# 3

# The Challenges of Measuring Violence Against Women

Diane R. Follingstad

The primary aim of measurement is to provide information about a phenomenon, and developing methodology for this purpose has been an important focus for many professionals in the area of violence against women (VAW) and intimate partner violence (IPV). Initially, the most important motivator for IPV researchers to engage in measurement was to document the frequency with which various forms of violence against women took place as a way to focus a spotlight on this societal problem. Quickly, however, this initial goal was expanded to measure and identify causal factors, risk factors, and/or prevention factors in the area of VAW that would have implications for developing intervention and prevention programs.

With such important goals and the good intentions of professionals to accomplish these goals, it seems almost heresy to critique the measurement strategies of VAW researchers. Nevertheless, measurement within the field of VAW is often problematic, at times resulting in contradictory findings or not allowing for definitive statements. With the goal in mind to sensitize policy makers and laypersons alike to this significant issue, choices are made as to which data are reported or highlighted that support a particular stance. Consequently, persons utilizing data from IPV literature may be *unaware* of potential problems with developed scales, interview strategies, statistical approaches, or even interpretations of data.

Unfortunately, there are numerous fallacies that many people hold about IPV measurement that, with some basic understanding, would allow them to be more critical consumers of the research literature in this area. The word *critical* is meant to imply that a person will bring an evaluative filter to his or her reading of IPV source material to make judgments regarding the relative quality of research studies to select the better ones on which to base conclusions. Being a more sophisticated consumer of research does *not* imply holding impossible standards for research areas such as IPV, which cannot attain the most rigorous requirements for experimentation, such as the government might require when testing the efficacy of new drugs. When women's lives are being studied, we often cannot apply the same standards as experimental research because typically, we are making assessments *after* a

violent experience. Also, a researcher would never place a person in an experimental condition where violence would occur in order to compare her or him to a person in a parallel control condition where violence did not occur. Rather, a critical consumer of this literature knows the difficulties and pitfalls of research and measurement, especially for understanding VAW and IPV, and is therefore able to make reasonable interpretations and avoid using flawed studies when making assertions.

The goal of this chapter is to raise the *critical* quotient for those wishing to read and use IPV and VAW literature in the most effective manner. To accomplish this, I will raise and address numerous fallacies, both general to social-science measurement and specific to IPV research. These fallacies stem from such issues as whether definitions exist for types of IPV, whether there is agreement about the definitions, and whether we can measure all forms of IPV through similar methods. The discussion will also include technical issues of how we might know whether the way a particular variable (e.g., psychological harm from physical abuse) was measured used a valid method; whether our use of surveys, questionnaires, and interview questions are appropriately scored or interpreted; and the difficulty of applying particular standards that are commonly used in the social sciences. For example, some discussion will center on the premise that not all statistical approaches employed by the social sciences for *reliability* (i.e., consistency of items in a scale and consistency of a person's answers from one time to the next) or *validity* (i.e., degree of accuracy of how well the researcher's approach actually measures the concept) may be appropriate for determining the quality of research on IPV and VAW topics. Combining theory, exposure of fallacious thinking, knowledge of scale development and data collection strategies, and psychometric considerations should result in an analytical perceptiveness leading to more refined conclusions for those willing to delve into the intriguing facets of measurement. It is my hope that this approach to considering the intricacies of measuring IPV and VAW will provide readers the background for developing a more sophisticated "filter" when utilizing IPV literature.

## Problematic Assumptions for Measuring IPV or Interpreting Research Findings

### Assumption 1: Most people agree on what violence and abuse are.

Although most people would agree that particularly horrendous behaviors (e.g., one spouse threatening to kill the children if the other spouse does not comply with demands) are absolute violations of the norms of interpersonal behavior, there can be substantial variability when researchers and/or practitioners label actions as abuse or violence, especially for less egregious behaviors. (Note: Although the terms *violence* and *abuse* are used interchangeably, *violence* is usually reserved for physical (and sexual) actions while *abuse* is more likely to be used in reference to nonphysical actions.) This is also true for laypersons (Follingstad, Helff, Binford, Runge, & White, 2004). Professionals demonstrate the most consistency when labeling behavior as violence or abuse when they are confronted with individuals using *physical* force toward their partners. Among mental-health professionals, practitioners, and advocates, there appears to be consensus that any use of physical force should be labeled *violence*. Some disagreement exists, however, when labeling actions of individuals who *threaten* to use physical force (Hegarty, Bush, & Sheehan, 2005) or who perform intimidating physical actions near their partner (e.g., hitting a wall next to them). Some researchers have suggested that physical force directed toward a partner in self-defense should not be counted as a violent action (Dobash, Dobash, Wilson, & Daly, 1992), but no agreed-upon criteria have been developed for helping a person screen his or her own actions to determine whether or not the potentially self-defensive use of force should be divulged on a survey. Further,

different behavioral patterns or characteristics of participants (e.g., gender) may influence whether physical force is labeled as violence; for instance, the more (unfortunately) common use of milder physical behaviors, such as slapping or pushing that is engaged in reciprocally by young adults and that recedes with maturity, is not as quickly labeled *violence* by professionals.

Deciding what is *violence* for adult *sexual* behaviors has been determined mostly by criminal statutes defining sexual assault. For reporting purposes, however, the concern is not whether there is agreement as to the violent and nonconsensual nature of the actions but, rather, whether persons experiencing sexual assault *within their intimate relationships* would recognize and/or label it as assault or violence. Some women have reported that they believed that their partners had a "marital right" to have sex with them even against their will or that because they had consented to sex within the relationship in the past, their partners have the right to force sex currently (e.g., E. K. Martin, Taft, & Resick, 2007). For behaviors that might not fit criteria for violent sexual assaults but still appear offensive or problematic, it is more difficult to define or produce consensus regarding these behaviors that fall below a criminal threshold that defines sexual assault as obtaining nonconsensual sex through physical coercion or when a person is incapacitated. The dialogue is just beginning for researchers who are attempting to distinguish lower-threshold sexually *abusive* behaviors from even more controversial examples of problematic, boorish, inept, or annoying partner behaviors, all the while attempting to avoid the pitfall of classifying differences in partner sexual preferences as abusive simply because one member of the couple finds the partner's preferences objectionable (e.g., a man who does not like oral sex but whose wife badgers him to engage in it with her).

Not surprisingly, determining consensus among professionals for which *psychological* behaviors constitute abuse is even more difficult than determining consensus for labeling physical or sexual violence. Although the literature abounds with descriptions of psychological actions that are considered *abuse*, practicing psychologists do not always label these as *abuse*, and certain factors (e.g., frequency or the intentions of the person engaging in the behavior) influence their willingness to label such behaviors as abuse (Follingstad & DeHart, 2000). One study determined that actions are more likely to be considered abuse when exhibited by a man toward his partner than when a woman engages in those actions toward her partner (Follingstad, DeHart, & Green, 2004), a phenomenon that would not be expected if a particular behavior is, in and of itself, *abusive*. Psychologists appeared to endorse some stereotypical sex roles, such as viewing a woman's monitoring of her husband's behavior as less problematic than the reverse, as well as believing that some actions by a woman would not have the same negative impact as the same actions engaged in by a man. Laypersons are more likely than psychologists to consider emotionally hurtful behaviors as abuse, whereas psychologists are more likely to consider behaviors indicative of severe jealousy to be abusive (Follingstad, Helff, et al., 2004). Overall, psychological maltreatment seems to be much more in the eye of the beholder if it does not occur at the extreme end of psychological cruelty so that consensus has not been achieved regarding definitions or conceptual anchors for the concept. More than other forms of IPV, psychological maltreatment often occurs reciprocally, which can cloud the assessment of who is the victim and who is the aggressor.

***Issue to be considered.*** *Abuse* is not a scientific term but, rather, a societal judgment that behavior has surpassed an acceptable threshold of conflict into deliberate attempts to harm. While it is not wrong for a society to make such judgments, asking science to determine a specific threshold for murkier behaviors occurring within intimate relationships is not only asking science to make moral determinations but can result in significant errors in the application of such a threshold.

Consider the following example: A psychologist develops a questionnaire listing 30 behaviors

identified in the literature as psychologically abusive and assigns weights to the items so that the 10 items that seem milder are given a score of 1, the 10 seen as moderate are scored with a 2, and the 10 more severe ones are given a score of 3. A person completing the questionnaire is directed to mark whether or not his or her partner ever carried out each behavior toward himself or herself. The psychologist decides that with a potential score of 0 to 60, any score over 20 would likely indicate the marital partner was a psychologically abusive person, and the person filling out the measure had been psychologically abused. Note that a person could have a score of more than 20 through any combination of milder, moderate, and/or severe behaviors being present. A different psychologist includes this "Measure of Psychological Abuse" in the group of measures that she administers when conducting a psychological evaluation with a person wishing to divorce her or his spouse. If the person being evaluated receives a score of 23, would that psychologist be justified in testifying in family court that the person she evaluated has been psychologically abused by his or her partner? Or that the person's spouse is a psychologically abusive individual? Should a judge make decisions about granting a divorce on grounds of mental cruelty or when assigning common family assets based on that score?

