Handbook of School Violence and School Safety
International Research and Practice
Shane R. Jimerson, Amanda B. Nickerson, Matthew J. Mayer, Michael J. Furlong

Assessment of Bullying

Publication details
Dewey Cornell, Joanna C. M. Cole
Published online on: 05 Dec 2011

How to cite :- Dewey Cornell, Joanna C. M. Cole. 05 Dec 2011, Assessment of Bullying from: Handbook of School Violence and School Safety, International Research and Practice Routledge
Accessed on: 27 Oct 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Assessment of Bullying

Dewey Cornell
CURRY SCHOOL OF EDUCATION, UNIVERSITY OF VIRGINIA

Joanna C. M. Cole
BOSTON UNIVERSITY SCHOOL OF MEDICINE

Abstract
Despite a large body of research, the field of bullying still needs standard, reliable, and independently validated measures for determining whether a student has been bullied (or has bullied others). This chapter reviews seven common problems with bullying surveys and makes recommendations for improved measurement and validation efforts. A systematic program of research is needed to compare bullying surveys with independent criteria and to assess their sensitivity, specificity, positive predictive power, and negative predictive power.

Assessment is the Achilles heel of bullying research and prevention efforts. Scientific progress is not possible in any field without standardized, reliable, and valid measurement of key constructs, and no intervention program can claim successful reduction in something it cannot accurately measure. Despite the tremendous increase in the study of bullying and efforts to reduce it in schools over the past decade, there continues insufficient evidence for the validity of the survey measures which are the principal means of assessing its prevalence. This chapter will review seven common problems that plague the assessment of bullying—both bullying others and being bullied—and recommend more rigorous standards for identifying students involved in bullying.

Problems in the Assessment of Bullying

(1) Definition of Bullying
In public health fields, it is axiomatic that one must have a standard definition of the problem, disease, or condition under study. However, efforts to study bullying are limited by a lack of consensus and consistency in defining this complex and abstract concept (Cornell & Bandypadhyay, 2010a; Furlong, Sharkey, Felix, Tanigawa, & Green, 2010; Swearer, Espelage, Vaillancourt, & Hymel, 2010). Bullying is a form of aggressive behavior that must be distinguished from the full range and variety of peer aggression and victimization. Although it is common for
children to tease, argue, and even fight with one another, the concept of bullying is intended to identify a particularly troubling form of peer aggression in which an identifiable victim is subject to repeated humiliation by a dominant aggressor. Unless this distinction is maintained, researchers and educators run the risk of obtaining inflated estimates of the prevalence of bullying. Moreover, interventions for bullying differ from those designed to resolve conflicts or mediate disputes between peers of comparable power or status (Olweus, Limber, & Mihalic, 1999).

Many authorities agree that there are three criteria for bullying: (a) action that intentionally causes harm or distress to the victim, (b) an imbalance of power between the aggressor and the victim, and (c) the harmful action occurs repeatedly (Furlong et al., 2010; Olweus, 1991; Swearer et al., 2010). Although most studies define bullying as an intentionally harmful act, many fail to restrict the term to situations that involve an imbalance of power and repetitive occurrence.

For example, a recent national study of child victimization in the United States (Finkelhor, Turner, Ormrod, & Hamby, 2010) did not present a definition of bullying, but measured bullying with two items (Finkelhor, Hamby, Ormrod, & Turner, 2004): (a) “In the last year, did any kids, even a brother or sister, pick on you by chasing you or grabbing your hair or clothes or by making you do something you didn’t want to do?” and (b) “In the last year, did you get scared or feel really bad because kids were calling you names, saying mean things to you, or saying they didn’t want you around?” These items appear to have good psychometric qualities as part of a general assessment of child victimization, but do not encompass the full range of bullying and do not distinguish bullying from other forms of peer conflict. Nevertheless, results from this study were used to report a nationwide decline in bullying (Finkelhor et al., 2010).

The Youth Risk Behavior Survey (YRBS) is administered in thousands of U.S. schools every year and used by the Centers for Disease Control and Prevention (CDC) to monitor national trends for a multitude of health-related behaviors. This survey asks students, “During the past 12 months, have you ever been bullied on school property?” with no definition of bullying (U.S. Department of Health and Human Services, 2009). A summary report of the methodology of this survey (Brener, McManus, Galuska, Lowry, & Wechsler, 2003) contains no information on the reliability or validity of questions about bullying. The only study that attempted to confirm the accuracy of YRBS self-reports concerned student height and weight. This study found that students, on average, over-reported their height by 2.7 inches (Brener et al., 2004).

