Four Measures of Partner Violence:
Construct Similarity and Classification Differences

Although measures of physical violence are commonly used to classify individuals as "having sustained violence" or "not having sustained violence," little is known about the interrelationships among violence measures and whether variation across measures is random or systematic. In this study, 224 female undergraduates completed four different partner violence measures. Confirmatory factor analysis indicated that a one-factor model best fits the data, but that the measures did not represent the construct equally well (i.e., the congeneric model was significantly better than tau-equivalent and parallel measurement models). Frequency measures were more strongly associated with the latent construct, "partner violence" than were severity measures. Some variation across measures appears to be systematic; participants reporting milder and more infrequent violence were classified most inconsistently. Single violence measures may classify individuals unreliable.

Physical violence between partners has been measured in many different ways. Although most measures rely, to some degree, on the reports of victims or perpetrators, differences among kinds of reports may lead to variations in what is identified as a physical assault (cf. Koss et al., 1994; Straus, 1990; Walker, 1984). For example, there has been considerable discussion of the merits and liabilities of using a predetermined set of violent acts when asking respondents about violence in their relationships (e.g., Straus, 1990). In addition to asking respondents about a set of behaviors, other common strategies include asking them to make a summary judgment about their total experience of violence (e.g., Bergman, 1992; Leonard & Blane, 1992) and asking for detailed descriptions of specific incidents (e.g., Dobash & Dobash, 1984; Frieze & McHugh, 1992; Walker, 1984). Major reviews of the field (e.g., Browne, 1993; Hotelling & Sugarman, 1986; Koss et al., 1994) often discuss methodological issues but accept the premise that each of the varying approaches to measuring violence refers to the same underlying construct (although not necessarily equally well).
Olson (1977) has provided a framework that helps to highlight some of the differences found in partner violence measures. His framework has two dimensions, frame of reference (insider or outsider) and type of data (subjective or objective). Insider reports are self-reports, and one of the advantages of his model is that he identifies two types of these. The Conflict Tactics Scales (CTS) and similar measures (e.g., Hudson & McIntosh, 1981; Shepard & Campbell, 1992), which focus respondents' attention on a specific set of behaviors, fall into Olson’s “objective insider” category. These types of behavioral self-reports have been the subject of both praise and criticism in the family violence literature (cf. Straus, 1990). Another common approach involves asking about the overall aggregate frequency of violence. These subjective self-reports typically allow respondents to determine what acts constitute violence. Critics of behavioral self-reports believe this approach, called “subjective insider” by Olson, will be more inclusive because unusual kinds of violence can be measured, but some loss of specificity occurs.

Pure outsider reports, of course, are difficult, if not impossible, to obtain in research on partner violence because of the practical and ethical problems related to observing such acts directly, especially in any systematic manner. Thus, most measures that try to adopt more of an outsider’s perspective are typically hybrids consisting of an outsider’s judgment of participants’ descriptions of violent episodes provided in interviews or questionnaires. In the partner violence literature, these judgments have typically been codified in some way (e.g., Frieze & McHugh, 1992) and, thus, belong on the objective side of the continuum. The most important advantages of outsider approaches to partner violence research are an increase in the amount of contextual information obtained from respondents and an attempt to reduce some of the potential biases of pure self-reports. The biases of the outsiders, however, may also affect these data.

Most commentators on methodology, including Olson (1977), have called for the use of more than one measurement approach in a single study (e.g., Cronbach & Meehl, 1955; Fiske, 1973; Pedhazur & Schmelkin, 1991). Occasionally, this suggestion is found in the family violence literature (e.g., Weis, 1989). Reliance on a single measure limits our ability to examine, both theoretically and empirically, what is meant by partner violence and physical abuse. Unfortunately, almost all studies of partner violence have relied on a single measure of physical violence (e.g., Bookwala, Frieze, Smith, & Ryan, 1992; Pan, Neidig, & O’Leary, 1994; Straus & Gelles, 1986; White & Humphrey, 1994; White & Koss, 1991). Many theoretical disputes are exacerbated by our reliance on a single operationalization of violence. For instance, the dispute over how broadly the term “physical violence” should be conceptualized (e.g., Caulfield & Riggs, 1992; Dobash & Dobash, 1988; Koss et al., 1994; Straus, 1990) could be addressed by comparing operationalizations of varying breadth and examining their ability to specify risk markers and predict treatment outcomes.