One must consider that there is no basis for knowing (1) whether those behaviors included on the measure truly tap the concept of *psychological abuse*; (2) whether these behaviors occurred only once or more often, which might influence a professional's interpretation of the results; (3) whether it is legitimate to establish a *threshold score* for which we state that we can scientifically determine when someone has achieved the status of being abused or being an abuser; and (4) whether we know that the person completing the questionnaire interpreted the questions correctly. Further, knowing that there is a chance the person completing the measure might harbor problematic motives (e.g., wishing to portray his or her partner as worse than the partner really is), we quickly realize the scientific quandaries present in trying to apply our research methods to a field often lacking in clear definitions and parameters.

Our goals may be better met if we avoid applying societal labels to scientific endeavors when we are seeking to better understand the darker side of interpersonal functioning. We lose nothing by focusing on data and research findings without applying moral designations of severity that cannot be scientifically justified. That is, we can assess milder, moderate, and more severe behaviors within various forms of IPV while avoiding ethical or moral determinations of when a label of *abuse* should be applied to the occurrence of particular behaviors, the combinations in which they occurred, or the frequency at which they were manifested. Science is not hampered by avoiding labels because IPV and VAW can be measured in ways that represent the continua of behaviors, which then can be used in conjunction with other factors to learn what important associations exist at the high, moderate, and lower ends of IPV spectra.

*Most importantly*, knowing that professionals do not work from a universally agreed-upon definition of abuse or violence, it is incumbent upon the consumer of literature to carefully read how the author defines the type(s) of abuse or violence that is being researched. What is the measurement device, how is it scored and interpreted, and how broad or restrictive is the definition? Knowing the answers to these questions will provide a critical lens for viewing the researcher's methodology, results, and conclusions.

## Assumption 2: Because we are measuring behaviors, we can easily document the existence of abuse and violence.

Keeping in mind the difficulty we have in confidently determining which behaviors clearly constitute the major forms of IPV, the idea that we are measuring accurate numbers of distinctly abusive behaviors can be quite misleading. Two

major developments that have occurred since initial attempts to measure IPV began in the 1970s illustrate this.

One of the earliest pioneering attempts to provide a structured, quantifiable method for assessing *physically* forceful behaviors toward an intimate partner (Straus, 1979) was quickly met with concerns that reading a checklist of physically violent acts and reporting how many times they occurred in the last year misses the *context* and the *outcomes*, which very much influence the interpretation of the behavior that occurred (e.g., National Center for Injury Prevention and Control, 2003). Several major contexts have subsequently been proposed that have the potential to significantly alter reported "numbers" of physically violent incidents. The first context involves ascertaining that the physical action that occurred was meant as a hostile or harmful behavior so that only those harmful behaviors would be reported. While this may initially sound strange as a filter for reporting on the use of physical force, when you consider the range of physical behaviors engaged in by couples that are done in a joking manner or as a fun interaction, this screening makes sense. We know that couples playfully wrestle, engage in playful physical dominance as foreplay, and use physical gestures, such as a gentle punch on the arm, to emphasize something that was said. Thus, some researchers have included in their instructions to those filling out physical-violence measures that they should not count behaviors that were done in fun or when playing around (Follingstad & Rogers, 2014); however, it is unknown whether most researchers ask their respondents to make this distinction. Without this type of instruction, numbers could be inflated over the true incidence of physical violence.

The second major context alluded to earlier—and championed by some researchers (e.g., Saunders, 1988)—is whether individuals should divulge the use of physical force that they perceive they engaged in due to self-defense, irrespective of whether the partner used physical force first. The issue appears to be that researchers may not want to label someone a perpetrator who used physical force against a partner out of necessity to protect himself or herself. Historically, the use of *self-defense* required that an action was committed because it was necessary for one's safety. To apply this filter or rule for reporting, must we require that the person who responded with physical force make an assessment as to whether she or he actually *needed* to hit the partner back to protect herself or himself, as opposed to hitting in retaliation? If it were determined that this is an important distinction, researchers would have to insert this layer of judgment for the person to use in deciding whether to report their use of physical force as self-defense because just hitting back after being hit would not suffice for labeling the action as self-defense. As you can imagine, the potentially complicated criteria, as well as the reliance on respondents to make unbiased judgments, quickly thwart reasonable data collection. And to confound matters further, if self-defense were extended to include situations in which you *perceived* that someone was intimidating to you (i.e., perception of dangerousness), how should researchers categorize your actions if you *preemptively* hit the other person while believing you did it in self-defense?

Hamby (2009) has recently focused on the context of *outcome* as a better explanatory factor and predictor than merely counting the occurrence of physically forceful behaviors. Her analysis challenges extremely low rates of female use of physical force in partner violence (e.g., Kurz, 1993; Ostoff, 2002) or extremely high rates (e.g., Archer, 2000) as not accurate and instead presents evidence that women's rates of partner violence are likely in line with other base rates for female violence. For example, using criminal behavior, delinquent behavior, and distressed-couple data, Hamby (2009) challenged data from checklists of IPV that suggest women's physical aggression in intimate relationships is significantly different from other data sources that produce approximately similar rates. In fact, she concluded from a "multimethod analysis of estimates for the incidence and prevalence of intimate and sexual aggression" (p. 149) that

behavioral checklists produce results that are not consistent with research findings using other methodology (Hamby, 2014).

Hamby's (2009) thoughtful analysis has identified several aspects of measurement that may be responsible for spurious outcome data when using checklists of physically violent behavior. She posits that any physical-abuse measure that does not collect data on sexual assault in intimate relationships is omitting highly significant data that need to be in the analysis. She agrees with other researchers that overreporting can occur if parameters for reporting behaviors are not specified to prevent false positives (e.g., stating in the instructions not to include behaviors done in fun). However, Hamby (2009) suggests that collecting information on the *outcome of injury* resulting from physical force is "critically important for understanding gender differences because it is universally agreed that men inflict injury against their intimate partners more than women" (p. 28). An international study of dating violence indicated that 4 times more men inflicted injury than women, and the infliction of severe injury showed a discrepancy of 7% of males versus 0.6% of females (Straus, 2005). It is Hamby's (2009) contention that other impacts or outcomes resulting from physical violence (e.g., psychological, social, or economic) would also better demonstrate gender differences when physical force is used in relationships that are not revealed when individuals respond to checklist formats to assess IPV.

Context would seem to be extremely relevant for understanding whether events are indicative of *psychological* abuse beyond those more egregious acts that appear unjustifiable on their face to any observer. College students responding to a measure of psychological aggression by Follingstad, Coyne, and Gambone (2005) reported that they perceived only 10% of the listed behaviors *that actually happened to them* to be *abusive*. Because these behaviors have been considered to be *abuse* by practitioners and researchers, it is possible that the context in which those behaviors occurred influenced the recipients' perceptions. Thus, it appears that behaviors that are briefly described on measures and that researchers have labeled *psychological abuse* for inclusion on surveys can occur in ways and contexts that people do not consider serious violations of intimate partner behavior (e.g., partner yelled at you).

Another context to consider is whether an abusive behavior occurs *within an exchange with both members of a couple participating* in aversive behaviors that can range from less problematic actions, such as swearing, rudeness, attempts to hurt the other person's feelings, or expressing contempt, all the way through more serious transgressions of intimate interactions. The reciprocal context makes it difficult to know whether a respondent completing a measure asking for relationship history of psychological aggressiveness should report a partner's abusive actions only if he or she did not initiate the exchange, whether he or she should still report being the recipient of abusive actions because the partner reciprocated in kind, or whether he or she should be considered a victim, a perpetrator, or both. Thus, the commonality of reciprocal psychological aggression adds another contextual factor that can influence whether behaviors are reported by research participants and whether researchers are underestimating or overestimating counts of psychologically abusive behaviors.

Although much of the discussion has concerned potential overreporting, other major developments have called into question whether we can know for sure that respondents are acknowledging behavioral occurrences in the area of abuse and violence. Working to measure accurate prevalence of sexual violence on college campuses, Koss and colleagues (Cook, Gidycz, Koss, & Murphy, 2011; Hamby & Koss, 2003) reported an evolution of their measurement based on their research experiences. Historically, the standard language for assessing sexual assault was to ask questions about the occurrence of *rape*. However, when sexual assault was assessed using behavioral descriptions (often based on legal definitions) rather than asking women if they had been raped, more

women endorsed being the victim of legally defined rape than when they were asked directly if they had been raped. Over time, the lesson learned from this area of research has been that accurate assessment of rates of sexual assault must come from instruments that describe sexual assault in technical terms because stereotypical elements of rape embedded in respondents' judgments may prevent them from realizing that they were sexually assaulted (e.g., Fisher, Daigle, & Cullen, 2009). For example, a woman raped by an acquaintance may view *rape* as only occurring if one is attacked by a stranger in public, or a woman who is married may believe that husbands have the right to force sex as an aspect of the marital contract. Thus, neither woman would be likely to endorse an item asking if she had been raped, but she would be likely to endorse an item describing sex that was nonconsensual and forced against her will.