Another well-respected national assessment of bullying is provided by the School Crime Supplement (SCS) of the National Crime Victimization Survey (DeVoe & Bauer, 2010). This survey is conducted by an interviewer who provides students with a broad definition of bullying: “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already.” The surveyor asks “During this school year, has any student bullied you?” This definition contains no language that distinguishes bullying from other forms of peer conflict, such as arguments between friends of comparable power or status. The survey then inquires about seven different forms of aggressive behavior that also do not distinguish bullying from any other kind of peer conflict. For example, the survey asks “That is, has another student: (a) made fun of you, called you names, or insulted you? (b) spread rumors about you? (c) threatened you with harm? (d), pushed you, shoved you, tripped you, or spit on you?...”

It is not feasible to compare results from surveys that use broad definitions of bullying with those that use a more restrictive definition. For example, a nationally representative study by the National Institute of Child Health and Human Development (Wang, Iannotti, & Nansel, 2009) used the revised Olweus Bully/Victim Questionnaire (BVQ; Olweus, 1996), which has a detailed definition of bullying that specifically excludes situations “when two students of about the same strength or power argue or fight.” Vaillancourt and colleagues (2010) demonstrated
that students were less likely to report being bullied when provided with this kind of more specific definition. Kert, Codding, Tryon, and Shiyko (2009) found that students who completed surveys that presented an adaptation of the Olweus definition of bullying and used the term “bully” reported significantly less bullying of others than students completing surveys that did not employ the term or a definition.

(2) **Student Understanding of Bullying**

Most assessment instruments rely on student reports of bullying, but consider the difficulty for students in deciding whether they have observed or experienced some form of bullying. The concept of bullying encompasses physical, verbal, and social behaviors, but must occur in a context of the aggressor having power or dominance over the victim. The difference in power may be a function of physical size, as in the stereotype of the “big bully” who physically dominates a smaller victim, but more subtle forms of social bullying may occur when the aggressor has an advantage in social status or popularity. The power imbalance between bully and victim could result simply from a disparity in self-confidence or verbal skills.

Physical bullying seems most easy to identify because it involves discrete acts of violence and can be readily observed. However, physical bullying often involves the threat of violence, and threats can be conveyed in words or even implied with a gesture or a glance. Verbal bullying refers to statements that tease or insult the victim, but do not threaten physical injury. Social or relational bullying may be the most varied and indirect form of bullying because it involves the manipulation of friendship patterns and social interactions to demean or exclude the victim from peers (Bjorkqvist, Lagerspetz, & Kaukinian, 1992; Crick & Grotpeter, 1995; Olweus, 1991). All of these varied forms of bullying require students to make judgments that might not be reliable or valid.

A related problem is that bullying must be intentionally harmful behavior. As a result, it can be difficult to distinguish bullying from playful behavior or other actions not intended to cause harm. Common horseplay and teasing among friends can seem like bullying to an observer. Even the participants may have differing perceptions of their behavior; teasing remarks can be misunderstood or taken more seriously than they were intended, and sometimes playful wrestling can escalate into physical bullying. Accused bullies may rely on the defense that they were “just playing around” or “didn’t mean” what they said.

There is little research demonstrating that students have an adequate understanding of the complex concept of bullying and apply that understanding accurately in completing self-report surveys. Several studies suggest that students report lower rates of bullying when they complete surveys that contain a detailed definition of bullying (Kert et al., 2009; Vaillancourt et al., 2010). Even brief educational interventions can have an impact on bullying reports. A study by Baly and Cornell (2011) randomly assigned classrooms of middle school students to watch or not watch a 6-minute instructional video about bullying that stressed the power imbalance present in bullying as distinguished from ordinary peer conflict. Although all surveys included a definition of bullying adapted from Olweus (1996), students who watched the video were 48% less likely to report being socially bullied. Boys who watched the video were 54% less likely to report physical bullying. Researchers are careful to train behavioral observers to achieve reliability in identifying bullying (e.g., Frey et al., 2005), but do not assess whether students can use the concept reliably. Vaillancourt et al. (2010) asked students to give their own definitions of bullying and found that, whereas almost all (92%) of the students included some mention of negative behavior in their definition, relatively few (26%) included the idea of a power imbalance. These observations suggest that an adequate measure of bullying must encompass several different forms of bullying.
as well as the ability to distinguish bullying from ordinary peer conflict between students of relatively equal status.