Issues of empirical import could also be more carefully addressed by using multiple measures. Little attention has been paid, for instance, to the reliability of violence classifications. Sugarman and Hotaling (1989) noted that different checklist measures do not appear to produce major differences in prevalence rates, but they also reported that prevalence estimates of violence among dating partners varied from 9% to 65%. (Of course, several factors probably contribute to these differences.) If multiple prevalence measures were available for each sample, then one could compare discrepant estimates directly and examine the extent to which individuals are classified differently on different measures. If estimates vary even slightly, then some individuals will be classified as having sustained violence on some measures and not on others. It is possible for this to be true even when prevalence estimates are identical. How these inconsistencies in classification affect partner violence research is largely unknown.

While some measurement error is unavoidable, the question arises whether inconsistencies in prevalence rates are random or systematic. Systematic inconsistencies are neither unique to particular individuals nor randomly distributed across individuals, but are regularly associated with the phenomenon of interest. Such inconsistencies complicate the interpretation of results. For instance, all individuals who have sustained any physical violence are typically included in one group. While we are not suggesting that “minor” and “infrequent” forms of violence are not of concern, it is possible that individuals’ reports of such experiences may not reach every measure’s criterion for physical violence. Even when violent behaviors are classified into minor and severe categories, there may be some ques-
tion about what behaviors meet the minimum criterion for these groups. One may then wonder how to determine whether a particular individual has, in fact, sustained violence.

Despite the theoretical and practical importance of knowledge that can be gained from the use of multiple measures of violence, very few studies have attempted to examine variation or consistency across measures. One exception is a recent study by Rodenburg and Fantuzzo (1993) that compared two measures in a sample of battered wives. Their own Measure of Wife Abuse—Physical Abuse factor was highly correlated with CTS Physical Violence scores ($r = .63$), despite the fact that the time referents for the two questionnaires were different. They did not, however, examine discrepancies between the measures.

Many studies use the criterion groups design to evaluate violence measures. Such studies do not directly address issues of consistency and cannot substitute for studies using multiple measures. Criterion group studies usually compare battered women to nonbattered women on a measure of partner violence (e.g., Hudson & McIntosh, 1981; Shepard & Campbell, 1992). Evidence that battered and nonbattered women differ on measures intended to assess partner violence provides little information about validity. Such evidence does not indicate that these instruments accurately measure partner violence or, for that matter, even measure partner violence at all. While it is certainly reasonable to expect that these two groups will differ on measures of partner violence, it does not follow that because the two groups differ on a measure, that the measure is one of partner violence. To say otherwise is to make the logical error of affirming the consequent (Pedhazur & Schmelkin, 1991). For example, battered and nonbattered women can be distinguished by questions that are labeled as measures of depression (e.g., Bergman, Larsson, Brismar, & Klang, 1987; Jaffe, Wolfe, Wilson, & Zak, 1986) or measures of alcohol use (e.g., Barnett & Fagan, 1993; Kaufman-Kantor & Asdigian, in press), in addition to questions labeled as measures of partner violence. The assertion that violence measures assess partner violence because they discriminate between battered and nonbattered women could also be made for depression and alcohol use measures, which make this same discrimination. Multiple measures are important because they help to establish construct validity. Data on the correspondence of multiple measures of partner violence are needed.

The present study is an examination of the similarities and differences among multiple measures of partner violence as reported by women in dating relationships. All measures of partner violence should be associated with the latent construct, "partner violence." We tested this hypothesis using confirmatory factor analysis (CFA). CFA enables one to determine whether all measures are equally good representations of the construct or whether some measures are better than others through a comparison of three measurement models: parallel, tau-equivalent, and congeneric (Pedhazur & Schmelkin, 1991). The parallel model assumes that all measures represent the latent construct equally well and have equal error. The tau-equivalent model also assumes that all measures represent the latent construct equally well, but error is allowed to vary. The congeneric is the least restrictive model because both the associations with the latent construct and error are allowed to vary. Accepting the appropriateness of this latter model indicates that some measures are better than others.

We further examined whether individual women are consistently classified across measures as having sustained or not having sustained violence, or whether some women are classified differently on different measures. We anticipated that classification inconsistencies would be somewhat systematic, and we examined the inconsistencies by exploring what incidents are reported by women who are classified inconsistently. We anticipated that more minor and infrequent violence would be most inconsistently classified because such acts hover near the threshold for violence.

**METHOD**

**Respondents**

The respondents were 224 female undergraduates at a large, state-supported Southeastern university. All were currently in heterosexual dating relationships that averaged under 2 years in length. For participating in this study, all respondents received credit for an introductory psychology course. The study included questions on relationship commitment, in addition to the questions on partner violence that are presented here.

