Associations between Peer Nominations, Teacher Ratings, Self-Reports, and Observations of Malicious and Disruptive Behavior

David B. Henry and The Metropolitan Area Child Study Research Group

Assessment 2006; 13; 241
DOI: 10.1177/1073191106287668

Associations Between Peer Nominations, Teacher Ratings, Self-Reports, and Observations of Malicious and Disruptive Behavior

David B. Henry
University of Illinois, Chicago

The Metropolitan Area Child Study Research Group

This study evaluates the validity of two aggression scales for predicting observations of malicious or disruptive behavior at school. Subgroups of a sample of 1,560 children (age 8.6 ± 1.5 years) were assessed using (a) peer nominations of aggression, (b) teacher reports on the Teacher Report Form (TRF) of the Child Behavior Checklist (CBCL) Aggression scale and the peer nomination items, or (c) self-reports on the peer nomination items. Criteria were observations of physical, verbal, initiated, retaliatory, malicious, and disruptive behaviors. Teacher report peer nominations predicted observed physical, verbal, initiated, and retaliatory aggression and disruptive behavior. Peer nominations predicted physical aggression, verbal aggression, initiation and disruptive behavior, and TRFs predicted verbal, initiated, and disruptive behavior. Self-reports did not significantly predict any behavior. Implications for assessment of aggression are discussed.

Keywords: test validity; aggression; behavioral observations

Criterion-related validity is the extent to which scores on a scale correspond to performance (Cascio, 1991, p. 154), and the difficulty in obtaining appropriate criteria for evaluation of measures is referred to as the criterion problem (Jenkins, 1946). Studies have found varying degrees of cross-informant correspondence between scales measuring externalizing, undercontrolled, or disruptive behavior (Achenbach, McConaughy, & Howell, 1987; Epkins, 1995a, 1995b, 1996; Epkins & Dedmon, 1999), but few have evaluated the validity of these scales against behavioral criteria among children. Criterion-related validity of such measures is difficult to establish due to the absence of a gold standard for its measurement. Although *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994) definitions exist for depression, inattention, and conduct problems, no such definition exists for aggression. Moreover, observers require substantial training to distinguish malicious aggressive behavior from disruptive or playful behavior (Pepler, Craig, & Roberts, 1998). Because of these obstacles and others, aggression scales are often validated against other aggression scales.

There are a variety of explicit or implicit definitions for aggression in the literature and, as a result, different measures of aggression may differ substantially in their

The Metropolitan Area Child Study Research Group is a collaboration of (in alphabetical order) Leonard Eron, University of Michigan; Nancy Guerra, University of California, Riverside; David B. Henry, University of Illinois at Chicago; L. Rowell Huesmann, University of Michigan; Patrick Tolan, University of Illinois at Chicago; and Richard VanAcker, University of Illinois at Chicago. This research was supported by grants from the National Institute of Mental Health (NIMH) and the Centers for Disease Control and Prevention (CDC). Communications regarding this article should be sent to David B. Henry, Department of Psychiatry, Institute for Juvenile Research, University of Illinois at Chicago, 1747 W. Roosevelt Road, Chicago, IL 60608; e-mail: dhenry@uic.edu.

Assessment, Volume 13, No. 3, September 2006 241-252
DOI: 10.1177/1073191106287668

content. Bandura and Walters (1963), following Buss (1961, p. 1), defined aggression as "the delivery of a noxious stimulus to another" (p. 113). Others add that for behavior to be regarded as aggressive, it must be intended to harm another (Carlson, Marcus-Newhall, & Miller, 1989; Coie & Dodge, 1998). Others define aggression in terms of the emotional state of the aggressor, differentiating between hostile and instrumental, or "hot" and "cold," aggression (Bushman & Anderson, 2001). Still other distinctions focus on whether aggressive behavior is proactive or reactive (Dodge & Coie, 1987), direct or indirect (McIntyre, 1972), physical or verbal (Minturn, 1967), or relational (Crick & Grotpeter, 1995).

The variation in definitions of aggression is reflected in the content of aggression measures. For this reason, one important question concerns variation in criterion-related validity due to differences in the content tapped by different measures. Measures include features of disruptive, oppositional, hyperactive, or inattentive behavior as well as malicious behavior intended to harm another. For example, the popular Child Behavior Checklist (CBCL) Aggression scale contains items on demanding attention, property destruction, talking out of turn, and jealousy, as well as other items that primarily measure direct physical aggression. The 14-item Aggression; scale of the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) taps bragging and poor sportsmanship as well as physical, verbal, and relational aggression. The self-report Aggression scale (Orpinas & Frankowski, 2001) includes items on anger, verbal aggression, relational aggression, and physical aggression; and the Peer Nominations of Aggression (Eron, Walder, & Lefkowitz, 1971) includes physical and verbal aggression items as well as items tapping hyperactivity and indirect aggression. In this article, we use the term "aggression" to refer to the broad class of behaviors tapped by aggression scales. We use the narrower term "malicious" to refer to behavior intended to harm another.

A second important question concerns differences in criterion-related validity across different sources of information. The degree of correspondence among ratings of peers, teachers, parents, and self-reports varies. Huesmann et al. (1996) found a strong correlation ($r > .6$) between the Teacher Report Form (TRF) Aggression scale (Achenbach, 1991) and Peer Nominations of Aggression. However, Achenbach et al. (1987), in a study of cross-informant consistency across multiple constructs and measures, found only moderate correlations between peer, teacher, parent, and self-reports. Cross-informant correlations were stronger for externalizing problems than for internalizing problems. Osterman, Bjorkqvist, Lagerspetz, and Kaukiainen (1994) found that children rated themselves as less aggressive than their classmates rated them on measures of physical, verbal, and indirect aggression. Studies in clinical samples suggest closer correspondence between parent and teacher measures than has been obtained in community samples (Biederman, Mick, & Faraone, 1998; Epkins, 1996). Many prevention programs depend on measures of aggression to identify high-risk students for selective interventions (e.g., Conduct Problems Prevention Research Group, 1999; Metropolitan Area Child Study Research Group, 2002). It is important to determine the extent to which different sets of items and different sources accurately predict aggressive behavior.

