measure is comprehensive in assessing different forms of maltreatment and complex in its operationalizations of severity. Furthermore, it has been used in tens of thousands of case assessments.

If civilian CPS agencies were routinely to include an assessment of the severity of all forms of family maltreatment for all maltreatment incidents, as the USAF-FAP does, it could greatly enhance our understanding of for whom and under what conditions relevant outcomes (e.g., re-abuse, termination of parental rights) occur. Such findings would be indicators of the measure's construct validity. However, for a measure to be valid, it must first be reliable.

To examine the reliability of ratings based on the Severity Index, we conducted two studies with complementary samples and methods. Taken together, these studies suggest that the Severity Index supports fair-to-good, but not excellent, levels of reliability. It is quite encouraging that fair-to-good levels of agreement could be obtained from masters-level caseworkers' ratings for a scale as complex as the Severity Index. This suggests that with minimal cost (e.g., no extra assessment or training time) investigating caseworkers can routinely assess and make fairly reliable ratings of the severity of seven forms of family maltreatment for each case they investigate.

Because the results were encouraging but not yet perfect, we recommend that work continue on refining and assessing Severity Index reliability. That is, given that the reliability results are typically in the fair to good range despite the complexity of the scale, it is likely that good to excellent reliability levels could be achieved. First, it is likely that a few minor modifications to the Severity Index would greatly enhance the reliability with which the scale could be applied to cases. Severity Index operationalizations are not always written in ways that clarify what a clinician should do when a case falls on the border between two ratings (e.g., it meets some of the criteria for "mild" and some criteria for "moderate"); more explicit guidance would certainly improve overall levels of agreement. Second, providing caseworkers with training materials, exposing them to "correct" application of the Severity Index to a variety of cases (e.g., via the vignettes and master ratings) would likely result in more consistent, reliable application of the scale.

The reliabilities of the Severity Index, combined with indications of face and content validity, are strong enough to make it the measure of choice for civilian agencies looking for an easy, reliable way to rate family maltreatment severity. The modifications noted above should result in good-to-excellent levels of reliability, making validity research both worthwhile and necessary.

In conclusion, the Severity Index is the first clinician-friendly, clinician-administered measure that comprehensively assesses the severity of all forms of partner and child maltreatment. The design of the Severity Index marks a substantial improvement over current practice (i.e., binary decisions on a single form of maltreatment) because it recognizes that (a) the multiple forms of family maltreatment often occur simultaneously in families; (b) the sequelae of family maltreatment are highly variable, assumably because of range of maltreatment severity; and (c) the interventions for family maltreatment are chosen based on severity. Given that these three facts are universally recognized within the otherwise fractious field of family maltreatment, use of the Severity Index—or other multifactorial measures of family maltreatment severity with documented reliability—is not only warranted but also needed.

APPENDIX

United States Air Force Family Advocacy Program's Family Violence Severity Index

Type	1. None	2. Mild	3. Moderate	4. Severe	5. Death
1. Child physical abuse	Not substantiated	Minor injury, or no medical treatment	Major/minor physical injury. Short-term medical treatment may be indicated	Major medical injury or long-term medical treatment or inpatient care.	Death due to nonaccidental physical injury.
2. Child sexual abuse	Not substantiated	No physical contact. No readily apparent physical or emotional harm. No medical/mental health treatment.	Physical contact which does not involve oral, vaginal, anal penetration, or physical injury. Short-term mental health or medical treatment. Verbal or physical threats.	Contact involves oral, vaginal, anal penetration, or physical injury. Long-term mental health or medical treatment. Severe verbal threats or emotional abuse.	Death due to sexually abusive behavior.
3. Child neglect	Not substantiated	Isolated incident, no repetitive pattern evident. No readily apparent physical or emotional harm to the child, but was placed in potential harm.	Repeated incident of neglectful behavior or the child suffers physical or emotional harm from the circumstances. Short-term medical treatment.	Pattern of neglectful behavior resulting in hospitalization or alternate placement for the safety of the child.	Death resulting from the neglectful behavior of the offender.
4. Child emotional abuse	Not substantiated	Isolated incident, no repetitive pattern evident. No readily apparent physical or emotional harm to the child, but was placed in potential harm.	Repeated incident of emotionally abusive behaviors or the child suffers physical or emotional harm from the behavior. Short-term medical treatment or mental health treatment.	Pattern of emotionally abusive behavior resulting in hospitalization or long-term medical or mental health treatment or alternate placement.	Death/suicide due to emotionally abusive behavior by the offender.
5. Spouse emotional abuse	Not substantiated	No repetitive pattern evident or no physical or emotional harm to the spouse, but spouse was exposed to potential harm.	Repeated incident of emotionally abusive behavior or the spouse suffers physical or emotional harm from the behavior. Short-term medical treatment or mental health treatment.	Pattern of emotionally abusive behavior resulting in hospitalization or long-term medical or mental health treatment or alternate placement.	Death/suicide due to emotionally abusive behavior by the offender.
6. Spouse physical abuse	Not substantiated	Minor physical injury (DoD). No medical treatment indicated.	Minor/major physical injury (DoD). Short-term medical treatment.	Major physical injury (DoD), or long-term medical treatment or mental health or alternate placement.	Death due to nonaccidental physical injury.
7. Spouse sexual abuse	Not substantiated	Nonconsensual touching.	Nonconsensual activity other than touching. Significant verbal or physical threats.	Nonconsensual sexual activity involving severe threats or physical injury.	Death due to sexually abusive behavior.