The sample consisted mostly of White, Protestant females who belong to families with upper-middle class incomes. Most of the sample indicated that their parents were currently married to
each other. They reported that their partners' sociodemographic characteristics were virtually identical to their own. These characteristics generally reflect the ethnic and socioeconomic composition of the undergraduate body at this university. See Table 1.

**Measures**

*Demographic variables*. The set of questionnaires that we used began by asking respondents to provide the sociodemographic information described above and presented in Table 1.

**Conflict Tactics Scales (CTS)**. Respondents completed a variant of Form N of the CTS (Straus, 1979), which is the most widely used measure to assess physical violence towards one's partner (Straus, 1990; Sugarman & Hotaling, 1989). The CTS includes a number of behaviors that an individual could use in an interaction with his or her partner. For each behavior, the respondent is asked to indicate how many times it has occurred, using the following seven categories: 0, 1, 2, 3-5, 6-10, 11-20, more than 20. Because other behavioral checklists include items that are identical or similar to those on the CTS (e.g., Hudson & McIntosh, 1981; Rodenburg & Fantuzzo, 1993; Shepard & Campbell, 1992), only one behavioral checklist was included.

**Judges' severity rating of description of worst incident (Judges' Severity)**. This measure adopts an approach to the measurement of violence different from that taken by the CTS. The format of the question bears some similarity to an interview probe used by other researchers (e.g., Dobash & Dobash, 1984; Frieze & McHugh, 1992; Walker, 1984), although the richness of responses to this measure does not approach what can be obtained through an extensive individual interview. The measure was intended to partly address the goals of those who advocate the use of more qualitative approaches to assessing partner violence and broad conceptualizations of physical violence (e.g., Dobash & Dobash, 1984; Koss et al., 1994).

Respondents read the following instructions: "We are trying to learn about the ways people act in different relationships, especially if and when partners use force. Please think about the most forceful, physically or verbally threatening episode directed towards you in your current relationship, even if you would not normally consider it violent. All respondents should answer these questions. Some examples of forceful acts that sometimes happen are pushing, hitting, shouting, slapping, slamming the door, biting, cursing, threatening with a weapon, etc." Examples were included because two earlier rounds of pretesting indicated

### Table 1. Demographic Characteristics of Respondents

<table>
<thead>
<tr>
<th>Age (in years)</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19.06</td>
<td>0.99</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>86%</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td>Family income (yearly estimated)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under $40,000</td>
<td>19%</td>
<td></td>
</tr>
<tr>
<td>$40,000 to $54,999</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>$55,000 to $69,999</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>$70,000 to $89,999</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>$90,000 and over</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Parents' marital status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married currently</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>27%</td>
<td></td>
</tr>
<tr>
<td>Relationship status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dating</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>Engaged</td>
<td>8%</td>
<td></td>
</tr>
<tr>
<td>Cohabiting</td>
<td>1%</td>
<td></td>
</tr>
<tr>
<td>Relationship length (in months)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>20.44</td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>15.32</td>
<td></td>
</tr>
</tbody>
</table>
that some respondents omitted descriptions of verbal aggression or described only the emotional consequences of the actions and not the acts.

The first question following these instructions read, "Describe what you and your partner did (or said), including what kind of force was used." Two female undergraduates were trained to rate the severity of the incident described. These judges were blind to other information about the respondents and to the hypotheses of the current study. Both judges rated all descriptions. They found that 98% of the descriptions could be rated on a 7-point scale that ranged from no force to severe force. A rating of 1 indicated that conflict and anger had occurred, but had been resolved without the use of verbal or physical aggression. A rating of 2 involved verbal aggression, including cursing, shouting, and name-calling. Ratings of 3 or higher involved some use of physical aggression, defined broadly to include any instances in which physical force was used to bodily control the respondent against her will in an attempt to intimidate or coerce her (i.e., to "win"). For example, blocking someone from leaving the room or driving recklessly to intimidate the respondent were classified as using physical aggression. Such acts received the lowest physical aggression rating of 3. This category also included actions in which minimally forceful contact was made, such as grabbing by the arm. A rating of 4 was assigned to actions such as pushing or shaking that moved the victim bodily but did not knock her off balance. A rating of 5 was assigned to actions that knocked the victim off balance but not to the ground, to slapping, and also to throwing something at the person that was unlikely to cause serious injury. Throwing objects such as knives or hot pans received higher ratings. A rating of 6 was used for actions that knocked the victim down. The highest rating, 7, was used for force that could easily or did result in physical injury—actions such as punching, hitting repeatedly, or using a weapon. The judges' ratings showed excellent reliability. The effective reliability or aggregate consistency of the two judges' scores (an inter-rater reliability score for continuous ratings that is based on the correlation coefficient, see Rosenthal & Rosnow, 1991) was .98. In order to make the format of these ratings comparable to the frequency measures (CTS and Aggregate Frequency), all nonphysical force (scores of 1 and 2) was grouped into one category.