Few studies have investigated the criterion-related validity of aggression measures, and even fewer have undertaken to validate measures of aggression against strict behavioral standards (Faraone & Tsuang, 1994), and most of these have studied adults. For example, Herzberg (2004) evaluated the criterion-related validity of a questionnaire on traffic behavior against traffic accidents, citations for aggressive driving, and driver's license suspensions. Criterion validity in studies of aggression measures in children often refers to validation against other rating scales (Baity & Hilsenroth, 2002; Casat, Norton, & Boyle-Whitesel, 1999; Flanagan, Alfonso, Primavera, Povall, & Higgins, 1996; Gadow, Sprafkin, & Grayson, 1990; Hilton, Harris, & Rice, 2003; Palme & Thakordas, 2005; Wang et al., 1997; Waschbusch, Willoughby, & Pelham, 1998). At times, the same questionnaire with different sources has been used (Epkins, 1995a; O'Connor, Archer, & Wu, 2001). Some studies have validated aggression measures by comparing scores from labeled groups with those of the general population (Garcia-Leon et al., 2002; Huesmann, Lefkowitz, & Eron, 1978). Only a few studies of children have evaluated aggression measures against behavioral criteria. One example is the study by Atkins, Pelham, and Licht (1989), who used observations of playground behavior to find that the Conners' Teacher Rating Scale could differentiate between observations of aggressive behavior and hyperactivity.

In this study, we evaluate the criterion-related validity of peer nominations of aggression, teacher ratings on two aggression scales, and self-reports of aggressive behavior against observations of malicious and disruptive behavior conducted in structured and nonstructured school settings. The analyses reported in this study were first conducted as preparation for the development of a composite aggression scale for a major prevention study (Huesmann et al., 1996; Metropolitan Area Child Study Research Group, 2002). In this report, we have expanded these analyses to compare the criterion-related validity of peer nominations, teacher ratings, and self-reports using identical items from the Peer

Nomination Inventory. In addition, we evaluate the association between subcategories of observed aggressive behavior and teacher reports on a widely used measure of aggression, the Teacher Report Form of the CBCL. The criterion measures used in this study were three mutually exclusive composite categories of aggressive behavior constructed from direct observations. The categories were malicious versus disruptive behavior, based on intent to harm, physical versus verbal aggression, and initiated versus retaliatory aggression, based on whether the behavior followed provocation. It is also important to evaluate differences in validity by gender, ethnicity, and age. Differential validity has been found for measures of academic achievement (Shields, Konold, & Glutting, 2004) and self-esteem (Nugent, 1994). At least one study found that items of an aggression rating scale performed differently in assessing aggression levels in men compared to women (Dunbar, 1999). Aggression levels also tend to differ by ethnicity, which may be due to confounding of the effects of ethnicity with those of stress due to poverty and inner-city residence (Guerra, Huesmann, Tolan, VanAcker, & Eron, 1995). Older children have more friends outside of their schools and spend more time with them (Cairns & Cairns, 1994, p. 122). In this study, we test for gender, ethnic, and age differences in predictive validity.

TABLE 1
Demographic Characteristics (*N* = 1,560)

Variable	N	%
Gender		
Female	621	39.8
Male	938	60.2
Unknown	1	0.0006
Ethnicity		
African American	703	45.2
American Indian	5	0.32
Asian	15	0.96
Non-Hispanic White	263	16.88
Hispanic	546	35.04
Other	6	0.39
Unknown	22	1.4
First-grade year		
1995	15	0.96
1994	1	0.06
1993	129	8.27
1992	264	16.92
1991	267	17.12
1990	290	18.59
1989	304	19.55
1988	289	18.53
Unknown	1	0.06

METHOD

Participants

The sample for the present study is composed of 1,560 children in Grades 1 to 6 who participated in the Metropolitan Area Child Study (MACS; Metropolitan Area Child Study Research Group, 2002), a multiwave prevention study. At the time of each child's entry into the study, risk for aggression was assessed using a composite measure derived from Peer Nominations of Aggression (Eron et al., 1971) and teacher reports on the CBCL (Achenbach, 1991). Students were designated as high risk if their scores on this composite aggression measure were in the upper 50% of students in their schools.

Students were included in the sample for the present study if they had nonmissing values on behavioral observations and any of the four measures of aggression described below at a single wave of measurement. All participants assented and had parental consent for participation. The study was approved by the Institutional Review Board of the University of Illinois at Chicago.

The demographic characteristics of the sample are reported in Table 1. As can be seen there, the sample was approximately 60% male, reflecting the fact that in early years of the MACS study, only high-risk students were observed, and the high-risk sample was predominantly male. The sample was primarily of African American and Hispanic ethnicity, reflecting the ethnic composition of the schools in the MACS study. The sample was drawn from 16 original schools in one large city and one mid-size city in a major metropolitan area. By the end of the MACS study, 23 schools had participated at some point in the 7 years of the project.