ACKNOWLEDGMENTS

This work was supported by the partnership of (a) the US Air Force Family Advocacy Program and (b) the US Department of Agriculture National Network on Family Resilience (contract CR-4953-545735). Thanks to Col. John Nelson (director of US Air Force Family Advocacy Program) for supporting the research and to the outgoing USAF-FAP research director, Lt. Col. Carla A. Monroe-Posey (ret.) for coordinating the data collection and for comments on an earlier summary of these results. Thanks also to current USAF-FAP research director, Lt. Col. Dari Tritt, for comments on drafts of this article, to USAF-FAP's Kay Webber and Fidencio Gonzales for assisting in the data collection, and to USAF-FAP staff worldwide who served as raters. Thanks also to our USDA-NNFR project administrators, Drs. Craig Allen (Iowa State) and Sandra Stith (Virginia Tech). Finally, we are indebted to Cheryl Van Dyke (Stony Brook) for her contribution to all phases of the project.

REFERENCES

Becker, J. V., Alpert, J. L., BigFoot, D. S., Bonner, B. L., Geddie, L. F., Henggeler, S. W., Kaufman, K. L., and Walker, C. E. (1995). Empirical research on child abuse treatment: Report by the Child Abuse and Neglect Treatment Working Group, American Psychological Association. *J. Clin. Child Psychol.* 24(Suppl): 23–46.

Black, D. A., Heyman, R. E., and Slep, A. M. S. (2001a). Risk factors for child physical abuse. Risk factors for child sexual abuse. *Aggression & Violent Behavior* 6: 121–188.

Black, D. A., Heyman, R. E., and Slep, A. M. S. (2001b). Risk factors for child psychological abuse. *Aggression & Violent Behavior* 6: 189–201.

Black, D. A., Heyman, R. E., and Slep, A. M. S. (2001c). Risk factors for child sexual abuse. *Aggression & Violent Behavior* 6: 203–229.

Black, D. A., Heyman, R. E., and Slep, A. M. S. (2001d). Risk factors for male-to-female partner sexual abuse. *Aggression & Violent Behavior* 6: 269–280.

Cascardi, M., O'Leary, K. D., Schlee, K. A., and Lawrence, E. E. (1995). Major depressive disorder and posttraumatic stress disorder in physically abuse women. *J. Consult. Clin. Psychol.* 63: 616–623.

Chaffin, M., Wherry, J. N., Newlin, C., Crutchfield, A., and Dykman, R. (1997). The Abuse Dimensions Inventory: Initial data on a research measure of abuse severity. *J. Interpers. Viol.* 12: 569–589.

Cicchetti, D. V. (1985). A critique of Whitehurst's "Interrater agreement for journal manuscript reviews": De omnibus disputandem est. *Am. Psychol.* 40: 563–568.

Cozby, P. C. (1981). *Methods in Behavioral Research*, 2nd ed., Mayfield Publishing Company, Palo Alto, CA.

Department of Defense. (1987). *Child and Spouse Abuse Report* (Instruction 6400.2), Department of Defense, Washington, DC.