Self-Report Severity Rating of the worst incident (Self-Report Severity). After they described the worst incident, respondents were asked to rate the behavior directed towards them during that incident on an 8-point scale from 0 (not forceful) to 7 (extremely forceful). We are not aware of other uses of this type of question, but included it to complement the Judges' Severity Ratings of the worst incident.

Self-Report Aggregate Frequency question (Aggregate Frequency). Using a single question with a response format analogous to that of the CTS, respondents were also asked how many times they had experienced physical force (not explicitly defined) during the course of their current relationship. The measurement characteristics of this kind of aggregate prompt are interesting for two reasons: (a) such questions are one of the most commonly used means of assessing partner violence, particularly in dating samples (cf. Sugarman & Hotaling, 1989; also see Bergman, 1992; Jackson, 1991; Leonard & Blane, 1992, for more recent examples) and (b) information obtained with this brief prompt can be compared with other violence measures.

Procedure

Respondents completed the study in groups of 15–20 women. Informed consent was obtained prior to the completion of all questionnaires. Of the violence measures, the CTS was administered first, followed by the written description of the worst incident. The respondents then rated the severity of this incident (Self-Report Severity) and lastly indicated the total frequency of force acts (Aggregate Frequency).

Results

The mean and standard deviation for each violence measure were as follows: For the CTS, $M = 0.92$, $SD = 2.49$; for Judges' Severity, $M = 0.92$, $SD = 1.36$; for Self-Report Severity, $M = 1.93$, $SD = 1.77$ (unlike other scales, this is a Likert scale with no zero-point); and for Aggregate Frequency, $M = 0.63$, $SD = 1.14$.

Variables in violence research are often skewed because many individuals report no violence. In this study, the two frequency measures, the CTS and Aggregate Frequency, were significantly positively skewed; that is, scores tended to cluster towards 0 (no violence). Judges' Severity was less skewed than the frequency measures, but more scores still occurred at the low end than would be expected for a normally distributed
TABLE 2. INTERCORRELATIONS OF VIOLENCE MEASURES

<table>
<thead>
<tr>
<th></th>
<th>CTS</th>
<th>Judges' Severity</th>
<th>Aggregate Frequency</th>
<th>Self-Report Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td></td>
<td>.55</td>
<td>.74</td>
<td>.51</td>
</tr>
<tr>
<td>Judges’ Severity</td>
<td></td>
<td></td>
<td>.59</td>
<td>.44</td>
</tr>
<tr>
<td>Aggregate Frequency</td>
<td></td>
<td></td>
<td></td>
<td>.57</td>
</tr>
<tr>
<td>Self-Report Severity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: All correlations are significant; \( p < .0001; \) \( n = 215 \).

variable. The mean for Self-Report Severity was also close to the low end for that scale, but skewness and kurtosis scores were not excessively elevated. The skew and kurtosis will tend to attenuate the correlations among variables, but the skewness of these measures probably reflects the skewed distribution of violence in the population.

**Similarities Among Violence Measures**

Similarities among measures are examined three ways. The intercorrelations of all four violence measures are presented to allow for bivariate comparisons. Confirmatory factor analysis is then used to examine whether all four measures belong to a single factor and whether all four measures are equivalent. Finally, the overall violence prevalence rates are compared.

**Intercorrelations Among Measures**

Zero-order Pearson product-moment correlations were computed for the four physical violence measures. The CTS violence scale correlated highly with the Aggregate Frequency question. All other measures of violence were moderately correlated with each other. See Table 2 for the correlation matrix for all force measures.

**Equivalency of Violence Measures**

Confirmatory factor analysis was used to address the equivalency of the four violence measures. The underlying latent construct was presumed to be physical violence by a male partner. Different measures of the same construct can be represented by one of three measurement models. The parallel model is the most restrictive and states that all measures have the same association with the latent construct (i.e., have the same weight) and that their error terms are also equal. The tau-equivalent model also postulates that the measures are equivalent measures of the latent construct, but their errors can vary. The congeneric model is the least restrictive and allows both the association with the latent construct and the errors to vary.

Confirmatory factor analysis provides a direct means of testing these models because the various constraints can be built directly into the hypothesized measurement model. If the observed data do not fit the hypothesized model well, that is evidence that the constraints are too restrictive. Because the models all have the same number of parameters, which are fixed in some models and free to vary in others, they are considered nested and can be compared directly using the chi-square difference test (see Jöreskog & Sörbom, 1993; Pedhazur & Schmelkin, 1991).