As the MACS study developed, different measures of aggression were employed at different times and for different purposes, resulting in the complex assessment schedule shown in Table 2. Of the 1,560 children in the sample, no subject had data on all four predictors because self-reports and teacher reports on the peer nomination items were never administered in the same wave of measurement. Five hundred seventy-five (575) had data on one predictor, 445 on two predictors, and 560 had data on three predictors. In this study, we selected subsamples of the overall MACS sample designed to maximize the number of participants who had nonmissing data on a predictor and on behavioral observations at the same wave of assessment. By this criterion, 905 participants were available for the TRF Aggression scale, 971 for the peer nominations, 773 for children's self-reports, and 416 for teacher predictions of peer nominations. We tested for

TABLE 2
MACS Assessment and Intervention Schedule

	Year						
Cohort	*Spring 1991*	*1991-1992 AY*	*1992-1993 AY*	*1993-1994 AY*	*1994-1995 AY*	*1995-1996 AY*	*1996-1997 AY*
Cohort 5	4th—PCG	5th-G_2B-y1-CGB	6th—y2-**PCGB**	7th	8th—TCSA	9th	10th—SA
Cohort 4	3rd—PCG		5th-**PCTB**-y1	6th—y2-PCGB	7th	8th—TCS	9th
Cohort 3	2nd—PC	3rd-GB-x1-**PCGB**	4th—x2-PCG	5th-B-y1-G	6th—y2-PCGB	7th	8th—TCSA
Cohort 2	1st—PCG	2nd-G_2B-x1-**PCGB**	3rd—x2-PCGB	4th	5th-B-y1-G	6th—y2-PCGB	7th
Cohort 1	Knd	1st—TC	2nd-PTB-x1	3rd-x2-**PCGB**	4th	5th-B-y1-CG	6th—y2-PCGB
Cohort 0		Knd	1st	2nd-**TGB**-x1-G	3rd-x2-PCGB	4th	5th
Cohort -1			Knd	1st—PC	2nd-**TGB**-x1-G	3rd—x2-PCGB	4th

NOTE: Assessments in bold type were used in this investigation. Each cell shows grade, fall assessments, intervention, and spring assessments. For assessments: C = Teacher's Report Form ratings of aggression; P = peer nominations of aggression; T = teacher predictions of peer nominations; G = self-report general assessments; B = behavior observations. For interventions: x1 = early intervention, 1st year; x2 = early intervention, 2nd year; y1 = late intervention, 1st year; y2 = late intervention, 2nd year.

differences in the gender and ethnic compositions of these samples using binomial and polytomous categorical models predicting gender and (African American/White/Hispanic) ethnicity by sample. These analyses found no gender differences and two small but significant differences in the ethnic compositions of the samples. The subsample for the self-reports had a higher than expected proportion of Hispanic children, $\chi^2(1, N = 1{,}560) = 5.7$, $p < .05$, and the sample for the TRFs had a lower than expected proportion of African American children, $\chi^2(1, N = 1{,}560) = 12.5$, $p < .05$, controlling for all other subsample memberships.

Measures

Criterion measure: Observations of student behavior. We gathered observations of student behaviors using a method for real-time, multiple-entry observations on laptop computers (Repp, Karsh, VanAcker, Felce, & Harman, 1989; VanAcker, Bush, Grant, & Getty, 1992). Observers coded class structure, academic interactions between students and teachers, and social behavior by pressing specific keys for time codes (such as structure) and event codes (such as a child hitting another child). The computer program recorded the beginning and end times of the time codes and the time in seconds after the beginning of the session when event codes were entered. The program also timed the observation sessions and signaled the observer when 20 min had elapsed.

In the first 3 years of the MACS study, direct observations of student and teacher behaviors for all targeted high-risk students were conducted in the fall and spring of each school year of intervention. In latter years, observations were made of all students regardless of risk status. Each observation session was 20 min long and done twice at randomly selected times during the school day to maximize the probability of gathering samples of low base-rate behaviors. To minimize reactive arrangements, observers spent large blocks of time (e.g., entire mornings or afternoons) following classes through class time, gym, recess, or lunch and transitions between these periods, observing several children during each block of time. Observation sessions could, thus, include multiple types of structured or unstructured time. Children to be observed were selected at random; thus, neither the teacher nor students were aware of which child was being observed or when the individual 20-min observation sessions began or ended. We attempted to observe each student at one structured (e.g., classes or gym) and one unstructured (e.g., transitions, lunch, or recess) time on different days. Of the time spend in observation, 41.3% was in large group structure, 31.0% in individual seat work, and 12.7% in transitions. The remainder was divided among six other structure types. Less than 1% of students (0.52%) were observed twice on the same day. The behavioral codes included information related to the instructional setting and structure, student task-related behavior, student compliance with teacher requests, social interaction, academic responding, and teacher feedback. In this study, we focused on codes for student malicious and disruptive behavior.

The observers were 31 male and female graduate students in education who participated throughout the 7 years of the MACS study. Prior to collecting observations, observers were trained for 1 month, first on code definitions until they could pass a criterion test on code definitions with 95% accuracy and then on videotapes and live observations until they achieved an 85% overall percentage agreement with another observer on three consecutive live observation sessions. Observers were not judged to be reliable until no individual code fell below an 80% overall percentage agreement within a ±2-s window. Random reliability checks were conducted across codes

TABLE 3
Observational Codes and Definitions

Base Observational Code[a]	*Composite Codes*			*Notes*
Physical malicious initiated	Physical	Initiation	Malicious	"Malicious" was coded for behavior obviously
Physical malicious retaliation	Physical	Retaliation	Malicious	intended to harm another, such as hitting
Verbal malicious initiated	Verbal	Initiation	Malicious	another child.
Verbal malicious retaliation	Verbal	Retaliation	Malicious	
Physical disruptive initiated	Physical	Initiation	Disruptive	"Disruptive" was coded for behaviors disruptive
Physical disruptive retaliation	Physical	Retaliation	Disruptive	of class but not intended to harm, such as
Verbal disruptive initiated	Verbal	Initiation	Disruptive	throwing a book on the floor.
Verbal disruptive retaliation	Verbal	Retaliation	Disruptive	

a. Codes were included for these acts directed at other children and at teachers. Peer and teacher targets are combined for these analyses.