(3) Timeframe for Bullying

A third problem in bullying assessment concerns the appropriate time period and frequency for questions about bullying. Surveys vary in asking about bullying in the past month, several months, or last year. Furlong and colleagues (2010) recommend asking students about experiences in the past month rather than longer time periods for which their memory might be less accurate. However, the timeframe for recalling events may influence student responses in unexpected ways. For example, Morrison and Furlong (2002) compared two versions of the California School Climate and Safety Survey inquiring about the past 30 days versus the past year. Surprisingly, for many items students reported more victim experiences in the past 30 days than in the past year. Perhaps the timeframe of the question affects the standards that students use to judge whether an event qualifies for reporting, and a shorter timeframe prompts students to consider less serious, more frequent events.

Surveys also vary in whether they ask about the frequency of bullying and in the cut-off point used to qualify as bullying. Solberg and Olweus (2003) recommended a frequency of “2 or 3 times a month or more” (p. 247). Their determination of this cut-off is based on analyses showing that students who endorse bullying at this frequency tend to be more reliably different from students who endorse less frequent bullying when compared on self-report measures of depressive tendencies, self-esteem, and feelings of acceptance by classmates.

(4) Scale Composition

There are potential discrepancies among surveys that rely on single items to measure the prevalence of bullying and those that use multiple items. From a psychometric perspective, multiple-item scales are likely to be more reliable than single-item measures. Olweus (personal communication, August 15, 2010) reported an Item Response Theory (IRT) analysis on Norwegian students that supported use of a scale constructed of eight questions about specific forms of bullying others (verbal, physical, and indirect/relational) on the Revised Olweus Bullying Questionnaire (OBQ). Olweus has also analyzed data on a large sample (> 76,000) of U.S. students in grades 4–11 showing a Cronbach’s alpha of .87 for a 10-item scale that included questions about the conventional physical, verbal, and relational forms of bullying, as well as bullying involving extortion, race, and sexual meaning, and also bullying through electronic communication (cyberbullying) (Olweus, personal communication August 15, 2010).

A natural question is how reports on multiple-item scales compare to use of single items. One study (Vaillancourt et al., 2010) found that the global, single-item questions about bullying others and being a victim of bullying used in the Olweus Bully/Victim Questionnaire yielded good specificity but poor sensitivity when compared to questions asking students about specific types of physical, verbal, social, and cyberbullying. In other words, many students denied being bullied or bullying others on the initial, global questions, but then admitted being bullied or bullying others in response to one of the more specific questions. It remains unclear whether multiple items about specific types of bullying are more accurate or perhaps over-inclusive in their assessment of bullying. Furthermore, why do some students fail to recognize the logical inconsistency between denying general involvement in bullying, yet admitting involvement in a specific form of bullying? Perhaps these students do not apply the restrictions in the definition of bullying to the questions about specific forms of bullying or perhaps the descriptions of specific forms of bullying stimulate new recollections of experiences they recognize as bullying.
Alternatively, some students might not comprehend the questions or take care to answer them consistently.

(5) Types of Bullying

Research on bullying has expanded considerably from a core interest in physical and verbal bullying to more indirect, relational or social forms of bullying (Björkqvist et al., 1992; Crick & Grotberg, 1995; Swearer, 2008), and more recently to include the use of electronic communication to engage in bullying, sometimes termed cyberbullying. Varjas, Henrich, and Meyers (2009) found evidence that cyberbullying may constitute a distinct form of victimization compared to other forms of bullying, while physical, verbal and relational bullying seemed more closely related. However, it should be noted that the term “cyberbullying” has been broadly used to identify many inappropriate or undesirable uses of electronic communication (such as heated arguments between peers) that may not meet the basic criteria for bullying (Kowalski, Limber, & Agatston, 2008).

There is also interest in bullying about sexual matters, which can range from overly aggressive flirtation to efforts to humiliate and threaten someone because of perceived sexual orientation. Ashbaugh and Cornell (2008) found that middle school boys as well as girls reported a high rate of sexual harassment in general. More narrowly, Swearer, Turner, Givens, and Pollack (2008) found that American boys teased as being gay (homosexual) reported higher levels of other forms of abusive treatment (such as being physically bullied, attacked, and excluded) than boys who were bullied for other reasons, and most importantly, experienced greater anxiety and depression. A national school climate survey (Kosciw, Diaz, & Greytak, 2008) found that approximately 85% of youth identifying themselves as lesbian, gay, bisexual, or transgender reported some form of bullying as well as other forms of victimization at school.

There is important work to be done in clarifying whether there are meaningful differences among the various forms of bullying sufficient to justify differences in prevention and intervention efforts. It would seem quite unlikely that the great diversity of experiences identified as physical, verbal, social, cyber, and sexual forms of bullying could be psychologically interchangeable, generated by the same causes and producing the same effects on students. Complicating this examination is the observation that many students experience overlapping forms of bullying, and there may be particular combinations or forms of bullying that do not fall neatly into one of our current linguistic categories. Instruments that measure bullying with a definition that encompasses all forms of bullying run the risk of collecting heterogeneous data that obscure important trends and correlations.