Language used to describe behaviors can impact the accuracy of reported data, especially when the language is value laden but even when a behavioral description is utilized. For example, different measures of physical violence have used various terms for serious physically aggressive incidents such as "beat you up" or "threw you around," but researchers developing these measures likely envisioned a similar extended physical assault. It is not difficult to imagine, however, that respondents could endorse being "thrown around" when completing one measure but not endorse being "beaten up" if they received a different measure using this terminology. This example of the difficulty of trying to compare language for *physical* incidents pales in comparison to attempts to match language across scales assessing psychological abuse when a person's *verbal* treatment of her or his partner is being assessed. Imagine the following item: "Your partner said _____ things to you." Now, individually insert the following adjectives: *spiteful*, *mean*, *nasty*, *hateful*, *cruel*, *abusive*, or *horrible*. First, a researcher cannot be sure whether individuals responding to this subjective language are interpreting the adjectives at the level he or she intended, and second, it is uncertain that researchers themselves know exactly what types of partner statements would fall under each descriptor. If an item of this sort is on a measure designed to identify *abuse*, the researcher cannot be certain of the scientific level that the particular adjective establishes, nor can he or she confirm that this measure of *abusive* verbal behavior toward one's partner matches the threshold of sanctionable actions (i.e., *abusive actions*) by our society. We can predict, however, that different terms across measures would likely produce *varying* rates of endorsement and that they probably cannot be considered comparable. Any behavioral descriptions given to individuals that contain words that are vague or subject to interpretation (by researchers as well as the individuals) typically raise more problems in interpreting the findings than they contribute to our knowledge in this area.

***Issue to be considered.*** It is paramount to understand that having persons respond to simple checklists of negative intimate relationship behaviors may not be the ideal way to provide sophisticated and accurate knowledge regarding rates of abuse or violence. Unfortunately, checklists provide much of the data upon which research literature is based, but reading carefully to determine whether researchers recognize the potential for the inaccuracy of numbers collected through this methodology is important for understanding the significance they attach to their results. Behavioral descriptions for sexual violence that fit legal definitions of criminal behavior appear likely to give the most accurate numbers when measuring sexual assault. Possibly more accurate measures can be developed in the future that may screen for abusive behaviors, whether physical, sexual, or psychological, based on checklist descriptions. Then, if a respondent endorses one of these descriptions, second-level questions could be used to eliminate behaviors for which contexts suggest they are not "true" abusive incidents or for which additional data (e.g., level of resulting injury) are necessary to understand the complexity and nuances of the behaviors or interactions. Ideally, some basic

agreement by researchers in the field of VAW on some terms and behaviors would be very helpful for comparing studies and understanding what factors influence discrepant data.

### Assumption 3: Self-reporting is appropriate, adequate, and accurate for collecting data on IPV abuse and violence.

How else, might you ask, can we get information about events that typically occur in private? Who better to ask than someone who had these things happen to him or her or who actually inflicted them on someone else?

The first critical lens we might use to answer that question is to remember that, except for sexual violence from a stranger or acquaintance, most of the data VAW researchers collect come from interactions with an intimate partner. Reporting on one's own behaviors is difficult enough. But reporting becomes an even more complex endeavor when the behaviors to be reported occurred within a relationship system that has a history, occurred within sequences of interactions, occurred when that person might have motivations to suppress or enhance self-report, and occurred concurrently with feeling states (e.g., anger) likely to influence one's interpretation of events. Not surprisingly, couples do not have good agreement rates when reporting on IPV, especially when they are dissatisfied with their relationship (Christensen & Nies, 1980; Jacobson & Moore, 1981; Margolin, 1987).

Second, historically, many surveys and interviews assessing IPV in the field were victim oriented—that is, they only inquired as to what abusive behaviors had been directed toward the respondent as a victim without any inquiry as to the respondent's own actions. Frequently, measurement of IPV has not exactly been *self-report* but, rather, *other report* because respondents only provided information about their partners' behaviors. Once researchers ventured into collecting victimization *and* perpetration data, the picture of who was the *recipient* and who was the *initiator* of force in a relationship became more blurred, especially with young high school and college populations, where reciprocal use of milder force is not uncommon (e.g., Follingstad, Wright, Lloyd, & Sebastian, 1991; Riggs, Murphy, & O'Leary, 1989). Regarding psychologically aggressive actions, it is not an uncommon pattern in larger representative populations that respondents admit to some use of these tactics but report that their partner has perpetrated *more* psychological abuse toward them (e.g., Follingstad & Edmundson, 2010).

Third, psychological research has demonstrated that persons, when asked about sensitive topics, may skew their responses based on how they wish to be perceived, even when their responses are anonymous. Maisto, McKay, and Connors (1990) stated early on that by virtue of IPV being a sensitive topic (whether you are a victim or perpetrator), disclosure and accuracy of self-report data in this area have been shown to be unreliable at times. As an example to support this contention, when college students were asked if they would report maltreatment of their intimate partner, they admitted they would be much less likely to reveal physical aggression than positive behaviors, much less likely to report their own physical aggression than that of their partner, and much less likely to report having engaged in severe forms of aggression (Riggs et al., 1989). Because of the potential for biased self-report and other report, it is surprising that few studies collecting IPV data assess either underreporting or overreporting tendencies in those providing the data (Follingstad & Bush, 2014; Follingstad & Rogers, 2013) when these may be important covariates to investigate.

***Issue to be considered.*** Generic concerns about self-report in social psychological research are of even greater concern in an area where individuals are reporting on sensitive topics involving intimate partners and their own behavior, which may even be illegal (e.g., assault). If researchers continue to conduct studies to measure IPV using the

same type of self- and other-report checklists, the research should at least include some attempts to assess participant response styles, as well as other personality traits potentially associated with reporting vulnerable data. Ideally, creativity in data collection might result in a methodology with multiple sources of validation so that researchers are not only relying on self-reports (e.g., police records). For example, data on the incidence of IPV resulting from a self-report survey or a structured interview could be matched against ratings by clinicians who have conducted an in-depth interview, which would be one step beyond what this area of research has produced in its unblinking reliance on and acceptance of self-report. Consumers of the literature would be wise to remember the problems with self-reporting and to give greater weight to results from studies that attempted to determine if self-reporting was influenced by biasing factors.

### Assumption 4: Qualitative methods are better for assessing IPV and VAW than quantitative methods.

The first thing to acknowledge when tackling this topic is that comparing these methods almost always implies that the respondent is either an interviewee (qualitative) or a survey taker (quantitative). Thus, this issue is more complicated than simply suggesting that a researcher is asking open-ended versus closed-ended questions because a methodology is crossed with subjective versus objective data collection. Interviews are not anonymous, the interviewer–interviewee interaction influences what is said, and what is focused on in the interview can be too broad or too specific to be useful. Surveys may produce more disclosure if the respondent believes her or his responses are anonymous, but specific questions may be difficult to answer when questioning a person about interpersonal interactions. Thus, advantages and disadvantages can be listed for these and other data collection methods as well (e.g., observation or coding of interactions).

However, it is too easy to forget that different methods can be utilized, at least to some degree, for both quantitative and qualitative data gathering—that is, both interviews and surveys can make use of objective and subjective data collection strategies. For example, after an open-ended question designed to produce spontaneous reactions, an interviewee could be asked to make several ratings on scales described by the interviewer that are pertinent to the subject matter. Researchers do not have to assume a rigid stance of using either qualitative or quantitative methods but should consider sequential use of the two strategies or a mixed-methods design in their investigation that could potentially produce the richest data.

Quantitative methodology has historically been the accepted approach for assessing psychosocial phenomena when the researcher has a specific hypothesis to be tested. Questions for the research respondent are typically presented via a survey, and statistical analyses are conducted on the data to assess the viability of the hypothesis. Although interviews were conducted early in the history of investigating IPV and VAW (D. Martin, 1979; Walker, 1979), they were typically used for anecdotal information because more sophisticated assessment strategies for qualitative data had not yet been developed. Thus, quantitative methods prevailed for many years as the methodology of choice for researchers hoping to establish VAW research as a legitimate field of inquiry and to document incidence and prevalence rates. By 2000, arguments were made claiming that women's experience of violence could not be captured through quantitative means that did not assess the context of violent incidents (Gavey, 2005; White, Smith, Koss, & Figueredo, 2000). In support of this position, Testa, Livingston, and VanZile-Tamsen (2011) stated, "Violence against women is a complex, multi-faceted phenomenon, occurring within a social context that is influenced by gender norms, interpersonal relationships, and sexual scripts" (p. 2).

Qualitative research examining VAW has since proliferated, especially with the development of software designed to aid the researcher's

organization of the qualitative data for analysis. However, it is important to know that these programs are still dependent upon the researcher's subjective and interpretive assignment of data to categories, although they are expected to increase the efficiency of the researcher. Therefore, qualitative findings are not technically empirically validated but, rather, are structured analyses by researchers based on emerging themes.

***Issue to be considered.*** Fortunately, researchers do not have to choose one method over the other. In fact, in response to research design experts calling for studies that involve mixed methods and mixed sources, combining data collection strategies may hold the most promise for IPV and VAW measurement. Some researchers suggest that beginning with qualitative strategies (e.g., focus groups or interviews) allows for more sophisticated quantitative hypotheses and instruments to be developed from the valuable data emanating from the less structured, open-ended methods (e.g., Noonan & Charles, 2009). Other researchers, upon discovering unexpected results from a quantitative study, subsequently used qualitative strategies to probe for greater understanding (e.g., Peterson & Muehlenhard, 2007). More importantly, Testa et al. (2011) suggest that using multiple methods when investigating IPV and VAW is likely to result in greater cross-validation of results, more sophisticated hypotheses to be tested, and potentially greater insight.

## Assumption 5: The same instruments and methodologies for identifying VAW can be used no matter the person's race or ethnicity, sexual orientation, social class, immigration status, or disability status. Violence is universal, so separate analyses by these demographic dimensions are unnecessary.