TABLE 4
Percentage of at Least One Occurrence of Each Composite Observational Code, by Sample

	Sample			
Composite Code	*TRF* n = *905*	*PRAGG* n = *973*	*CSR* n = *773*	*TPPRAG* n = *416*
Physical	9.39	9.78	6.08	10.34
Verbal	8.95	8.96	5.69	8.41
Initiation	9.94	10.50	6.73	11.54
Retaliation	7.07	6.80	4.14	6.25
Malicious	5.86	5.56	2.98	6.33
Disruptive	10.83	11.23	8.02	11.06

NOTE: TRF = Teacher Report Form of the Child Behavior Checklist; PRAGG = Peer Nominated Aggression scale; CSR = Child Self-Report; TPPRAG = Teacher Predictions of Peer Nominations.

throughout the data collection period on approximately 10% of the total sessions to avoid decay in reliabilities.

Base codes were summed to produce composite categories for physical or verbal, initiated or retaliation, and malicious or disruptive behavior. Table 3 reports the base codes and how they were aggregated, and Table 4 reports the frequencies of the composite codes by sample.

Kappa coefficients were calculated for each reliability session by comparing the data streams on a second-by-second basis (MacLean, Tapp, & Johnson, 1985). Kappa coefficients for the composite codes used in this study were derived from 108 reliability sessions and are reported in Table 5. As can be seen there, the kappa reliabilities for the composite behavioral codes used in this investigation ranged from .79 to 1.0.

Counts of behaviors from two observation sessions on each student were summed to form scores on each base code. Base codes tallied acts according to whether the behavior was physical or verbal, initiated or in retaliation for provocation, and malicious or disruptive. Correctly coding malicious versus disruptive behaviors required substantial training and testing of observers. Malicious behavior was coded if the behavior clearly was intended to deliver a noxious stimulus to another person, whereas acts were coded as disruptive if the behavior delivered a noxious stimulus but was either not intended for another person or intended to be playful. A child throwing a pencil intending to hit another child would have been coded as malicious, but a child throwing a pencil across the room not aimed at another child would have been coded as disruptive.

Predictors: Measures of aggression. The predictors used in this study were the Aggression scales of the Peer Nomination Inventory (Eron et al., 1971) and the CBCL (Achenbach, 1978). The peer nomination items were completed by peers, by teachers, and as self-reports. For ease in comparing the content of these measures, we list their items in Table 6. Means, standard deviations, pairwise correlations, and their associated numbers of observations are reported in Table 7.

Classmates' nominations were used to assess aggression through the Peer Nomination Inventory (Eron et al., 1971). Children were asked to nominate their classmates who engaged in the behaviors listed in the left column of Table 6. Children were allowed to nominate as many classmates as they wished. The Peer Nomination Inventory was designed for school-age children and assesses popularity, rejection, victimization, and prosocial behavior in addition to aggression. The relevance, reliability, and validity of the inventory have been demonstrated in numerous studies (Eron et al., 1971; Eron, Laulicht, Walder, Farber, & Spiegel, 1961; Guerra et al., 1995; Huesmann & Eron, 1986; Huesmann, Eron, Lefkowitz, & Walder, 1984; Lefkowitz, Eron, Walder, & Huesmann, 1977). Internal consistency of the Peer Nominated Aggression scale (PRAGG) by coefficient alpha was .98 in this sample, and 2-year stability was .62. The PRAGG scale provides ratios indicating the number of times a child was nominated for each of 10

TABLE 5
Kappa Reliabilities, Means, Standard Deviations, Minima, Maxima, and Spearman Correlations of Composite Observed Behavior Codes

						Spearman Correlations					
Observed Behavior	*Kappa*	M	SD	*% With Zero*	*Max*	*1*	*2*	*3*	*4*	*5*	*6*
1. Physical	.905	0.16	0.47	93.67	4	1.00					
2. Verbal	.859	0.16	0.70	94.01	14	.40	1.00				
3. Initiation	.797	0.21[a]	0.82	93.02	17	.68	.68	1.00			
4. Retaliation	1.000	0.10[a]	0.37	95.01	5	.59	.56	.30	1.00		
5. Malicious	.795	0.06[b]	0.25	97.26	3	.63	.38	.48	.44	1.00	
6. Disruptive	.887	0.25[b]	0.85	91.27	14	.65	.80	.78	.60	.27	1.00

NOTE: The means are the mean sums of observed behaviors in two 20-min observation sessions. Maxima are maximum numbers observed in two 20-min observation sessions. All correlations are significant at $p < .001$.
a. Wilcoxon $Z = 5.17$, $p < .001$.
b. Wilcoxon $Z = 8.46$, $p < .001$.

physically and verbally aggressive behaviors (e.g., pushing and shoving other children, taking other children's belongings without asking, or yelling at other children) compared with the number of times the child could have maximally been nominated by peers.

We also asked students to rate their own behavior using the same items as the peer nominations on the Child Self-Report of Aggression (CSRPRAG). Item stems were changed from "Who . . ." to "How often do you . . ." and children answered each item on a 4-point scale consisting of *never*, *not very often*, *sometimes*, and *a lot*. Internal consistency for this measure was .83 by Cronbach's alpha and, as can be seen in Table 7, it correlated significantly with peer-nominated aggression ($r = .26$, $p < .01$) and TRF aggression ($r = .21$, $p < .01$) scores.