There also may be important differences in the frequency, intensity, and duration of bullying that overshadow differences in its form. Most studies currently pay little attention to such differences and simply identify groups of victims, bullies, and in some studies, students who appear to be both bullies and victims (Solberg, Olweus, & Endresen, 2007). The task of differentiating subgroups of bully victims (or perpetrators of bullying) along dimensions of types of bullying as well as frequency, intensity, and duration of each type may outstrip the sample size and measurement capacity of most studies, as well as generate daunting analytic complexity.

(6) Anonymous Versus Confidential Reporting

Even if students understand the concept of bullying, they may be unwilling to report it. Many studies have noted the code of silence that discourages students from identifying bullies or seeking help for bullying (Oliver & Candappa, 2007; Smith & Shu, 2000; Unnever & Cornell, 2004). There is some evidence that students are more willing to seek help for bullying from
teachers when they perceive the school climate to be more supportive and believe that teachers will take effective action (Eliot, Cornell, Gregory, & Fan, 2010; Williams & Cornell, 2006).

Many bullying surveys are administered on an anonymous basis because of the belief that students will be more forthcoming (Olweus, 2010; Solberg & Olweus, 2003). On the contrary, several studies challenge the assumption that an anonymous survey is necessary in order for students to admit involvement in bullying. Chan, Myron, & Crawshaw (2005) studied students who were randomly assigned to take a survey anonymously or to write their names on the survey. There were no statistically significant differences in rates of endorsement of behaviors that reflected bullying others and being victims of bullying (i.e., hitting, teasing, and lying about other students) between these two groups. O’Malley, Johnston, Bachman, and Schulenberg (2000) examined differences between anonymous and non-anonymous adolescent reporting of drug use and illegal behaviors (i.e., stealing and weapon carrying) on the Monitoring the Future survey. In this study, one group answered the survey anonymously and the other group was not assured of anonymity and was required to report names and addresses to researchers, but was told that their answers would be confidential. Again there were little or no group differences in endorsement rates for sensitive information.

More studies examining the difference between anonymous and confidential administration of bullying surveys are needed, especially because anonymous survey administration prevents researchers from linking survey results to other student information, such as academic performance. Our research group at the University of Virginia has routinely used data from confidential survey administration to show that reports of bullying involvement are stable over time, correspond to peer perceptions, and predict behavioral and emotional adjustment (Branson & Cornell, 2009; Carlson & Cornell, 2008; Cole, Cornell, & Sheras, 2006; Cornell & Brockenbrough, 2004; McConville & Cornell, 2003).

(7) Survey Screening for Invalid Responses

Another reporting problem is that some students may not take the survey seriously and could engage in either careless or intentionally dishonest responding to survey questions. Because bullying and victimization generally occur in a small percentage of students, careless or inattentive marking by students will increase their frequency (e.g., random responses to a yes-no question will generate a 50% prevalence rate). Provocative adolescents will produce even higher rates if they intentionally choose the most extreme or unexpected response. Furlong, Sharkey, Bates, and Smith (2004) identified a group of respondents on the Youth Risk Behavior Surveillance survey (YRBS) who claimed to have carried a weapon to school six or more times in the past month (the most extreme response). Many of these weapon-carrying students also claimed to exercise every day, eat plenty of carrots, and drink lots of milk, but also to make frequent suicide attempts, use heroin, sniff glue, and take steroids. The researchers concluded that these students gave extreme responses to survey questions regardless of item content.

Validity screening procedures can substantially reduce estimates of the prevalence of student involvement in high-risk behavior such as fights, drug use, and gangs. In a survey of 10,909 middle and high school students, Cornell and Loper (1998) found that one-fourth of the surveys failed to meet validity screening criteria that included detection of students who omitted demographic information, marked a series of items all in the same way, and gave inappropriate answers to validity questions (e.g., answering “No” to “I am telling the truth on this survey”). The deletion of invalid self-report surveys reduced estimates of fighting at school from 29% to 19%, drug use from 25% to 15%, gang membership from 8% to 5%, and carrying a knife at school from 18% to 8%.