The most restrictive parallel model was examined first. It did not fit the data particularly well, as can be seen by the description of the model in Table 3. The significant chi-square indicates that the model is different from the data. Theoretically, the fit indices, the adjusted goodness-of-fit index (AGFI) and the normed fit index (NFI), should be extremely close to 1. Their values are somewhat lower than this for the parallel model. The root mean square residual (RMR) is not as close to 0 as would be desirable. Further, one of the standardized residuals (for the association between CTS and Aggregate Frequency) exceeds the level of significance at 3.19 (critical value = 2.58). Thus, a less restrictive model was examined.

The next model examined, the tau-equivalent, allows for the estimation of individual error terms for each measure. This model is significantly better than the parallel model, as indicated by a significant chi-square for the difference in fit between the two models. Nonetheless, the tau-equivalent model still does not fit the data especially well. The overall chi-square for this model is still significant, and the RMR is actually somewhat higher. The standardized residual for the CTS and Aggregate Frequency is still significantly elevated (2.77).
Next, the congeneric model was examined. This model improves significantly over the tau-equivalent model, as evidenced by a significant chi-square difference, and is also significantly better than the parallel model. With a nonsignificant chi-square and very good fit and residual indices, this model also fits the data very well. The congeneric model reveals that all four measures have strong associations with the Partner Violence factor that was hypothesized to underlie each of these measures. The model also indicates, however, that the measures are not equally related to the Partner Violence factor. The two frequency measures, Aggregate Frequency and the CTS, load more highly on the Partner Violence factor than the two severity measures (Judge's Severity and Self-Report Severity).

Because of the pattern of differences in the magnitude of the loadings across measures, two additional models were examined post hoc: a model that hypothesizes that the severity measures do not load on the Partner Violence factor and a two-factor model that postulates separate Frequency and Severity factors. The first, the frequency-only model, was a very poor fit for the data. The two-factor model fit the data about as well as the congeneric model, but the intercorrelation between the two factors was .97, clearly indicating that a one-factor solution that includes all four measures is the most elegant conceptualization of these data. The one-factor congeneric model, therefore, was retained as the preferred model. The parameter estimates for the accepted congeneric model are in Table 4.

**Similarity of Prevalence Rates**

Commonly, measures of partner violence are used to classify participants as those who sustained and those who did not sustain violence. Three of the measures used in this study, Aggregate Frequency, CTS, and Judges' Severity, can also classify participants into dichotomous groups because in each case a distinct point is identified as representing violence. For Aggregate Frequency and the CTS, this is a frequency of one incident or more, and for Judges' Severity, this includes all episodes in which physical force was used to bodily control the respondent. The Self-Report Severity question cannot be used in a comparable manner to determine the number of respondents who had experienced or had not experienced physical violence because there is no defined cutoff for violence on that scale. Thus, this measure is not used to classify participants. The violence prevalence rates were calculated for each of the three other measures. The measures yielded relatively similar estimates: The CTS rate was 27%, the Aggregate Frequency rate was 32%, and the Judge's Severity rate was 38%.

**Differences Among Violence Measures**

Differences among violence measures were examined in two ways. First, a profile of classifications...
was constructed for each individual. This profile was used to identify three groups of participants: those classified as having sustained violence on every measure, those classified as not having sustained violence on every measure, and those classified differently on different measures. These groups were then used to identify the behaviors that most often led to inconsistent classifications.

**Individual Classification Differences Across Measures**

A profile of classifications across the three dichotomous measures (CTS, Aggregate Frequency, and Judges’ Severity) was created for each individual and then summed across individuals to identify the obtained frequency of the eight possible classification profiles. The results can be found in Table 5. The largest group of respondents, 51% of the sample, consisted of women who were classified as not having sustained violence on all three measures (Profile 1). The next largest group was the 17% classified as having sustained violence on all three measures (Profile 8). These two groups, however, accounted for only 68% of all respondents. A considerable proportion (32%) of respondents fell into remaining six profiles (2 through 7), indicating that they were classified inconsistently (i.e., were placed in different groups by different measures). Slightly less than half of this inconsistent group (14% of all respondents) was classified as having sustained violence on two of the three measures.