Teachers rated students' aggressive behavior on two scales. The Teacher Predictions of Peer Nominations (TPPRAGG; Huesmann, Eron, Guerra, & Crawshaw, 1994) asked teachers to predict what proportion of a child's peers would nominate the child for each of the questions of the Peer Nomination Inventory. The TPPRAGG inventory has high internal consistency for measuring aggression ($\alpha = .97$) and correlates strongly with peer nominations (Huesmann et al., 1994). In this sample, it correlated at .46 with peer-nominated aggression and .58 with TRF aggression.

Teachers also rated students' aggression on the TRF of the CBCL (Achenbach, 1978, 1991). The reliability and validity of the TRF are well-documented. In this sample, the internal consistency by coefficient alpha for the TRF Aggression scale was .96.

Demographic variables. Information on gender and ethnicity was initially gathered from teacher reports on the TRF. Throughout the MACS study, this information was confirmed and corrected using school records, direct observations, and parent reports.

Procedure

Table 2 details the intervention and assessment schedule for the Metropolitan Area Child Study. Three increasingly intensive combinations of intervention were conducted throughout a period of 7 years in a total of 23 schools in a major metropolitan area. One fourth of the schools were assigned to a no-treatment control condition. Another fourth were assigned to receive a general classroom enhancement intervention (Level A). Another fourth were assigned to receive the Level A intervention plus a small group social skills training intervention for high-risk students (Level B). The remaining fourth received the Level B intervention plus a family intervention for the families of high-risk students (Level C). Schools were randomly assigned to the four conditions. Interventions were delivered in second and third grades (early intervention, x in Table 2) and in fifth and sixth grades (late intervention, y in Table 2). One cohort had the early intervention delivered in third and fourth grades rather than second and third grades.

For this study, we used data from assessments of seven birth cohorts of children. In Table 2, we report the cohorts and grades assessed by Peer Nominations (P), TRFs (C), Teacher Predictions of Peer Nominations (T), Self-Reports of Aggression (G), and Behavioral Observations (B). Slightly more than half of the children (51.8%) in the MACS study participated in both Grades 2 to 3 and 5 to 6. Nevertheless, this investigation uses only a single assessment from each child's data.

Teacher predictions of peer nominations were administered as a substitute for peer nominations under two conditions: (a) when there were less than 5 participating students in a class (46 classrooms and 133 students, 32%) and (b) in two cohorts for whom peer nominations were not administered due to cost factors (33 classrooms and 283 students, 68%). In the final years of the MACS study,

TABLE 6
Items on the Peer Nominations of Aggression and the Child Behavior Checklist Aggression Subscale

Peer Nominations of Aggression	*TRF*
Who takes other children's things without asking?	Cruelty, bullying, or meanness to others
Who starts a fight over nothing?	Destroys his or her own things
Who pushes and shoves other children?	Destroys property belonging to others
Who gives dirty looks or sticks out their tongue at other children?	Disobedient at school
Who often says "Give me that!"?	Easily jealous
Who makes up stories and lies to get other children in trouble?	Argues a lot
Who does things that bother other children?	Gets in many fights
Who says mean things?	Defiant, talks back to staff
Who are the children you wish were not in class?	Disturbs other people
Who is always getting into trouble?	Talks out of turn
	Physically attacks people
	Bragging, boasting
	Demands a lot of attention
	Showing off or clowning
	Explosive or unpredictable behavior
	Demands must be met immediately; easily frustrated
	Stubborn, sullen, or irritable
	Sudden change in mood or feelings
	Talks too much
	Teases a lot
	Temper tantrums or hot temper
	Threatens people
	Unusually loud
	Disrupts class discipline
	Screams a lot

NOTE: All Peer Nomination Items were also on the Teacher Predictions of Peer Nominations (TPPRAGG) and the Children's Self-Report of Peer Nominated Aggression (CSRPRAG). TRF = Teacher Report Form of the Child Behavior Checklist.

TABLE 7
Means, Standard Deviations, and Pairwise Correlations Among Predictor Variables

				Correlations		
Predictor	M	SD	T *Score*	*1*	*2*	*3*
1. Peer nominations of aggression	0.31	0.19		—		
2. TRF aggression						
r	12.14	12.64	59.90	.56***	—	
n				878		
3. Self-reports of aggression						
r	2.23	0.64		.26***	.21***	—
n				486	418	
4. Teacher predictions of peer nominations						
r	15.12	19.84		.44***	.42***	—
n				161	163	0

NOTE: Teacher predictions of peer nominations and self-reports on the peer nomination items were not administered in the same wave of measurement. TRF = Teacher Report Form of the Child Behavior Checklist.
***$p < .001$

teacher predictions of peer nominations were administered to all students.

The analyses reported here were begun in preparation for developing a composite aggression scale to be used as the primary outcome variable for the MACS project. Their aim was to determine the validity of peer nominations, teacher report forms of the CBCL, and self-reports of aggression for possible inclusion in the composite

aggression measure. Analyses of the Teacher Predictions of Peer Nominations were added and the original analyses updated for this report.

RESULTS

In Table 5, we report the means, standard deviations, percent with zero, and maxima for the aggregate behavioral observation categories and the Spearman correlations among the behavioral codes. Malicious behavior was observed less frequently than disruptive behavior. Table 5 also reports comparisons using Wilcoxon Z test between mutually exclusive aggregate categories of behavior. These comparisons showed that initiation was more frequently observed than retaliation and that disruptive behavior was observed more frequently than malicious behavior, in which a student directed a noxious stimulus at another with intent to harm.