Cross and Newman-Gonchar (2004) screened three different school surveys for the presence of inconsistent responses to items with the same content (e.g., answering “never” when asked
what age they belonged to a gang and “yes” to the question, “Have you ever belonged to a
gang?”) and extreme responses (e.g., claiming to have used LSD 20 or more times in the past 30
days). Surveys with three or more inconsistent and/or extreme responses were identified as “sus-
psect.” Although fewer than 5% of students were identified as suspect, deletion of these suspect
surveys reduced estimates of students carrying a handgun at school by a magnitude of 30—from
3.2 to 0.1%. In one high school, the proportion of students who reported having been bullied
was 46%, but after suspect surveys were removed from the sample, the proportion dropped to
25%, which is a reduction of more than 45%. In other words, the error in survey results that
could be attributable to inconsistent and extreme responding—not considering other forms of
error such as limitations in memory or concentration—is larger than the typical reductions
reported by many bully prevention programs (as estimated by Smith & Ananiadou, 2003 and
Ttofi & Farrington, 2009).

Cross and Newman-Gonchar (2004) observed striking diff erences in survey results between
schools that used trained versus untrained survey administrators. In some cases the teachers were
not given adequate instructions or advance notice that they would be administering a lengthy
survey in their classroom. Although this was not a controlled study, their post hoc observations
were provocative; 28% of surveys obtained by untrained administrators failed to meet valid-
ity standards, whereas only 3% of those obtained by trained administrators were considered
invalid. The findings by Cross and Newman-Gonchar (2004) raise concern that teachers should
be appropriately prepared to administer any classroom survey and that survey results should be
carefully scrutinized for invalid responses. Teachers must be motivated to administer the survey,
they must have clear instructions and adequate time, and they must be willing and able to engage
the students so that they take the survey seriously and put forth a reasonable eff ort to complete it
accurately. The survey should not be so laborious that students lose interest, fail to concentrate,
or begin marking answers at random.

The Need for More Rigorous Validity Studies

These seven problem areas demonstrate the need for more rigorous measurement research on the
assessment of bullying, especially the use of student self-report surveys. Because most studies use
anonymous surveys, researchers cannot validate student reports with independent sources of cor-
roboration. Instead, researchers have presented data on the internal consistency of student reports
or its correspondence with other criteria obtained from the same self-report instrument. For
example, two studies of the Olweus BVQ (Solberg & Olweus, 2003; Solberg, Olweus, & Endre-
sen, 2007) show that two self-report questions (one about bullying others and another about
being bullied) correlate with subsequent questions about specific forms of bullying on the BVQ
and also additional survey questions about externalizing and internalizing behaviors. Although
these findings show consistency in student reports on the BVQ, they are not suffi  cient to establish
validity. In each analysis, there is a lack of independence between the bullying questions and the
criterion items. Results from such analyses are confounded and inflated to an unknown degree
by shared method variance—correlations produced by the measurement method rather than the

Shared method variance can occur when two measures or scales drawn from one reporter
(e.g., a student) are correlated with one another because of the consistency in how the reporter
answers questions rather than an underlying relationship between the constructs being mea-
sured. Method effects for a self-report bullying survey can be produced by a variety of factors,
including student reading level, mood, and attitude toward completing the survey. For example,
a defensive student might mark items to minimize his or her involvement in bullying, while
a more provocative student might choose to exaggerate his or her claims. In both these cases,
an apparent relationship between bullying and some criterion measure (e.g., externalizing or internalizing behavior) simply represents student consistency in answering questions. Thus, studies that rely solely on correlations between survey results from the same source do not achieve the independence between predictor and criterion variables necessary to provide rigorous evidence of self-report accuracy. Shared method variance is often overlooked or discounted in bullying research, but is widely recognized as a serious problem in other fields (Podsakoff et al., 2003).

**Effect on Program Evaluation Efforts**

The fundamental question is, How can we know that a student’s report about bullying is accurate? Unreliable measures can have a devastating effect on efforts to evaluate the effectiveness of a bullying prevention program. For example, a meta-analysis of bullying prevention studies (Ttofi & Farrington, 2009) suggested that effective programs typically reduce the number of victims of bullying by approximately 20%. To illustrate this effect, they offered the example of 100 students in a bullying prevention program versus 100 students in a control condition. There might be 25 victims among the 100 students in the control condition versus 20 victims in the prevention program, which is a reduction of 5 victims or 20%. An effect of this size, even if statistically significant, is fragile and might not be detected with unreliable measures. If only a handful of students in the treatment condition over-reported victimization, the apparent treatment effect would be obscured and it would appear that the program was unsuccessful.