One unfortunate aspect of the study of IPV and VAW is that Caucasian women were the initial focus of advocates and researchers in the United States and the United Kingdom. In the 1980s, white women were more likely to find their way to shelters, which was a primary source for research participants. More recently, the major critique of the assessment of VAW within the United States has focused on the need for ethnically and culturally sensitive research.

White, Yuan, Cook, and Abbey (2013) clearly made the case for IPV measurement to be sufficient in its range of abuse experiences to capture potential violence directed toward a full spectrum of racial and ethnic minorities. They challenged the stereotype that all women from major racial or even ethnic groups are alike, suggesting that researchers developing VAW measurement may need better knowledge of subgroups' experiences in order to be more accurate in their data collection and interpretation. For example, White et al. (2013) report that there are 566 federally recognized tribes within the racial classification of American Indian and Alaskan Native, for which rates of IPV can vary widely. Ethnic-minority women within the same broad classification may also vary by religion, immigration status, and social class.

A more difficult problem for IPV researchers to tackle may be ethnic women's greater hesitancy to report abuse experiences within their intimate relationships. White et al. (2013) suggested that due to cultural norms or motivations, such women may not encode an abusive experience as problematic, they may not perceive the IPV instrument's language of the questions as matching their experience, they may not disclose due to their perceptions of intimate relationships as private, or they may believe that disclosure of abuse has the potential to harm others' views of their community. Because these elements greatly challenge researchers' abilities to produce accurate measurement, working with ethnic communities in the development and implementation of IPV and VAW measurement appears not only plausible but possibly paramount (White et al., 2013).

In addition to race and ethnicity, we must consider other dimensions of women's experiences that could potentially suggest the need for

different instrumentation, expanded instrumentation, the need for separate analyses, and/or a re-evaluation based on changing cultural norms. Baker, Buick, Kim, Moniz, and Nava (2013) stressed the significance of both including and separately investigating IPV within same-sex couples. A re-examination of violence within same-sex couples is long overdue because historically, same-sex research was fraught with convenience sampling (i.e., using persons in your research sample who were willing to speak with you, rather than drawing a random sample in which every member of the group had an equal chance of being asked to participate in the study), fears of labeling, concerns that identifying oneself as homosexual would lead to a range of problems for the respondent, and distrust that researchers would not be able to appropriately represent same-sex participants' responses. However, Baker et al. (2013) also suggest that the inclusion of same-sex couples in larger assessments of IPV can control for the influence of gender in order to "discover [IPV] is a function of a complex interaction of culture, social structures, social status, and interpersonal dynamics" (p. 182).

Two additional groups that are likely to meet criteria for special consideration for instrumentation or analysis are immigrant women and women with disabilities. Menjívar and Salcido (2002) reported that immigrant women are more vulnerable to violence because they are more likely to be isolated and stressed due to lack of language proficiency and lack of social networks, and they may be more easily threatened if they are not legally documented. Women with disabilities can have unique risks for experiencing abuse or violence that may be a function of the severity, as well as the type, of their disability. Brownridge's (2006) research demonstrated that women with disabilities were significantly more likely to report physical victimization and forced sexual activity by an intimate partner.

***Issue to be considered.*** Researchers who educate on the range of factors influencing a woman's likelihood of experiencing abuse or the types of abuse she may experience are important for reminding us that "one size does not fit all." The first message seems to be that researchers need to expand their sampling techniques to include a greater number of multiethnic women (Mechanic & Pole, 2013) and to work within communities to specifically assess subgroups of multiethnic women (White et al., 2013). When researchers have more diverse samples, it is vital that they assess for potential differences by variables such as ethnicity, sexual orientation, social class, disability, or immigration status. When reading a description of a researcher's sample of participants, question whether the researcher intends for the study findings to be generalizable to a larger population, and determine how successfully the researcher accomplishes that goal. It is highly appropriate to study broad phenomena if the sampling allows for generalizability. It is also highly appropriate to study subgroups within the larger population to determine whether specialized instrumentation or services might be needed. However, researchers should be sensitive to broadening their samples if the result is oversampling of IPV victims who have been the target of most research to date (e.g., Caucasian women). Also, if researchers include reasonable subsamples within their participant group, the consumer of the research should assess whether the researcher made use of this opportunity by comparing the subgroups in the sample based on their specific experiences or factors influencing their experiences.

## Assumption 6: There is consensus as to the most accurate method for scoring behaviors to assess abuse and violence.

Because much of the IPV and VAW research is quantitative and relies on numbers for statistical analyses (e.g., mean scores and frequency counts), the way in which a survey instrument *measures* violence or abuse is more critical an issue than most people realize. To familiarize you with what researchers mean when they report how they *scored* a measure of IPV or

VAW for analysis, consider the many different ways that a person could receive a *score* after completing a measure asking whether she or he had been subjected to physical violence by a partner. First, one person could be assigned a score of 0 if he or she indicated that *nothing* ever happened while another person could be assigned a score of 1 if he or she indicated that *some form* of physical violence happened to him or her. This is called *dichotomizing the sample* in order to place individuals in one of two categories for analysis. Second, because violence measures most likely ask respondents to indicate *how often* a physically violent action was directed toward them, not just whether or not it happened, respondents could receive scores by adding up the number of times physical incidents occurred (*frequency*), by adding up the types of violent actions that occurred (*types*), or even by combining frequency rates with the types of violence to get a combined score (*frequency and type*). Third, the researcher could designate violent physical actions as milder, moderate, or severe and therefore assign respondents a score based on the most violent physical action that happened to them. Using that scoring strategy, a person who only received milder forms would receive a score of 1, persons who experienced any moderate form of physical violence would receive a score of 2, persons experiencing any severe form would receive a score of 3 (*level of severity*), and so forth. Hopefully, it has become apparent that applying a score to a research participant is not a fixed procedure but could be derived in a variety of ways.

Because consensus in defining abuse has not been achieved, it follows that devising the best method for scoring instruments to determine abuse has also not achieved consensus. Setting aside the questionable goal of labeling people as abused versus labeling them abusers, the most common reason for wrestling with scoring decisions has been to establish reasonable comparison groups to investigate differences that might inform ideas for intervention in and the prevention of IPV. However, with this reasonable goal in mind, many scoring efforts unfortunately end up establishing "abused" versus "not abused" groups for comparison. This is typically accomplished by placing anyone who has experienced at least *one* IPV behavior *once* into the abused group and everyone else with zero in every category into the nonabused group. Two-group comparisons are easier to interpret, but unfortunately, they can be quite misleading. This dichotomous classification for victimization (abused versus nonabused) lumps together individuals who have experienced something once with individuals who have had extensive victimization. Also, classifying perpetrators into two groups (abuser versus nonabuser) combines individuals who may have engaged in very minor incidents (e.g., shoved someone once) with individuals who have engaged in extremely serious forms of physical violence (e.g., used a knife on a partner). Several research studies have shown that individuals experiencing very small amounts of IPV generally appear to be much more similar to those experiencing *no* IPV behaviors (Follingstad, Bradley, Laughlin, & Burke, 1999). Thus, dichotomization based on experiencing *any* IPV or perpetrating *any* IPV is likely misleading, and the possibility of missing effects associated with a higher threshold of abuse in a relationship is likely when individuals experiencing that higher threshold are averaged in with those experiencing extremely little IPV.

As mentioned earlier, most studies assessing IPV collect frequency information to learn more than whether IPV occurred at all. Although researchers must consider the accuracy of recall over long periods of time (e.g., "In the last 12 months . . ."), the important question for this discussion is how the frequency data are used to establish whether a participant is classified for research purposes. Not surprisingly, these decisions as to how to score and classify participants can significantly impact numbers and, subsequently, research results. Follingstad, Coker, and Fisher (2014) presented results from a data set of high school students in which students were classified as *victims* based on three methods of scoring. Using participant data for eight types of

harassment, abuse, and violence, people were classified three times as to whether they were a victim based on different scoring criteria: (1) whether the person experienced at least one item on a victimization scale at least one time (at least once), (2) whether the person experienced at least one item three to five times (higher frequency), and (3) whether the person either experienced at least one item three to five times (higher frequency) *or* the person experienced more than one item at least one to two times (more than one type). The authors compared the number of victims produced by each scoring method across the eight scales and found that the first scoring method (i.e., the at-least-once method) always yielded the highest percentage of victims, the second (i.e., higher frequency) always yielded the lowest percentage (about half of the first scoring method), and the third (i.e., either higher frequency or more than one type) was in between. Requiring the highest *frequency* before assigning victim status (i.e., Scoring Method 2) ruled out the most students as *polyvictims* (i.e., victims of more than one type of violence) when students' scores were summed across the eight forms of victimization. Scoring Method 1 was the most liberal at assigning participants in the study to the category of polyvictim. Individuals classified as victims using more stringent criteria (e.g., higher frequency or more types of violence required) were more likely to demonstrate the kinds of problems that can occur following victimization, such as emotional difficulties, which suggest the stringent criteria were more appropriate for classification than the more inclusive criteria, which appeared to classify too many people into victim status. Specificity and sensitivity, which are statistical methods assessing the appropriate classification of individuals into categories, demonstrated that Scoring Method 3 (i.e., higher frequency or experiencing more types of IPV) was consistently better at predicting the negative behavioral and emotional outcomes. What the reader should hopefully take from this detailed analysis is that scoring methods are extremely important not only for classifying individuals as victims but also for the types of research findings that result from that classification.