Drew, C. J., Hardman, M. L., and Hart, A. W. (1996). *Designing and Conducting Research: Inquiry in Education and Social Science*. Needham Heights, MA: Allyn and Bacon.

Haynes, S. N., Richard, D. C., and Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychol. Assess.* 7: 238–247.

Heyman, R. E., and Neidig, P. H. (1997). Physical aggression couples treatment. In Halford, W. K., and Markman, H. J. (eds.), *Clinical Handbook of Marriage and Couples Intervention*, Wiley, New York, pp. 589–617.

Kaufman, J., Jones, B., Stieglitz, E., Vitulano, L., and Mannarino, A. P. (1994). The use of multiple informants to assess children's maltreatment experiences. *J. Fam. Viol.* 9: 227–248.

McGee, R. A., and Wolfe, D. A. (1991). Psychological maltreatment: Toward an operational definition. *Dev. Psychopathol.* 3: 3–18.

McGee, R. A., Wolfe, D. A., Yuen, S. A., Wilson, S. K., and Carnochan, J. (1995). The measurement of maltreatment: A comparison of approaches. *Child Abuse Neglect* 19: 233–249.

Mollerstrom, W. W., Patchner, M. A., and Milner, J. S. (1995). Child maltreatment: The United States Air Force's response. *Child Abuse Neglect* 19: 325–334.

O'Leary, K. D. (1993). Through a psychological lens: Personality traits, personality disorders, and levels of violence. In Gelles, R. J., and Loseke, D. R. (eds.), *Current Controversies on Family Violence*, Sage, Newbury Park, CA, pp. 7–30.

O'Leary, K. D., Heyman, R. E., and Neidig, P. H. (1999). Treatment of wife abuse: A comparison of gender-specific and couples approaches. *Behav. Ther.* 30: 475–505.

Schumacher, J. A., Slep, A. M. S., and Heyman, R. E. (2001a). Risk factors for child neglect. *Aggression & Violent Behavior* 6: 231–254.

Schumacher, J. A., Feldbau-Kohn, S., Slep, A. M. S., and Heyman, R. E. (2001b). Risk factors for male-to-female partner physical abuse. *Aggression & Violent Behavior* 6: 281–352.

Schumacher, J. A., Slep, A. M. S., and Heyman, R. E. (2001c). Risk factors for male-to-female partner psychological abuse. *Aggression & Violent Behavior* 6: 255–268.

Sedlak, A. J., and Broadhurst, D. D. (1993). *The Third National Incidence Study of Child Abuse and Neglect*, U.S. Department of Health and Human Services, Washington, DC.

Stets, J. E., and Straus, M. A. (1990). Gender differences in reporting of marital violence and its medical and psychological consequences. In Straus, M. A., and Gelles, R. J. (eds.), *Physical Violence in American Families: Risk Factors and Adaptation to Violence in 8,145 families*, Transaction, New Brunswick, NJ, pp. 151–165.

Straus, M. A. (1979). Measuring intrafamily conflict and violence; The Conflict Tactics (CT) Scales. *J. Marriage Fam.* 41: 75–78.

Straus, M. A., and Gelles, R. J. (1990). *Physical Violence in American Families: Risk Factors and Adaptations to Violence in 8,145 Families*, Transaction, New Brunswick, NJ.

Straus, M. A., Hamby, S. L., Boney-McCoy, S., and Sugarman, D. B. (1996). The Revised Conflict Tactics Scales (CTS2): Development and preliminary psychometric data. *J. Fam. Issues* 17: 283–316.

U.S. Department of Health and Human Services. (2000) *Child Maltreatment 1998: Reports From the States to the National Child Abuse and Neglect Data System*, U.S. Government Printing Office, Washington, DC.

Vivian, D., and Langhinrichsen-Rohling, J. (1994). Are bi-directionally violent couples mutually victimized? A gender sensitive comparison. *Viol. Victims* 9: 107–124.

Wiggins, J. S. (1973) *Personality and Prediction: Principles of Personality Assessment*, Robert E. Krieger Publishing, Malabar, FL.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *Am. Psychol.* 39: 22–28.

Whitehurst, G. J. (1985). On lies, damned lies, and statistics: Measuring interrater agreement. *Am. Psychol.* 40: 568–569.