It is usually assumed that measurement errors are randomly distributed across groups, so that the net effect is negligible in most cases. However, measurement error due to over- or under-reporting might not be randomly distributed in bullying studies. One problem noted by Smith and Ananiadou (2003) is that a bullying prevention program may sensitize students to report bullying, resulting in an increase in self-reports that does not reflect the actual decline in bullying achieved by a successful intervention. Ironically, the more effective the program, the more pronounced the sensitization—resulting in the paradoxical finding that student reports of bullying have increased rather than decreased. In a noteworthy randomized controlled study of the *Steps to Respect* program, Frey and colleagues (2005) found that the intervention produced declines in bullying and argumentative behavior according to trained observers, but no changes in self-reports of those behaviors.

A contrasting problem can occur when students respond to demand effects of the bullying prevention program. After being repeatedly lectured and reminded about the undesirability of bullying, some students may learn to disavow and deny bullying on subsequent administrations of the survey, even if they have not changed their behavior with peers. If the self-report survey is the only source of information to indicate a reduction in bullying, it would not be possible to rule out that an apparent decline in bullying simply represents student acquiescence to an expected response pattern. If both sensitization and demand effects occur, one might see an increase in self-reports of bullying shortly after a program is initiated, due to sensitization effects, followed by a decline as students become compliant with program expectations.

Research on bullying owes much to the groundbreaking work of Scandinavian researcher Daniel Olweus, who in 1983 implemented a nationwide program to reduce bullying in Norway. His program has served as the model for bully prevention efforts throughout the world, including the United States (Olweus et al., 1999). The Olweus program received international recognition because it produced reductions in bullying of 50% or more. Critical to this recognition is that Olweus (1996) had an assessment instrument that could document the success of his program. Without a reliable measure of bullying, Olweus would not have received due credit for the success of his program. However, there has been limited published information on both the reliability and validity of the Olweus BVQ. Few studies have conducted traditional criterion-
related validity studies using independent criteria for bullying. Because the BVQ is so widely used and serves as a model for other instruments, it is especially important that its validity be established to a high degree of methodological rigor.

Lee and Cornell (2010) published a direct examination of the concurrent validity of the BVQ using independent criteria. Based on a relatively small sample of 202 middle school students, they found that self-reported bullying on the BVQ was significantly correlated with peer nominations for bullying (r = .12, p < .05) and academic grades (r = -.15, p < .05), but not disciplinary infractions. Self-reported victimization was significantly correlated with peer nominations for victimization (r = .42, p < .01) and academic grades (r = -.12, p < .01). More studies with larger and more diverse samples are needed, but these results lend only modest support to the validity of the BVQ.

**Need for Diagnostic Accuracy Statistics**

Although the correlations obtained by Lee and Cornell (2010) are similar in magnitude to other correlations obtained in cross-informant studies (Achenbach, McConaughy, & Howell, 1987), they do not provide strong evidence for test accuracy, which is a higher standard than simply examining the size of correlation coefficients. The assessment of a measure’s accuracy requires more than a statistically significant correlation, because a correlation does not indicate what percentage of students who are victims (or bullies) in a school are accurately identified as victims (or bullies), or what percentage of students who are not victims (or bullies) in a school are accurately identified, either.

It has long been a standard practice in epidemiology, medicine, and psychological assessment (McNeil, Keller, & Adelstein, 1975; Swets, 1986) to assess the accuracy of diagnostic tests for the presence/absence of any condition by calculating four important statistical indices: the sensitivity, specificity, positive predictive power, and negative predictive power of a test score. Sensitivity refers to the capacity of the test to identify all cases with the condition in the study population, while specificity refers to the capacity to identify cases without the condition. The sensitivity of a bullying survey for victims would be the percentage of all bullied students that are identified by the survey. The specificity of the survey would be the percentage of all nonvictims of bullying whom the survey identifies as nonvictims. Together, sensitivity and specificity tell us how many students in the underlying population (e.g., student body of the school) are accurately identified as victims or nonvictims.

In contrast, positive predictive power and negative predictive power tell us about the accuracy of test results. Positive predictive power is the percentage of cases identified as victims of bullying by the survey who are actually victims, while the negative predictive power is the percentage of cases identified as nonvictims who are actually nonvictims. In combination, these four statistics are essential to assess the accuracy of any assessment instrument designed to identify the presence/absence of any population condition such as bullying.

When bullying is measured on a continuous scale, there may be questions about the appropriate cut-point to use for classifying students as victims (or perpetrators) of bullying. Receiver Operating Characteristic (ROC) analysis is a standard procedure for examining the sensitivity and specificity of a test across the full range of possible cut-points (Swets, 1986). ROC analyses are needed for both self-report and peer-report measures.