Even knowing that different scoring methods yield discrepant numbers of individuals for classification as victims, the more significant issue than which algorithm to use is to remember that classification continues to reside in the hands of the individual researcher. Thus, you might find a researcher who classifies *any* incident of even the mildest form of IPV as abuse (e.g., one shove over 30 years of marriage). Another researcher may require a particular *threshold* of either frequency or severity that he or she determines must be met for classification in his or her research. Another researcher may determine that a *pattern* of behavior, usually based on some frequency consideration, must occur rather than a one-time occurrence of an action. Because there is no accepted standard, we know more about these researchers' philosophy regarding victimization than we know that their classification systems are scientific and verifiable.

***Issue to be considered.*** It is somewhat ironic that many studies collecting frequency data on forms of IPV conduct their data analyses using dichotomous groups, with anyone experiencing even one behavioral incident of IPV being combined into the abused group. Using individual participant scores without classifying them into two exclusive groups is one strategy for avoiding the pitfall of assigning participants as victims or nonvictims. Another strategy for avoid dichotomizing participants could be to establish more categories of participants than abused versus nonabused to potentially determine richer results. At the very least, researchers should be encouraged to look at the same data using a variety of classification schemes to present a range of data for comparison. It is recommended that consumers of the IPV and VAW literature pay attention to criteria by which researchers form their comparison groups because researchers' classification schemes establish the lens that the reader must apply when considering the statistical results and their discussion of those results.

## Assumption 7: Reliability and validity have been established for instrumentation used in the field of IPV.

### Reliability

One way that researchers scientifically demonstrate the quality of their methods of measurement is to show statistically that they are reliable. The best synonym for statistical reliability is *consistency*, and there are a variety of ways in which measurement can be assessed for how consistently it measures a construct. *Test–retest reliability* refers to the degree to which respondents' initial responses match their later responses when a particular measure is given twice within a span of time between the two administrations. Test–retest reliability is important for determining if the measure shows stability by producing similar or the same results. This might be especially important if a trait, such as sociability, were being measured because you would expect that a person's extroverted or introverted style would remain consistent over several months' time. If you were measuring respondents' experiences over time, however, you would not necessarily expect that a measure of those experiences (e.g., being sexually assaulted) would show consistency, or test–retest reliability, over time. Therefore, for most IPV and VAW measurement, the type of consistency that is more important in assessing the quality of a scale is whether the items on a scale are consistent with each other for tapping into the phenomenon being measured. This form of consistency is referred to as *internal reliability*. For example, it is likely that a marital partner who is willing to use one psychologically aggressive tactic toward his or her partner would be willing to use other psychologically aggressive strategies. The consistency among the items (i.e., internal consistency) can be statistically calculated to determine whether the items generally appear to be assessing the same concept. Typically, research studies report the latter—the degree of internal consistency (typically called *Cronbach's alpha*) of the IPV or VAW measures that they used. Reports on test–retest reliability are less likely to be found in IPV and VAW research because many studies consist of cross-sectional samples, meaning that they only gathered information at one time period, during which they assessed a cross-section of the population.

Although reporting on the consistency of a measure used in a study is useful information, reliability indicators are not to be mistaken for an indication of validity. If a researcher could show that his or her measure was consistently answered by research participants (i.e., internally consistent), this may be an indicator of people's consistency in viewing and reporting their interpersonal experience. However, we still would not know whether their reporting was accurate. For example, a person with a paranoid orientation may view many actions of her or his partner as intending harm, but she or he might be wrong in this interpretation. Another person might be sensitive to certain cues in his or her relationship because of early life experiences so that he or she pays attention selectively to events that happen and thus reports consistently but in a skewed manner. Thus, even if the internal consistency of a scale in a study reaches the threshold of reasonable reliability standards, that scale must still be assessed for the likelihood that the data are accurate (i.e., valid).

Reliability statistics for assessing how consistently something is measured may not be appropriate at times for IPV or VAW research. In college samples where there are many persons reporting no or almost no dating violence, the data are skewed because the data do not conform to normal distribution in the form of a bell curve. Calculating reliability estimates relies on the assumption that the data being assessed have a normal distribution, so when the data are skewed from many respondents not reporting violence, serious problems for calculating forms of reliability occur (Ryan, 2013). If intimate partners do not engage in a range of types of violence (e.g., different types of physical abuse or multiple acts of psychological maltreatment) but, rather, concentrate on one behavior to mistreat their partner (e.g., treating the person as an inferior as one form of psychological abuse), internal

consistency is seriously reduced to levels that would typically suggest that the measurement device is flawed. If partners engaging in abusive acts do not repeat those acts, test–retest reliability is compromised.

The last form of reliability to be discussed is the statistical indicator of consistency or agreement by two individuals (i.e., *kappa coefficients*), whether they are reporting on events or rating phenomena. This type of reliability for IPV and VAW would apply to estimates of consistency between members of a couple who are reporting on incidents of IPV in their relationship. As mentioned earlier, couples average around 50% agreement for specific events and experiences of violence, which would be considered unacceptable rates of reliability for conducting analyses on data. Again, if the sample being analyzed consisted of many couples with zero incidents of IPV and only a small portion of couples reporting IPV, those data would provide skewed reliability estimates.

***Issue to be considered.*** One of the most thorough considerations of reliability estimates in IPV research was written by Ryan (2013). Her analysis of reliability issues led her to suggest that for reporting research from dating populations, a consideration of "the method used to estimate reliability, the standard deviations of scores, the standard error of measurement, and a sample description" (p. 143) may be more informative in assessing the quality of the data than consulting reliability estimates of internal consistency or rest–retest reliability. Ryan also suggested that a more parsimonious approach to calculating percentage agreement and kappa coefficients with couple data would be to score items or scales for each person as indicating *presence* versus *absence* of partner violence enacted at least once for a designated time period and using those data to calculate reliability estimates, rather than requiring agreement on total frequency of IPV events. Thus, an overall evaluation of the data—rather than rigidly insisting on high internal consistency or test–retest estimates that may not be appropriate for IPV and VAW assessment—appears warranted for determining the quality of the results.

### Validity

For a researcher to be confident that his or her data accurately capture the "true" picture of what is being measured has proven to be a more complex and challenging endeavor than researchers in this area originally anticipated. The types of tests for the validity of measures or scales that researchers created were originally devised for the development of attitude or trait measures (e.g., how anxious a person is or whether a person holds more traditional or liberal views toward women's roles), even though such types of validity are also applied to measures of behaviors at times.

A major test of the validity of a measure is whether it accurately covers the content of the concept being assessed (i.e., *content validity*). This does not mean that every possible question that could be asked must be asked but, rather, that the items in the scale or measure generally address all aspects of the phenomenon (e.g., major forms of psychological abuse are all included in a scale purporting to measure psychological abuse). The second major type of validity requires the researcher to demonstrate that the concept (i.e., *construct validity*) being assessed is accurately captured by the items that are included on the scale. For example, are the items on the psychological-abuse scale actually examples of aversive interpersonal behavior, such as rudeness or boorishness, rather than truly abusive actions? And, a third form of validity entails a demonstration by researchers that their measures accurately predict some outcome (i.e., *predictive validity*). A person experiencing serious forms of psychological maltreatment at high rates on a psychological-abuse scale might be expected to exhibit depression or anxiety.

Similar to reliability, we may not be able to require demonstration of all types of validity for IPV or VAW instrumentation that are typically viewed as necessary during scale development. But before investigating IPV measures' success

in addressing the standard forms of validity listed earlier, a metaissue needs to be deliberated first. Discussions of the validation of IPV measures are not always clear regarding which are the most important types of validity that should be established. Is it most vital that we have comprehensively covered the domain of each form of IPV? That we are able to establish a threshold score by which we can indicate that a person has truly been abused? That a person's score will predict the likelihood that he or she will engage in abusive behavior again? That a particular behavior occurred for which we can predict outcomes of emotional injury or physical harm? Or that we know for sure that a person's report is accurate? These different aims would likely result in a different set of items being developed, potentially different methodologies to research the question, or a different focus for the IPV instrumentation, all of which can relate to various standards of validity. When measuring IPV, the difficulty in determining the exact domain to assess, the interpersonal privacy of the subject matter, the extent of factors impinging on successful data collection, and ambiguity as to *what exactly* is being validated have resulted in less-than-stellar confirmations of validity and sometimes abandonment of that step in measurement construction (Follingstad & Rogers, 2013).

Content validity is the most likely form that can be accomplished in IPV measurement. However, until there is effort among scale developers to work toward consensus of what should be measured or which items are most representative of all possible items for the types of IPV (Follingstad & Bush, 2014), we will continue to be plagued with research results based on the proliferation of currently existing scales. Construct validity raises problems for researchers when devising measures purporting to assess abuse when the threshold for moving some actions from *aversive* into *abusive* is far from being established. Physical, sexual, and psychological IPV cannot easily be used as establishing validity for each other (i.e., *concurrent validity*) because they do not always co-occur in the same relationship. The continued lack of a majority consensus for definitions and thresholds for IPV types (Follingstad, 2007; Maiuro, 2001) limits the ability to test the adequacy of IPV scales in relation to other IPV or related measures (i.e., *concurrent validity*).