**Types of Behaviors Inconsistently Classified**

Profiles 5, 6, and 7 (see Table 5) represent the 23 individuals who were identified as having sustained violence by the CTS, but were considered nonabused on at least one of the other two measures. Because the CTS lists specific acts, one can determine which behaviors were reported on the CTS that were not defined as violent by another measure. “Minor” and infrequent forms of force were more likely to be classified inconsistently. By far the most commonly reported CTS item for these 23 individuals was “pushed, grabbed, or shoved you”; 20 people (87%) in this group indicated this had happened to them. Two of the three others reported “tried to hit you with something,” and the remaining person reported “threatened you with a knife or gun.” Only three individuals (13%) in this group reported more than one class of behavior (i.e., received a nonzero score for more than one CTS item). Most individuals in this group also reported that violence occurred infrequently; 17 (74%) reported only one occurrence of violent behavior.

More minor forms of force were also more likely to lead to individuals falling into profile groups 2, 4, and 6. These were individuals who were identified as reporting violence by the judges’ rating, but not by at least one of the other two measures. Eight of the 46 individuals in these groups (17%) were judged to have experienced “mild” violence, such as grabbing or being physically prevented from leaving the room. Four of these eight reported being grabbed. Most of them (63%) reported experiencing “mild-to-moderate” violence, including pushing, shoving, or shaking lightly. The CTS and other measures define such behaviors as minor violence. Some individuals (20%), however, described more violent incidents, such as slapping, being picked up and dropped on cement, and being thrown across the room. Thus, 91% of these behaviors theoretically would meet the threshold for violence on the CTS.

Profiles 3, 4, and 7 represent 32 individuals who were classified as having sustained violence on Aggregate Frequency but not on one of the other measures. Again, infrequent forms of force were more likely to be classified inconsistently. Severity cannot be identified because the specific behaviors cannot be identified with this measure. Twenty women (63%) reported experiencing one violent incident, 8 (25%) reported two, and 4 (12%) reported three to five violent incidents.

**Discussion**

**Conclusions**

The measures of partner violence used in this study had much in common. The measures were
Measures of Partner Violence

strongly intercorrelated with one another, and they yielded relatively similar prevalence rates. Furthermore, they all belonged to the same underlying Partner Violence factor. Factor models that hypothesized otherwise did not represent the data as well as the one-factor, four-measure model. All of these findings indicate that despite the somewhat different conceptions of violence represented by these measures, a common thread unites them. These results corroborate the conventional judgment that these different measures are operationalizations of the same construct (e.g., Browne, 1993; Hotaling & Sugarman, 1986; Koss et al., 1994; Sugarman & Hotaling, 1989).

The differences among measures were also substantial. Even though the overall prevalence rates obtained with different measures were similar, they did not represent the partner violence construct equally well. In this data set, the Aggregate Frequency measure was most closely related to this construct. Most importantly, almost a third of participants were classified inconsistently across measures. The high rate of inconsistent classification is especially surprising in a sample that is relatively homogeneous in other respects. Several trends in the inconsistencies across measures are apparent, and there were several possible sources of these trends. For instance, the inconsistencies could be due to differences in the threshold used to classify violent acts across different measures. Alternatively, inconsistencies could be due to differences in the range of acts that can be identified that are above that threshold. Finally, there may be differences between respondents’ thresholds for violence and the measures’ thresholds.

There was little evidence that threshold differences among measures caused the inconsistencies. Surprisingly, threshold differences appeared to have a minimal impact on the overall prevalence estimate from Judges’ Severity ratings, despite the fact that measure has the lowest definitional threshold. Most of the classification differences (over 90%) between Judges’ Severity and the other measures were for acts that would meet the violence threshold on the CTS (and most other behavioral self-report measures). Most incidents identified by Judges’ Severity (but not by other measures) were acts that were at least as serious as “pushed, grabbed, or shoved you,” the first violent item on the CTS. To put it another way, only 2% of respondents reported acts that fell between the threshold of the Judges’ Severity definition of violence and the CTS definition.

On the other hand, inclusiveness issues appeared to have some impact on inconsistencies. The CTS has been criticized for limiting its definition of violence to a predetermined set of acts (cf. Straus, 1990), and some differences between the CTS and other measures were apparently due to the fact that some reported behaviors—such as being shaken, being dropped on cement, and being thrown across the room (identified from Judges’ Severity descriptions)—do not have CTS categories. This may be one reason that the CTS gave the lowest prevalence rate in this study, even though it is the only measure that asks multiple questions about the occurrence of violence.