We used generalized linear models with a negative binomial probability distribution for error and logarithmic linkage to estimate the effects of the aggression scales on the aggregate categories of observed behavior. Such models may be more appropriate than models assuming normal or Poisson error distributions because of the extreme skewness of the behavioral observations, because the observational data were counts of behaviors, and because the sparseness of these behaviors might result in misfit (underdispersion or overdispersion) in Poisson models (Adejumo, Heumann, & Toutenburg, 2004). We believed negative binomial models were preferable to recoding the data to binary indicators and using logistic regression because the negative binomial regression models preserved variation among individuals displaying the coded behaviors. These models, fit through SAS PROC GENMOD, included terms controlling for gender, ethnicity, and grade.

To produce standardized estimates of the effects of each aggression measure on the observed behaviors, we adapted a formula for conversion of chi-square statistics to Pearson's r that is used in meta-analysis (Wolf, 1986). The ratio of the square of the parameter estimate (β-hat) to the square of its estimated standard error (σ-hat) is the Wald chi-square, and the square root of a single degree of freedom chi-square divided by the number of observations is equal to r. Therefore,

$$\gamma = \sqrt{\left(\frac{\left(\frac{\hat{\beta}^2}{\hat{\sigma}^2}\right)}{N}\right)}, \text{ keeping the sign of } \hat{\beta}.$$

Using this formula, we calculated point estimates and 95% confidence intervals for the effect sizes from the parameter estimates and the 95% confidence intervals for the parameter estimates, respectively. The parameter estimates, standard errors, effect sizes, and confidence intervals for each aggression measure predicting each type of observed behavior are reported in Table 8.

As can be seen in Table 8, three of the measures were significantly associated with most forms of observed behavior. None of the measures predicted malicious (as distinct from disruptive) behavior significantly. Only the TPPRAGG had significant and positive effect sizes for all other forms of observed behavior. The TRF Aggression subscale had significant positive effects on verbal, initiated, and disruptive behavior. PRAGG predicted all observed behaviors except retaliation and malicious behavior.

Self-reports of aggression (CSRPRAG) had no effect size whose 95% confidence interval did not include zero. The strongest positive effects of self-reports were for verbal behavior and retaliation, but these effects were not significant. Both teacher measures and the peer nominations had stronger effects for predicting disruptive behavior than for malicious behavior.

Moderated Effects

We also evaluated the possibility of differential validity by gender, ethnicity, and grade. For these analyses, we used generalized linear models to which terms for the moderator (gender, ethnicity, or grade) and the interaction between the moderator and the aggression scale being tested were added. Gender and ethnicity were dummy-coded for this purpose so that moderated effects could be straightforwardly interpreted. In moderated analyses by gender, males were coded 1 and females were coded 0. In one contrast for the moderated analyses by ethnicity, non-Hispanic White was coded 1, and in the other, Hispanic was coded 1. African American was the comparison level, coded 0, in both contrasts. These analyses found no evidence for moderated effects by gender or ethnicity. There were two marginal effects for moderation by grade. The interactions between grade and self-reports suggested a decreasing effect of self-reports on verbal aggression with increasing grade, $B = -0.32$, $\chi^2(1) = 3.41$, $p < .10$, and a decreasing effect of self-reports on disruptive behavior with increasing grade, $B = -0.26$, $\chi^2(1) = 3.12$, $p < .10$.

Supplementary Analyses

We could not test the association between all of the aggression measures and behavioral observations with a single sample because no sample had both self-reports and

TABLE 8
Estimates and Effect Sizes From Generalized Linear Models Predicting Observed Behaviors From Aggression Scales

Predictor	N	*Estimate*	SE	*Effect Size* (r)	*95% Confidence Interval for* r	
					Lower	*Upper*
			Physical			
CSRPRAG	773	–0.01	0.12	.00	–0.07	0.07
PRAGG**	971	1.36	0.49	.09	0.03	0.15
TRF	905	0.01	0.00	.04	–0.02	0.11
TPPRAGG**	416	0.02	0.01	.15	0.05	0.25
			Verbal			
CSRPRAG	773	0.33	0.25	.05	–0.02	0.12
PRAGG***	971	2.39	0.50	.15	0.09	0.22
TRF***	905	0.03	0.01	.12	0.06	0.19
TPPRAGG*	416	0.03	0.01	.12	0.03	0.22
			Initiation			
CSRPRAG	773	–0.04	0.23	–.01	–0.08	0.06
PRAGG***	971	2.59	0.48	.17	0.11	0.24
TRF***	905	0.04	0.01	.19	0.12	0.25
TPPRAGG**	416	0.03	0.01	.15	0.05	0.24
			Retaliation			
CSRPRAG	773	0.39	0.28	.05	–0.02	0.12
PRAGG	971	0.31	0.34	.03	–0.03	0.09
TRF	905	0.00	0.00	.00	–0.06	0.07
TPPRAGG*	416	0.02	0.01	.12	0.03	0.22
			Malicious			
CSRPRAG	773	0.05	0.15	.01	–0.06	0.08
PRAGG	971	0.28	0.31	.03	–0.03	0.09
TRF	905	0.01	0.00	.04	–0.02	0.11
TPPRAGG	416	0.00	0.00	.02	–0.07	0.12
			Disruptive			
CSRPRAG	773	0.11	0.21	.02	–0.05	0.09
PRAGG***	971	2.05	0.44	.15	0.09	0.21
TRF***	905	0.03	0.01	.14	0.07	0.20
TPPRAGG**	416	0.03	0.01	.16	0.06	0.25

NOTE: CSRPRAG = Children's Self-Report of Peer Nominated Aggression; PRAGG = Peer Nominated Aggression; TRF = Teacher's Report Form of the Child Behavior Checklist; TPPRAGG = Teacher Predictions of Peer Nominations.
*$p < .05$. **$p < .01$. ***$p < .001$, by Wald chi-square tests.

teacher reports on the peer nomination items and only a small sample had both teacher predictions of peer nominations and TRFs. A sample of 413 cases had self-reports, peer-nominations, and TRFs, allowing us to test the associations between these aggression scales and observed behavior. Analyses using the negative binomial distribution found significant effects for the TRF Aggression scale on verbal aggression, initiation, and disruptive behavior and no effects for either the peer nominations or self-reports on any of the observed behaviors. Effect sizes for the peer nominations on verbal, initiated, and disruptive behavior were smaller than those obtained in the main analysis.[1]

DISCUSSION

This study evaluated the association between teacher, peer, and self-report measures of aggressive behavior and

observations taken in structured and unstructured times at school. We found effects for teacher ratings on most categories of observed behaviors and effects for peer nominations on all categories with the exception of retaliation and malicious behavior. Surprisingly, we found no effects for self-reports on any category of observed behavior. The effects of different teacher ratings varied, as did the effects of different informants completing the same items.