Predictive validity, the demonstration that a measure can estimate an expected criterion behavior external to itself (Nunnally & Bernstein, 1994) is a powerful form of validity that does not always map successfully onto IPV assessment. IPV scales measure *past actions* rather than traits or attitudes, and criterion variables may be difficult to establish. For example, should past physical force predict hospital visits, calls to the police, or injuries, which can be low base rate outcomes of low base rate behaviors? Research on IPV tends to favor cross-sectional designs to suggest associations of the history of IPV with variables of interest (e.g., marital satisfaction or preexisting risk factors), but this approach does not establish predictive validity. IPV measurement may fall into a category of cases for which Nunnally and Bernstein (1994) believe predictive validity may not apply because "there is no sensible criterion available or the available ones may be obviously biased and/or unreliable," and the researcher must "fall back on content and construct validity" (p. 109) as validity indicators for the field of inquiry.

A final approach to validity, which may become increasingly useful, is establishing a scale's sensitivity and specificity for accurate classification. Reporting errors may result in false positives (someone being identified as abused or abusing when he or she is not) or in false negatives (someone *not* being identified as abused or abusing when she or he actually is), which certainly compromise accurate IPV assessment. Sensitivity and specificity are statistical ratios that allow for the researcher (and those reliant on that researcher's findings) to understand how accurate a scale is for correct classification.

Likely, the most significant concern for the measurement of IPV—and perhaps the most difficult element to ascertain—is the accuracy of self-reporting of IPV. This is not meant to suggest that people are not truthful but, rather, that many elements of self-reports can be subject to inaccuracy. Often, we ask people to rely on memory for

events occurring "within the last year" or if they "ever happened." If we require respondents to report on behavior that was verbal or subjective, we trust, without knowing whether we should, the accuracy of their representation of events or their interpretation of events occurring in their intimate relationships. Whether there is a "reality" against which a person's reporting of events could be weighed is yet to be determined, but in the meantime, it would be wise to simultaneously assess response styles and personality traits (e.g., social desirability, self-handicapping, overreporting tendencies, interpersonal sensitivity, and neuroticism), which may be useful as covariates when forming conclusions from the data.

***Issue to be considered.*** First, validity cannot be assumed from decent reliability indicators. Second, the accuracy of self-reporting is the most significant issue regarding validity that researchers continue to grapple with but often ignore. Reliance on self-reporting without some checks (e.g., secondary-level assessment once an item is endorsed) or additional information that could provide "corrective" factors in the form of covariates (e.g., response styles or personality traits) should result in a more skeptical lens placed on the results and discussion of them. Third, look for evidence that those developing IPV scales attempted to establish validity through more than a correlation with a different IPV scale (i.e., concurrent validity), which may instead measure a form of consistency (i.e., reliability) rather than validity if the same person reports on the same events on the two scales. More creativity for establishing validity would raise confidence in a measure of IPV, such as identifying a national sample to rate the scale items in terms of severity (e.g., Follingstad, 2011) or using identified criterion groups to demonstrate clinically significant differences on the measure (e.g., Tolman, 1989).

## Conclusion

Contemplating the myriad of issues surrounding the measurement of IPV and VAW should not make one feel discouraged and disheartened. Rather, questioning the methods and subsequent results of VAW research can be an intriguing investigation in which the consumer of research develops an increasing ability to discern quality differences among the results. Hopefully, the sections titled "issue to be considered," following the discussion of each fallacy, have provided some clear directions for how to understand and evaluate individual studies and reports. Thus, having digested the discussions provided in this chapter, the consumer of research should feel less confused when similar studies present dissimilar results because that person knows what to identify in the research that would explain why those differences exist. Possibly, two general fallacies that have been exposed throughout this chapter about psychosocial research are (1) that the publication of a study should not guarantee that its findings should be uncritically accepted and (2) that not all studies are of the same quality.

## Discussion Questions

1. To understand the difficulty in determining the types of partners' actions that should be added onto a *psychological-abuse measure*, discuss what types of behaviors you would include if you had to decide if the behavior was *psychological maltreatment*, *psychological aggression*, *psychological abuse*, *mental cruelty*, or *psychological mistreatment*.
2. Discuss whether you know of racial, ethnic, or cultural differences in defining violence and abuse within intimate relationships that would create dilemmas for devising general violence or abuse scales.
3. Discuss whether you believe that a threshold or absolute standard for determining when violence or abuse within intimate relationships can be established from a legal standpoint, a sociological standpoint, and a psychological standpoint.
4. Discuss whether the various forms of IPV and VAW might be better investigated through different research methodologies.

## Resources for Further Study

Breiding, M. J., Basile, K. C., Smith, S. G., Black, M. C., & Mahendra, R. (2015). *Intimate partner violence surveillance: Uniform definitions and recommended data elements* (Version 2.0). Atlanta, GA: Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/violenceprevention/pdf/intimatepartnerviolence.pdf

## References


Archer, J. (2000). Sex differences in aggression between heterosexual partners: A meta-analytic review. *Psychological Bulletin*, *126*, 651–680. doi:10/1037/0033-2909.126.5.651

Arriaga, X. B. (2002). Joking violence among highly committed individuals. *Journal of Interpersonal Violence*, *17*, 591–610.

Baker, N. L., Buick, J. D., Kim, S. R., Moniz, S., & Nava, K. L. (2013). Lessons from examining same-sex intimate partner violence. *Sex Roles*, *69*, 182–192. doi: 10.1007/s11199-012-0218-3

Brownridge, D. A. (2006). Partner violence against women with disabilities: Prevalence, risk, and explanations. *Violence and Victims*, *12*, 805–822. doi:10.1177/1077801206292681

Christensen, A., & Nies, D. C. (1980). The spouse observation checklist: Empirical analysis and critique. *American Journal of Family Therapy*, *8*, 69–79. doi:10.1080/01926188008250357

Cook, S. L., Gidycz, C. A., Koss, M. P., & Murphy, M. (2011). Emerging issues in the measurement of rape victimization. *Violence Against Women*, *17*, 201–218. doi:10.1177/1077801210397741

DeKeseredy, W. S., & Schwartz, M. D. (1998). *Measuring the extent of woman abuse in intimate heterosexual relationships: A critique of the Conflict Tactics Scales*. Minneapolis, MN: VAWnet.

Dobash, R. P., Dobash, R. E., Wilson, M., & Daly, M. (1992). The myth of sexual symmetry in marital violence. *Social Problems*, *39*, 71–90. doi:10.2307/3096914

Dobash, R. P., Dobash, R. E., Wilson, M., & Daly, M. (1992). The myth of sexual symmetry in marital violence. *Social Problems*, *39*, 71–91.

Fisher, B. S., Daigle, L. E., & Cullen, F. T. (2009). *Unsafe in the ivory tower: The sexual victimization of college women*. Thousand Oaks, CA: Sage.

Fernández-González, L., O'Leary, K. D., & Muñoz-Rivas, M. J. (2013). We are not joking: Need for controls in reports of dating violence. *Journal of Interpersonal Violence*, *28*, 602–620.

Follingstad, D. R. (2007). Rethinking current approaches to psychological abuse: Conceptual and methodological issues. *Aggression and Violent Behavior*, *12*, 439–458. doi:10.1016/j.avb.2006.07.004

Follingstad, D. R. (2011). A measure of severe psychological abuse normed on a nationally representative sample of adults. *Journal of Interpersonal Violence*, *26*, 1194–1214. doi:10.1177/0886260510368157

Follingstad, D. R., Bradley, R. G., Laughlin, J. E., & Burke, L. (1999). Risk factors and correlates of dating violence: The relevance of examining frequency and severity levels in a college sample. *Violence and Victims*, *14*, 365–380.

Follingstad, D. R., & Bush, H. M. (2014). Measurement of intimate partner violence: A model for developing the gold standard. *Psychology of Violence*, *4*, 369–383. doi:10.1037/a0037515

Follingstad, D. R., Coker, A. L., & Fisher, B. S. (2014). *IPV polyvictimization in a Kentucky high school population: Does it matter how victimization is defined?* Paper presented at the 2014 American Society of Criminology Conference, San Francisco, CA.

Follingstad, D. R., Coyne, S., & Gambone, L. (2005). A representative measure of psychological aggression and its severity. *Violence and Victims*, *20*, 25–38. doi:10.1891/vivi.2005.20.1.25

Follingstad, D. R., & DeHart, D. D. (2000). Defining psychological abuse of husbands toward wives: Contexts, behaviors, and typologies. *Journal of Interpersonal Violence*, *15*, 891–920. doi:10.1177/088626000015009001

Follingstad, D. R., DeHart, D. D., & Green, E. P. (2004). Psychologists' judgments of psychologically aggressive actions when perpetrated by a husband versus a wife. *Violence and Victims*, *19*, 435–452. doi:10.1891/vivi.19.4.435.64165

Follingstad, D. R., & Edmundson, M. (2010). Is psychological abuse reciprocal in intimate relationships? Data from a national sample of American adults. *Journal of Family Violence*, *25*, 495–508. doi:10.1007/s10896–010–9311-y

Follingstad, D. R., Helff, C. M., Binford, R. V., Runge, M. M., & White, J. D. (2004). Lay persons' versus psychologists' judgments of psychologically aggressive actions. *Journal of Interpersonal Violence*, *19*, 916–942. doi:10.1177/0886260504266229

Follingstad, D. R., & Rogers, M. J. (2013). Validity concerns in the measurement of women's and men's

report of intimate partner violence. *Sex Roles*, *69*, 149–167. doi:10.1007/s11199-013-0264-5

Follingstad, D. R., & Rogers, M. J. (2014). The nature and prevalence of partner psychological abuse in a national sample of adults. *Violence and Victims*, *29*, 3–23. doi:10.1891/0886-6708.09-160

Follingstad, D. R., Wright, S., Lloyd, S., & Sebastian, J. A. (1991). Sex differences in motivations and effects in dating violence. *Family Relations*, *40*, 51–57. doi:10.2307/585658

Gavey, N. (2005). *Just sex: The cultural scaffolding of rape*. New York, NY: Routledge.