Most of the inconsistencies, however, appear to be related to differences in respondents’ thresholds versus those adopted by the measures. Objective definitions of violence, which influence to some extent the CTS and Judges’ Severity, appear to include more minor and infrequent acts of violence than subjective definitions include, as far as can be determined by these data. “Minor” acts of violence were much more likely to be classified inconsistently. For example, most differences between the CTS and the other measures (87%) were associated with the “pushed, grabbed, or shoved you” item. It was reported by several respondents on the CTS who indicated on at least one other measure that they had not sustained violence. Similarly, most differences between Judges’ Severity and other measures (80%) were for acts that are usually classified as “minor violence.”

Infrequent acts were also much more likely to be classified inconsistently, again suggesting differences between respondents’ criteria for violence and the measures’ criteria. Most (74%) individuals classified as having sustained violence on the CTS but not on other measures fell into the sustained-violence group on the CTS because of a single report of violence. Although it is impossible to determine what kinds of acts respondents are including on Aggregate Frequency that they have omitted from other measures, it is evident that single episodes of violence were much more likely to lead to inconsistencies between Aggregate Frequency and the other measures. In other words, most inconsistently classified respondents reported having experienced one act of violence.

Thus, although the similarities among the measures are substantial, there is reason to be concerned about the usual use of these measures to represent individuals’ experiences of violent relationships. The implications of these findings
are discussed below, but first we draw attention to some of the limitations of the current study.

Limitations

Three limitations of the current study should be noted. The nature of our sample is the first. It is possible that the same findings would not be obtained with respondents with different backgrounds in relationship status, sex, ethnicity, or socioeconomic position. This sample was rather homogeneous in terms of sex, race, and socioeconomic status, and that, along with, presumably, the respondents' considerable experience in completing questionnaires, might have contributed to a greater consistency of results than would otherwise be found. The sample probably generalizes to other studies of courtship violence among women because much of the research in this growing field is based on undergraduate samples. The extent to which dating violence is comparable to marital violence has been debated, but there appears to be some link between the two (Sugarman & Hotaling, 1989; White & Koss, 1991). Thus, these data may generalize to that population to some extent. Also, because administering four measures of the same construct is time consuming, undergraduate samples may be an appropriate starting point for investigating methodological issues that have received little empirical attention elsewhere (cf. White & Koss, 1991).

Because the questionnaires were administered in the same order for all respondents, order effects are also an issue. In particular, one might question whether the CTS provided cues that led to greater endorsement of violence on other measures. Two other studies (Hamby & Gray-Little, 1995; Nicolas, 1991) provide evidence that this is not the case. In both, respondents did not complete the CTS but did complete the other measures, and prevalence rates were very similar to the rates reported in this study. In the Hamby and Gray-Little study (1995), Judges' Severity yielded a prevalence estimate of 34%, and Aggregate Frequency produced an estimate of 31% for those in courtship relationships. In the Nicolas study (1991), the Judges' Severity estimate was 39%. Aggregate Frequency yielded an estimate of 48%, which is higher than the one found in the current sample (32%). In the Nicolas study, however, Aggregate Frequency was asked in a context of mutual violence, and the higher rate was consistent with expectations. While these studies do not provide conclusive evidence that order effects are unimportant, they suggest that comparable estimates are obtained for at least some order combinations and also indicate that the prevalence estimates obtained with some measures were not inflated due to the prior completion of the CTS.

Finally, the short duration of many courtship relationships poses special measurement problems. In the marital violence literature, violence is often assessed for the past year, which makes the time sampling of events consistent. This reduces one source of measurement error (Huston & Robins, 1982). But this approach is not without its costs. The low frequency of violence can create base rate problems, and it may be as important, conceptually, to identify the most serious incidents or all incidents as it is to identify recent ones. Thus, some marital research, especially interview-based research, has focused on the entire relationship (Frieze & McHugh, 1992; Walker, 1984). For dating relationships, the choice of referent is even more difficult because many relationships have not lasted for an entire year. Shorter referents (such as one month) would be consistent but would exacerbate base rate problems and probably miss a considerable amount of relevant data. Because the construct of interest is usually conceptualized as "violent relationships," short referents also may be conceptually further from the phenomenon. Longer referents offer the opportunity to document more instances of violence, but the length of the referent period can vary across respondents.