The effects of teacher ratings on observed behavior varied according to whether teachers were completing the PRAGG items or the TRF items. This difference was found on physical aggression and retaliation, for which the 95% confidence intervals for the TRF Aggression scale included zero. This difference may have been due to the somewhat greater bandwidth of the TRF Aggression scale compared to the PRAGG scale. The TRF items included disruption of class discipline and demanding attention, as well as items reflecting aggressive behavior, whereas the PRAGG items focused more narrowly on aggression, including physical, verbal, relational, disruptive, and malicious behavior.

Effects also varied within informants completing the same items. Although the PRAGG items had effects on most categories, it is noteworthy that PRAGG items, completed by peers, did not significantly predict physical or retaliatory behavior, whereas TPPRAGG, completed by teachers, significantly predicted both. It is possible that students, unlike teachers, perceive retaliation as justified, and not aggressive. Henry, Cartland, Ruch-Ross, and Monahan (2004) found, in a study of children's norms for aggression among rural, urban, and suburban grade school and high school youth, that aggression in retaliation did not fall on the same dimension as initiated aggression and was generally more acceptable among students. The difference between students and teachers in the association between the PRAGG items and physical aggression may reflect teachers' greater concern with class disruption, coupled with students' tendency to view some aggressive acts as playful rather than malicious (Pellegrini, 2003).

Self-reports on CSRPRAG, using the same items as PRAGG and TPPRAGG, did not have significant effects on any variable, even though self-reports correlated modestly with peer nominations and scores on the TRF Aggression scale. Examination of the 95% confidence interval shows that the effect of self-reports on observations of verbal behavior narrowly missed statistical significance. The validity of self-reports for measuring externalizing behaviors has been called into question (Osterman et al., 1994), and this contention may be supported by these findings. It is also possible that self-reports are tapping aggressive behavior in situations other than school. Because of teacher supervision and norms against aggression (Henry et al., 2000, 2004), aggression may be much more likely in the community and at home than in school. As children age and spend more unsupervised time with friends, self-reports may be more sensitive to aggressive behavior that occurs outside of school than among younger children. This may be related to the marginal age-moderated effects of self-reports on verbal and disruptive behavior found in this investigation. Nevertheless, determining whether cross-situational inconsistency is responsible for differences in validity would require observations in multiple contexts.

Taken together, these results suggest that teacher ratings, even on different aggression scales, have validity for measuring many behaviors related to aggression in the school context. Teacher ratings are less costly and often more practical to administer than peer nominations. These characteristics are important as prevention programs that require identification of children at risk for aggression are implemented widely (cf. Henry, Miller-Johnson, Simon, Schoeny, & The Multisite Violence Prevention Project, in press). The importance of context in the measurement of aggression also is suggested by these results. There is a need for more research on the criterion-related validity of peer and self-report aggression measures outside of school. Within the school context, however, these results call the validity of self-reports of aggression into question.

There are limitations to this study that should be considered when interpreting these results. The low base rates of observed behaviors related to aggression may have influenced the findings of this study. The absence of any association between the aggression measures and observed malicious behavior may have been due to the relatively low base rates of malicious behavior in each sample. Arguing against this possibility is the finding that the base rate of malicious behavior was higher in the sample for teacher predictions of peer nominations (TPPRAGG) than the base rate for retaliation, yet TPPRAGG was significantly associated with retaliation but not with malicious behavior.

A second limitation is the limited overlap between the samples used to evaluate the associations between measures of aggression and observed behavior. We believed that constructing the samples in this manner was necessary because the peer, teacher, parent, and self-report measures used in this study were used at different times in the MACS study for different purposes. All four measures were not administered to the same children at the same time. However, this decision does leave open the possibility that had identical samples been available for all measures at the same point in time, the results would have differed. A test with a small sample having three of the measures found smaller effect sizes for peer nominations on verbal, initiated, and disruptive behaviors than were found in the main analysis, but no substantial

differences in the effects of TRFs and self-reports. Finally, because this study did not address all subtypes of aggression, there is a need for future research exploring the criterion validity of aggression measures for predicting types of aggression that were not observed in this study.

Despite these limitations, this study has three implications for measurement of aggression. The first is to recommend the use of measures that are narrowly and specifically focused on aggressive behavior as opposed to more general measures. Although measures that tap inattention and generally disruptive behaviors are commonly used as measures of aggression, the more narrowly focused peer nomination items appeared, in this investigation, to be more closely associated with observed physical, retaliatory, and disruptive behavior than were the broader TRF items. Second, the results of this study are consistent with previous research that has called into question the validity of self-reports for predicting externalizing behaviors. The failure of self-reports to predict observed behaviors and the marginal moderated effects of age and self-reports may have been due to self-reports tapping behaviors outside of the school setting, but if so, peer nominations also would have been expected to show a similar deficit because peers are likely to be aware of behavior outside of school. A third implication is the concern that none of the measures or sources tested in this investigation significantly predicted behavior that was clearly malicious or intended to harm another. As is noted above, this may have been due to the relative rarity of such behavior in the school setting. Nevertheless, further research into the validity of aggression measures for differentiating behavior intended to harm from disruptive behavior seems warranted.