Hamby, S. (2005). Measuring gender differences in partner violence: Implications from research on other forms of violent and socially undesirable behavior. *Sex Roles*, *52*, 725–742.

Hamby, S. (2009). The gender debate on intimate partner violence: Solutions and dead ends. *Psychological Trauma*, *1*(1), 24–34.

Hamby, S. (2014a). *Battered women's protective strategies: Stronger than you know*. New York, NY: Oxford University Press.

Hamby, S. (2014b). Intimate partner and sexual violence research: Scientific progress, scientific challenges, and gender. *Trauma, Violence, & Abuse*, *15*, 149–158.

Hamby, S. (2016). Self-report measures that do not produce gender parity in intimate partner violence: A multi-study investigation. *Psychology of Violence*, *6*, 323–335.

Hamby, S., & Turner, H. (2013). Measuring teen dating violence in males and females: Insights from the national survey of children's exposure to violence. *Psychology of Violence*, *3*, 323–339.

Hamby, S. L. (2009). The gender debate about intimate partner violence: Solutions and dead ends. *Psychological Trauma: Theory, Research, Practice, and Policy*, *1*, 24–34. doi:10.1037/a0015066

Hamby, S. L. (2014). Intimate partner and sexual violence research: Scientific progress, scientific challenges, and gender. *Trauma, Violence, and Abuse*, *15*, 149–158. doi:10.1177/1524838014520723

Hamby, S. L., & Koss, M. P. (2003). Shades of gray: A qualitative study of terms used in the measurement of sexual victimization. *Psychology of Women Quarterly*, *27*, 243–255.

Hegarty, K., Bush, R., & Sheehan, M. (2005). The Composite Abuse Scale: Further development and assessment of reliability and validity of a multidimensional partner abuse measure in clinical settings. *Violence and Victims*, *20*, 529–547. doi:10:1891/vivi2005.20.5.529

Jacobson, N. S., & Moore, D. (1981). Spouses as observers of the events in their relationships. *Journal of Consulting and Clinical Psychology*, *49*, 269–277. doi:10.1037/0022-006X.49.2.269

Jouriles, E., Garrido, E., McDonald, R., & Rosenfield, D. (2009). Psychological and physical aggression in adolescent romantic relationships: Links to psychological distress. *Child Abuse & Neglect*, *33*, 451–460.

Kurz, D. (1993). Physical assaults by husbands: A major social problem. In R. Gelles & D. Loseke (Eds.), *Current controversies on family violence* (pp. 88–103). Newbury Park, CA: Sage.

Maisto, S. A., McKay, J. R., & Connors, G. J. (1990). Self-report issues in substance abuse: State of the art and future directions. *Behavioral Assessment*, *12*, 117–134.

Maiuro, R. D. (2001). Sticks and stones may break my bones, but names will also hurt me: Psychological abuse in domestically violent relationships. In K. D. O'Leary & R. D. Maiuro (Eds.), *Psychological abuse in violent domestic relations* (pp. ix–xx). New York, NY: Springer.

Margolin, G. (1987). The multiple forms of aggressiveness between marital partners: How do we identify them? *Journal of Marital & Family Therapy*, *13*, 77–84. doi:10.1111/j.1752-0606.1987.tb00684.x

Martin, D. (1976). *Battered wives*. San Francisco, CA: Glide Publications.

Martin, E. K., Taft, C. T., & Resick, P. A. (2007). A review of marital rape. *Aggression and Violent Behavior*, *12*, 329–347. doi:10.1016/j.avb.2006.10.003

Mechanic, M. B., & Pole, N. (2013). Methodological considerations in conducting ethnoculturally sensitive research on intimate partner abuse and its multidimensional consequences. *Sex Roles*, *69*, 205–225. doi:10.1007/s11199-012-0246-z

Menjívar, C., & Salcido, O. (2002). Immigrant women and domestic violence: Common experiences in different countries. *Gender & Society*, *16*, 898–920. doi:10.1177/089124302237894

Meyer, E., & Post, L. (2013). Collateral intimate partner homicide. *SAGE Open*, *3*(2). doi:10.1177/2158244013484235

Miller, A. B., Wall, C., Baines, C. J., Sun, P., To, T., & Narod, S. A. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: Randomised screening trial. *BMJ: British Medical Journal*, *348*.

National Center for Injury Prevention and Control. (2003). *Costs of intimate partner violence against*

*women in the United States*. Atlanta, GA: Centers for Disease Control and Prevention.

Noonan, R. K., & Charles, D. (2009). Developing teen dating violence prevention strategies: Formative research with middle school youth. *Violence Against Women*, *15*, 1087–1105. doi:10.1177/1077801209340761

Nunnally, J. D., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Ostoff, S. (2002). But, Gertrude, I beg to differ, a hit is not a hit is not a hit. *Violence Against Women*, *8*, 1521–1544. doi:10.1177/107780102237968

Peterson, Z. D., & Muehlenhard, C. L. (2007). Conceptualizing the "wantedness" of women's consensual and nonconsensual sexual experiences: Implications for how women label their experiences with rape. *Journal of Sex Research*, *44*, 72–88.

Riggs, D. S., Murphy, C. M., & O'Leary, K. D. (1989). Intentional falsification in reports of interpartner aggression. *Journal of Interpersonal Violence*, *4*, 220–232. doi:10.1177/088626089004002006

Ryan, K. M. (2013). Issues of reliability in measuring intimate partner violence during courtship. *Sex Roles*, *69*, 131–148. doi:10.1007/s11199-012-0233-4

Saunders, D. G. (1988). Wife abuse, husband abuse, or mutual combat: A feminist perspective on the empirical findings. In K. Yllö & M. Bograd (Eds.), *Feminist perspectives on wife abuse* (pp. 90–113). Newbury Park, CA: Sage.

Smith, S. G., Fowler, K. A., & Niolon, P. H. (2014). Intimate partner homicide and corollary victims in 16 States: National Violent Death Reporting System, 2003–2009. *American Journal of Public Health*, *104*, 461–466. doi:10.2105/ajph.2013.301582

Straus, M. (2012). Blaming the messenger for the bad news about partner violence by women: The methodological, theoretical, and value basis of the purported invalidity of the Conflict Tactics Scales. *Behavioral Sciences & the Law*, *30*, 538–556.

Straus, M. A. (1979). Measuring intrafamily conflict and violence: The Conflict Tactics (CT) Scales. *Journal of Marriage and Family*, *41*, 74–85. doi:10.2307/351733

Straus, M. A. (2005). *Gender and partner violence in world perspective: Some results from the International Dating Violence Study*. Paper presented at the Ninth International Family Violence Research Conference, Durham, NH.

Straus, M., Gelles, R. J., & Steinmetz, S. K. (1980). *Behind closed doors: Violence in the American family*. New York, NY: Doubleday.

Testa, M., Livingston, J. A., & VanZile-Tamsen, C. (2011). Advancing the study of violence against women using mixed methods: Integrating qualitative methods into a quantitative research program. *Violence Against Women*, *17*, 236–250. doi:10.1177/1077801210397744

Tolman, R. M. (1989). The development of a measure of psychological maltreatment of women by their male partners. *Violence and Victims*, *4*, 159–177.

Vagi, K. J., O'Malley Olsen, E., Basile, K. C., & Vivolo-Kantor, A. M. (2015). Teen dating violence (physical and sexual) among us high school students: Findings from the 2013 National Youth Risk Behavior Survey. *JAMA Pediatrics*, *169*, 474–482.

Walker, L. E. (1979). *The battered woman*. New York, NY: Harper & Row.

White, J. W., Smith, P. H., Koss, M. P., & Figueredo, A. J. (2000). Intimate partner aggression—What have we learned? Comment on Archer (2000). *Psychological Bulletin*, *126*, 690–696. doi:10.1037/0033-2909.126.5.690

White, J. W., Yuan, N. P., Cook, S. L., & Abbey, A. (2013). Ethnic minority women's experiences with intimate partner violence: Using community-based participatory research to ask the right questions. *Sex Roles*, *69*, 226–236. doi:10.1007/s11199-012-0237-0

## Biographical Statement

**Diane R. Follingstad**, PhD, is the director and Women's Circle Endowed Chair in the Center for Research on Violence Against Women and a professor of clinical and forensic psychology in the Department of Psychiatry, College of Medicine, at the University of Kentucky (with a joint appointment in the Department of Psychology). Dr. Follingstad's research in the area of intimate partner violence has covered issues related to battered women, physical dating violence, and factors impacting jury verdicts in cases where battered women killed a partner. Most recently, her research efforts have led to more sophisticated

measurement of psychological aggression and abuse, and she has published a critique regarding the problems of measurement in this field. Her focus on psychological abuse within relationships has led to the investigation of the impact of controlling and interfering partner behaviors on women's recovery from cancer and women's ability to remain abstinent following treatment for opioid abuse. Dr. Follingstad is board certified as a forensic psychologist and has served as president of the American Board of Forensic Psychology. She has also served as secretary of APA's Division of Psychology and Law, which awarded her honorary fellow status. She was awarded the Distinguished Contributions Award in Forensic Psychology from the American Academy of Forensic Psychology in 2009 and the Linda Saltzman Memorial Intimate Partner Violence Researcher Award in 2012 from the Institute on Violence, Abuse and Trauma.