The critical requirement for conducting research on the accuracy of an assessment instrument is the existence of some independent standard of truth. A major problem in bullying research is that there is no standard or definitive procedure for deciding that a student has actually been bullied (or has bullied others). From a practice standpoint, school authorities make this determination by interviewing students or observing their behavior and making a judgment.
We recently compared middle school student responses to a self-report bullying survey with counselor interviews of those students (Cornell & Bandyopadhyay, 2010b). In this school, students were advised that if survey results indicated that they were victims of bullying, the researchers would notify one of the two school counselors to speak with them. A total of 43 students identified themselves as victims of bullying “about once a week” or “several times a week” in the past month. The students and counselors were familiar with the concept of bullying through the school’s participation in the Olweus Bullying Prevention program. After interviewing the students, the counselors concluded that only 24 (56%) of the 43 students were accurately self-reported as victims of bullying. The remaining students included two who had been victims of bullying prior to the timeframe of the survey question, 13 who were involved in a peer conflict that was not bullying, such as a disagreement with a friend, and four who claimed to have marked the survey in error.

**Peer Reports of Bullying**

Peer reports represent an important alternative to self-report assessment of bullying. The peer report or nomination method usually involves asking students to identify classmates who match a descriptive statement or definition (Pakaslahti & Keltikangas-Jarvinen, 2000; Ladd & Kochenderfer-Ladd, 2002; Nabuzoka, 2003; Cornell & Brockenbrough, 2004). There are variations to this method in which students are asked to nominate a fixed number of classmates or to assign frequency ratings (e.g., never, sometimes, often) to each of their classmates.

Olweus (2010) pointed out a number of measurement problems associated with peer reports, including the lack of consistency across studies in how peer ratings are obtained, the difficulty of determining standard cutting scores to identify bullies and victims, and uncertainty whether peers will know that some of their classmates are being bullied. These legitimate concerns merit systematic standardization research. For example, the study design used by Solberg and Olweus (2003) to identify cutting scores for self-reported frequency of bullying could be used to determine the optimal number of peer nominations that identify a victim or perpetrator of bullying. Indeed, many of the measurement concerns about peer reports mirror the problems of self-report and speak to the broader need to establish the validity—and especially the accuracy—of bullying measures using independent criteria.

Nevertheless, many reliability and validity questions about peer reports have been addressed in the related field of peer aggression research, where peer nomination is a highly regarded, standard method of identifying aggressive students and their victims (Hawker & Boulton, 2000; Ladd & Kochenderfer-Ladd, 2002; Leff, Kupersmidt, Patterson, & Power, 1999; Pellegrini, Bartini, & Brooks, 1999; Perry, Kusel, & Perry, 1988). Peer reports have been found useful in assessing a wide variety of emotional and behavior problems, including peer aggression, delinquency, hyperactivity, anxiety, and depression (Huesmann, Eron, Guerra, & Crawshaw, 1994; Weiss, Harris, & Catron, 2004). Peer reports have also been used in studies of indirect peer aggression, social exclusion, and interpersonal problems (Crick & Bigbee, 1998; Hill, Zrull, & McIntire, 1998; Pakaslahti & Keltikangas-Jarvinen, 2000).

Several studies demonstrate the value of peer reports in studies of bullying and victimization (Ladd & Kochenderfer-Ladd, 2002). A Korean study (Kim, Leventhal, Koh, Hubbard, & Boyce, 2006) examined the causal relationship between bullying nominations by peers and self-reported psychopathology. Based on data collected from seventh- and eighth-grade students on two occasions 10 months apart, the authors found that symptoms of emotional and behavioral maladjustment were better understood as a consequence of being bullied (or bullying others) rather than a contributing factor. Childhood victimization is linked to subsequent emotional disorder, social maladjustment, academic difficulties, and other adverse outcomes (Hawker & Boulton, 2000;
Ladd & Ladd, 2001; Rigby, 2001). Rigby (2001) demonstrated that both self-reported and peer-reported victimization were associated with suicidal ideation in Australian adolescents.

A Canadian study by Chan (2006) asked victims of bullying to name their aggressors. Among 435 students named as bullies were 94 students who engaged in serial bullying, defined as bullying more than one victim. The serial bullies accounted for nearly 70% of the total victim population and were also the most likely to engage in physical bullying. This study demonstrated that peer nominations can yield quantitative estimates of the scope of a student's bullying that are not obtained from self-report.

Thunfors and Cornell (2008) investigated the peer popularity of American middle school students identified as bullies or victims. A middle school sample consisting of 379 students completed the standard peer nomination form on the SCBS (Cornell, 2011) and an additional question asking them to identify up to 10 of the most popular boys and girls in their grade. Over the course of the school year, the students identified in the fall as bullies by at least two classmates earned lower grades, accrued more discipline violations, and were more likely to be suspended from school than other students. However, bullies received substantially more endorsements as popular students (mean 20.6) than victims (3.6) or other students (12.8).