Researchers of courtship violence have solved the referent problem several ways. Some have used the relationship as the referent (e.g., Bird, Stith, & Schladele, 1991; Henton, Cate, Koval, Lloyd, & Christopher, 1983), some have used a lifetime referent (e.g., Bergman, 1992; White & Koss, 1991), and some have used a 1-year referent, even though not all respondents have dated that long (e.g., Stets & Henderson, 1991). There have not been, as far as we are aware, empirical studies of the differences obtained using these various referents. We adopted a relationship referent, in part, because we were also interested in assessing other variables specific to the current relationship (e.g., commitment), in part, because we did not want to exclude women in shorter relationships (short relationships are common in the dating population), and, in part, because it is a referent that has been used in marital research. In future research, however, it would be worth assessing if there are persuasive reasons for using one referent instead of another.
Implications

These findings suggest that the tendency to perceive violence as an all-or-nothing phenomenon oversimplifies women's experiences. While simplification is a feature of any classification scheme, the systematic nature of these inconsistencies suggests that further attention to this issue is warranted. Other authors (e.g., Koss et al., 1994; Ylio, 1988) have questioned the underlying logic in placing individuals who have experienced a single push or one grab in the same category as those who have experienced repeated and severe beatings. In fact, such an argument is often associated with some of the major controversies in the field of partner violence. For instance, research on the mutuality of male and female violence is often based on dichotomous prevalence rates (e.g., White & Kowalski, 1994). The oft-reported inconsistencies between spouses' reports may also relate to these issues (e.g., Langhinrichsen-Rohling & Vivian, 1994; McLaughlin, Leonard, & Senchak, 1992), as well as findings indicating that women often do not label their violent experiences as abusive (e.g., Koss, Diner, Seibel, & Cox, 1988; Lloyd, 1991). The data from the current study suggest that dichotomous prevalence rates based on single measures may be relatively unstable (i.e., one-third may change status across measures), and, thus, should be interpreted with caution.

On the other hand, the suggestion that only severe and repeated violence "counts" or even "counts more" seems to come disturbingly close to a normalization of minor violence. There is evidence that minor violence is associated with greater depression and poor family functioning (Meredith, Abbott, & Adams, 1986; Sugarman, Aldarondo, & Boney-McCoy, in press). Minor violence is also associated with greater risk for severe assaults later (Feld & Straus, 1989). Likewise, there is considerable evidence that partner violence is often associated with other violent behaviors and personality traits (cf. Holtzworth-Munroe & Stuart, 1994, for a review). There is little reason to discount the occurrence of minor and infrequent violence. Further, from a practical perspective, a narrower definition of violence would only exacerbate the problem of low base rates that already plague epidemiological research on partner violence and adds considerably to the cost of such research. The rate in this sample is halved when only consistently classified participants are included.

A consideration of different degrees of violence is one solution to these difficulties. Most studies of partner violence include only two groups, those who have and those who have not sustained violence. Rather than adopting dichotomous classification schemes, however, researchers should study partner violence as a continuum or as a three-level (none, minor, severe) categorical variable. Johnson (1995) has suggested that "terroristic" violence be included as a separate level. The causes and consequences of partner violence could be studied in more detail if violence were analyzed as a continuum. For example, Sugarman and his colleagues (in press) have recently demonstrated that many factors (such as alcohol use) associated with partner violence increase linearly as violence increases, and Pan et al. (1994) have documented that risk factors may vary for minor and severe violence. Also, the Feld and Straus (1989) finding that minor violence poses a risk of escalation could only be determined by distinguishing among degrees of violence. One should be cautious, however, about analytic strategies that first combine minor and severe violence and then examine the severe group only because the severe group may account for most of the effect of the comparison of the combined violent group with the nonviolent group.

Research on partner violence has been conducted for more than two decades now, and it is time we paid more attention to basic psychometric issues. Two specific suggestions about how to measure partner violence are indicated by this study. First, allowing participants to define their experiences of physical force adds important information to that obtained by fixed-format behavioral checklists. Other researchers (Berger, Fischer, & Rose, 1994; Bograd, 1988) have also highlighted the importance of obtaining women's own perceptions of their experiences in order to comprehensively measure their experiences. Second, the results provide empirical evidence that measures of frequency are more central to the construct of "partner violence" than measures that focus on the severity of a single episode. Because most existing measures are behavioral checklists, the majority does emphasize the frequency of violence. This finding also raises theoretical questions about the definition of partner violence and whether repeated acts of violence are important to the definition of violence. Lastly, these data highlight the importance of using multiple measures to assess partner violence. There
appear to be systematic difficulties associated with the measure of partner violence, particularly the measure of minor and infrequent violence, and we need to give consideration to results obtained with different measures.

NOTE
Support for this study was provided to the first author by the Family Research Laboratory, University of New Hampshire, and NIMH grant 5-T32-MH15161. The authors would like to acknowledge David Finkelhor, Vicki Banyard, Murray Straus, C. Albert Bardi, the members of the Family Research Seminar, and two anonymous reviewers for their comments on an earlier draft of this article. Please address correspondence concerning this article to Sherry L. Hamby.

REFERENCES