# Are Women Really as Violent as Men? The "Gender Symmetry" Controversy

*Sherry Hamby*

Although it may surprise many people in the general public, there has been a controversy raging for some time in the domestic violence (DV, also known as intimate partner violence) field about whether women are as violent as men (DeKeseredy & Schwartz, 1998; Dobash, Dobash, Wilson, & Daly, 1992; Hamby, 2005, 2009, 2014b, 2016; Straus, 2012; Straus, Gelles, & Steinmetz, 1980). Women compose the vast majority of people who seek help from domestic violence from law enforcement, shelters, and other services. Many research studies also show that there are more female than male victims. However, some types of surveys find gender "symmetry"—similar rates of male and female victimization. Because the reasons for these differences were not well understood, there have been arguments about which data are "right."

To understand this controversy, first we need to address the "straw man" argument that pro-symmetry authors often use. What is a "straw man" argument"? This refers to the debating technique of making your opponent's argument seem more simplistic than it really is. This makes it easier to knock down the argument, just as it is easier to knock down a straw man versus a real one. (Note: Historically, the name for this rhetorical device has been gendered, but for the rest of the article I will refer to it as the *straw person* argument.)

In DV scholarship, the straw person argument that is often used by proponents of symmetry is to speak as if the nonsymmetry position is equivalent to saying that women are *never* violent. This could not be further from the truth. Most measures of violence find female perpetrators, even measures of the most extreme forms of crime, such as murder and rape. No one thinks that women are never violent.

This is what the asymmetry hypothesis actually states: There is overwhelming evidence that men commit domestic violence more often than women, just as they commit other types of violent crime more often than women (Hamby, 2009). Most indicators fall in the range of about 2 to 4 male perpetrators for 1 female perpetrator, although a few indicators have even lower rates of female-perpetrated violence. Just because a single methodology produces a different result does not mean all these other data are wrong—it is more likely that the opposite is true. Recent scientific developments have identified some of the causes of this discrepancy, and these indicate that it was methodological problems in the definition and operationalization of DV that produced the spurious (that is, mistaken or false) result that women are as violent as men (Hamby, 2014b).

## Ways That DV Measures Systematically Underrepresent Men's Violence

*Example 1: Domestic violence homicide*. If you have read elsewhere in this book carefully, you might be surprised to see DV homicide on this list because DV homicide never shows gender parity. Instead, DV homicide rates indicate about 3 to 4 men murder their partner for every female murderer. However, even this statistic systematically underrepresents the homicidal

behavior of male versus female batterers. Why? This is because our current tracking of DV homicide only counts the actual intimate partner as the victim. However, some batterers do not just kill their partner. Some batterers also kill the victim's children, extended family members, and, most commonly, the new partner of women trying to move on (Meyer & Post, 2013; Smith, Fowler, & Niolon, 2014). Unfortunately, these murders are often put into the "acquaintance" or even "stranger" category in homicide statistics (Smith et al., 2014).

Multiple homicides involving intimate partners are far more gendered than single intimate partner homicides. Although it is hard to get precise estimates because of the flaws in our national surveillance, one study of 17 years of Michigan murders found that 100% of the intimate partner murders involving "collateral" victims were perpetrated by men (Meyer & Post, 2013). A more comprehensive understanding of DV-motivated homicide shows that abusive males are responsible for even more intimate-partner-related murders than we already recognize.

*Example 2: The Youth Risk Behavior Survey*. Every year, the Centers for Disease Control (CDC) administers a large, school-based survey on health issues, including teen dating violence (Vagi, O'Malley Olsen, Basile, & Vivolo-Kantor, 2015). Because they ask about so many issues, there was only one question on dating violence. For many years, the answers to this question seemed to support gender symmetry, with approximately 10% of males and 10% of females reporting victimization by a dating partner. However, recently they reworded the item on physical victimization and added an item on sexual victimization. Because the defining characteristic of intimate relationships is intimacy, it is extremely important to include sexual victimization in the operationalization of dating or domestic violence (Hamby, 2009, 2014b). These two changes dramatically altered the results, producing a 20% rate of dating victimization for females and 10% for males. Further, more female than male victimization was found for physical and sexual victimization (Vagi et al., 2015).

Estimates of domestic violence that ignore sexual victimization are systematically underrepresenting female victimization and creating the false illusion of more symmetry than there actually is in domestic violence.

*Example 3: Using Science to Solve Scientific Controversies*. This controversy has raged for 40 years without resolution, in part because both sides have relied on single-measure studies or indirect comparisons of measures (Hamby, 2016). For example, although the YRBS data in Example 2 are compelling, it is possible that other changes, such as changes in the sampling or even random variation, explain the findings. The best way to pinpoint the reasons for different results across methods is through direct comparisons of measures. Surprisingly, there have been few experimental studies of DV measurement. However, recent studies match the results from YRBS. Increasing the precision of the questions or the scoring eliminates gender symmetry (Hamby, 2016; Hamby & Turner, 2013).

One important way to increase precision is to set a threshold to reduce false positives. A *false positive* occurs when a test says a problem exists when that person does not actually have a problem. Some well-known examples include medical screening tools, such as mammograms, which might indicate a woman has breast cancer when further testing, such as with a more precise biopsy, indicates that no cancer is present (Miller et al., 2014).

*(Continued)*

## Are Women Really as Violent as Men? The "Gender Symmetry" Controversy (Continued)

In violence, false positives occur when behavior that is not really violence gets reported to a questionnaire designed to measure violence. Violence is intentionally causing unwanted harm. Unfortunately, many measures do not assess intent, do not assess wantedness, and do not assess harm. Thus, all sorts of events, ranging from accidents to pillow fights, sometimes get reported and scored as violence. Without assessing the intent to cause unwanted harm, even acts such as tackling in football or pushing someone out of harm's way could be scored as violence. Several studies indicate that horseplay and joking around are often mistakenly included in responses to many violence measures (Arriaga, 2002; Fernández-González, O'Leary, & Muñoz-Rivas, 2013; Jouriles, Garrido, McDonald, & Rosenfield, 2009). Research has shown that it is relatively easy to develop reliable, valid tools that reduce these kinds of false positives. One new scale, the Partner Victimization Scale (PVS), produced a lifetime DV rate of 34.1% for women and 18.7% for men, bringing self-report data more in line with what we know about DV from other sources, such as reports to law enforcement, arrests, and help seeking (Hamby, 2016).

### The Next Generation of Domestic Violence Science

Fortunately, the good news is that these problems with old data are fairly easy and straightforward to fix, and several second-generation measures are already emerging, including the freely available, five-item Partner Victimization Scale and the revised YRBS items (Hamby, 2016; Vagi et al., 2015). These more accurate measures will help on several fronts. Hopefully, they will help, at long last, to provide a scientific conclusion to the long-running arguments about gender symmetry versus asymmetry and help everyone in the field recognize that the best description of domestic violence is a pattern of moderate gender asymmetry. Perhaps even more importantly, these more accurate measures will help us better develop and evaluate prevention and intervention. For example, understanding that female victims of serious battering may fear not only for their own lives but also the lives of their children and other loved ones is important to designing better safety planning (Hamby, 2014a). More sensitive and specific assessment will also hopefully help us better understand which prevention programs are best for reducing domestic violence and eventually reducing the burden that domestic violence places on too many families and on societies across the world.

### Discussion Questions

1. How does some research underestimate the level of female victimization?
2. What are some ways to improve the measurement of intimate partner violence?

### Resources for Further Study

Follingstad, D. R., & Bush, H. M. (2014). Measurement of intimate partner violence: A model for developing the gold standard. *Psychology of Violence*, *4*, 369–383.

Hamby, S. (2016). Self-report measures that do not produce gender parity in intimate partner violence: A multi-study investigation. *Psychology of Violence*, *6*, 323–335.

Lerhner, A., & Allen, N. E. (2014). Construct validity of the Conflict Tactics Scales: A mixed-method investigation of women's intimate partner violence. *Psychology of Violence*, *4*, 477–490.

McHugh, M. C., Livingston, N. A., & Ford, A. (2005). A postmodern approach to women's use of violence: Developing multiple and complex conceptualizations. *Psychology of Women Quarterly*, *29*, 323–336.

## Author Biography

Sherry Hamby, PhD, is research professor of psychology at the University of the South and director of the Life Paths Appalachian Research Center. She is also founding editor of the APA journal *Psychology of Violence*. A licensed clinical psychologist, Dr. Hamby has worked for more than 20 years on the problem of violence, including frontline crisis intervention and treatment, involvement in grassroots organizations, and research leading to the publication of more than 150 articles and books. She is the recipient of numerous honors. Her most recent book is *Battered Women's Protective Strategies: Stronger Than You Know* (Oxford University Press, 2014).