Branson and Cornell (2009) found modest correspondence between self- and peer reports of bullying others ($r = .18$) and bullying victimization ($r = .32$) in a sample of 355 middle school students. Despite their limited agreement, both self- and peer-reported bullying/victimization were predictive of school maladjustment, as measured by discipline referrals, school suspensions, and aggressive attitudes. Regression analyses showed that both self and peer report provided unique predictive value for school maladjustment measures. These findings support the view that multiple measures, preferably using multiple informants, are most useful in the assessment of bullying (Juvonen, Nishina, & Graham, 2001; Ladd & Kochenderfer-Ladd, 2002; Swearer et al., 2010).

The simple advantage of peer report over self-report is that scores are based on data aggregated from multiple sources, which tends to decrease measurement error and produce a more reliable result. Although some children may make an erroneous judgment about a classmate's involvement in bullying, the combined judgment of the group should be more accurate. The most common reservation about peer nomination is that teachers are reluctant to ask students to make judgments about one another, fearing that the exercise will stimulate teasing or cause anxiety. Based on our observations and experience with peer nominations over the past 10 years, with appropriate classroom supervision, a peer nomination survey can be administered without such problems.

The choice of instrument to assess bullying can have a powerful effect on the nature and course of the intervention. If school authorities or researchers choose to rely exclusively on an anonymous self-report measure to assess the prevalence of bullying, they may learn how much bullying is occurring, but they will not know who is being bullied and by whom. With this limited knowledge, interventions naturally focus on schoolwide rules and curriculum units on bullying. Meanwhile, counselors must wait for bullying to be reported before they can intervene with specific students. Unfortunately, many students do not seek help for bullying and teachers often do not detect it (Eliot et al., 2010; Unnever & Cornell, 2003, 2004). Moreover, as noted above, self-reports may be inflated if students do not understand the concept of bullying, are sensitized to report bullying by initial intervention efforts, or mark the survey in a careless or intentionally exaggerated manner.

The peer nomination method may be especially valuable for school counselors attempting to reduce bullying because it can focus and expedite their intervention efforts by identifying specific students who are perceived to be victims and perpetrators of bullying. Assisted by this information, counselors can observe or interview these students to confirm their involvement.
Peer nomination data in which one or more students are identified as a victim by a large number of classmates can be useful in convincing teachers that bullying is a problem in their classroom and motivating them to take action.

**Future Directions in Bullying Assessment**

In conclusion, the field of bullying research needs standard, reliable, and independently validated measures for determining whether a student has been bullied (or has bullied others). Bullying measures should incorporate standard definitions that use the three criteria of intentional harm, power imbalance, and repetition. A systematic program of research is needed to establish the most reliable and valid way to survey students about bullying, including investigations of student comprehension of survey questions and their willingness to respond in a truthful manner.

The use of anonymous surveys has prevented researchers from validating student responses against independent criteria. However, there is evidence that confidential surveys may be a viable alternative to anonymous surveys and can produce findings that are not inflated by shared method variance and other confounding factors. Peer nominations also represent an important assessment strategy that should be more widely used. Peer nominations have the virtue of providing school staff with direct and verifiable information about victims of bullying. Nevertheless, research on peer reports regarding bullying is needed to address similar issues of reliability, validity, and accuracy.

Efforts to validate surveys should go beyond correlations with other measures to include assessment of accuracy using conventional indices of sensitivity, specificity, positive predictive power, and negative predictive power. These statistics will make it possible to assess how accurately surveys measure the prevalence of bullying and whether they can be used with confidence to evaluate the effectiveness of prevention programs. However, this kind of validation effort would require substantial research to develop an independent criterion for bullying. Counselor interviews, or perhaps research interviews conducted with trained interviewers, could provide one source of criterion information, although such efforts would require reliability training and careful implementation.

**Table 22.1  Implications for Practice: Recommendations for the Assessment of Bullying**

1. There is insufficient evidence for the validity and accuracy of instruments used to assess the prevalence of bullying and bully victimization.
2. Shared method variance is a common problem in bullying research; survey measures require validation with independent criteria.
3. Validity studies should investigate the accuracy of bullying measures with attention to sensitivity, specificity, and positive predictive value, and negative predictive value.
4. To enhance survey reliability and validity:
   - Students need systematic education about the definition of bullying before completing survey measures;
   - Teachers should be well-prepared for survey administration and engage students in taking the survey seriously;
   - Student surveys should be screened for careless or exaggerated responding.
5. Peer nominations can be a useful additional source of information.
References