NOTE

1. A table containing the results of this analysis is available upon request from the first author.

REFERENCES

Achenbach, T. M. (1978). The child behavior profile: I. Boys aged 6-11. *Journal of Consulting and Clinical Psychology*, *46*, 478-488.

Achenbach, T. M. (1991). *Manual for the Teacher's Report Form and 1991 profile.* Burlington, VT: Associates in Psychiatry.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232.

Adejumo, A. O., Heumann, C., & Toutenburg, H. (2004). Modeling negative binomial as a substitute model to Poisson for raters agreement on ordinal scales with sparse data. Retrieved September 20, 2005, from the Institut für Statistik, Munich, Germany, Web site: http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper387.ps

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed). Washington, DC: Author.

Atkins, M. S., Pelham, W. E., & Licht, M. H. (1989). The differential validity of teacher ratings of inattention/overactivity and aggression. *Journal of Abnormal Child Psychology*, *17*, 423-435.

Baity, M. R., & Hilsenroth, M. J. (2002). Rorschach Aggressive Content (AgC) variable: A study of criterion validity. *Journal of Personality Assessment*, *78*, 275-287.

Bandura, A., & Walters, R. H. (1963). *Social learning and personality development*. New York: Holt, Rinehart and Winston.

Biederman, J., Mick, E., & Faraone, S. V. (1998). Biased maternal reporting of child psychopathology? *Journal of the American Academy of Child & Adolescent Psychiatry*, *37*, 10-12.

Bushman, B. J., & Anderson, C. A. (2001). Is it time to pull the plug on hostile versus instrumental aggression dichotomy? *Psychological Review*, *108*, 273-279.

Buss, A. (1961). *The psychology of aggression*. New York: Wiley.

Cairns, R. B., & Cairns, B. D. (1994). *Lifelines and risks: Pathways of youth in our time.* Hertfordshire, UK: Harvester-Wheatsheaf.

Carlson, M., Marcus-Newhall, A., & Miller, N. (1989). Evidence for a general construct of aggression. *Personality & Social Psychology Bulletin*, *15*, 377-389.

Casat, C. D., Norton, H. J., & Boyle-Whitesel, M. (1999). Identification of elementary school children at risk for disruptive behavioral disturbance: Validation of a combined screening method. *Journal of the American Academy of Child & Adolescent Psychiatry*, *38*, 1246-1253.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Coie, J. D., & Dodge, K. A. (1998). Aggression and antisocial behavior. In W. Damon & N. Eisenberg (Eds.), *Handbook of child psychology: Social, emotional, and personality development* (5th ed., Vol. 3, pp. 779-862). Englewood Cliffs, NJ: John Wiley.

Conduct Problems Prevention Research Group. (1999). Initial impact of the fast track prevention trial for conduct problems: I. The high-risk sample. *Journal of Consulting & Clinical Psychology*, *67*, 631-647.

Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development*, *66*, 710-722.

Dodge, K. A., & Coie, J. D. (1987). Social-information-processing factors in reactive and proactive aggression in children's peer groups. *Journal of Personality and Social Psychology*, *53*, 1146–1158.

Dunbar, J. L. (1999). Differential item performance by gender on the externalizing scales of the Behavior Assessment System for Children. *Dissertation Abstracts International Section A: Humanities & Social Sciences*, *60*(6-A), 1902.

Epkins, C. C. (1995a). Peer ratings of internalizing and externalizing problems in inpatient and elementary school children: Correspondence with parallel child self-report and teacher ratings. *Journal of Emotional & Behavioral Disorders*, *3*, 203-213.

Epkins, C. C. (1995b). Teachers' ratings of inpatient children's depression, anxiety, and aggression: A preliminary comparison between inpatient-facility and community-based teachers' ratings and their correspondence with children's self-reports. *Journal of Clinical Child Psychology*, *24*, 63-70.

Epkins, C. C. (1996). Parent ratings of children's depression, anxiety, and aggression: A cross-sample analysis of agreement and differences with child and teacher ratings. *Journal of Clinical Psychology*, *52*, 599-608.

Epkins, C. C., & Dedmon, A. M. (1999). An initial look at sibling reports on children's behavior: Comparisons with children's self-reports and relations with siblings' self-reports and sibling relationships. *Journal of Abnormal Child Psychology*, *27*, 371-381.

Eron, L. D., Laulicht, J. H., Walder, L. O., Farber, I. E., & Spiegel, J. P. (1961). Application of role and learning theories to the study of the development of aggression in children. *Psychological Reports*, *9*, 291-334.

Eron, L. D., Walder, L. O., & Lefkowitz, M. M. (1971). *The learning of aggression in children.* Boston: Little Brown.

Faraone, S., & Tsuang, M. (1994). Measuring diagnostic accuracy in the absence of a "gold standard." *American Journal of Psychiatry*, *151*, 650-657.

Flanagan, D. P., Alfonso, V. C., Primavera, L. H., Povall, L., & Higgins, D. (1996). Convergent validity of the BASC and SSRS: Implications for social skills assessment. *Psychology in the Schools*, *33*, 13-23.

Gadow, K. D., Sprafkin, J., & Grayson, P. (1990). The Helpurt game as a measure of aggression in children with learning and behavior disorders. *Learning & Individual Differences*, *2*, 337-351.

Garcia-Leon, A., Reyes, G. A., Vila, J., Perez, N., Robles, H., & Ramos, M. M. (2002). The Aggression Questionnaire: A validation study in student samples. *Spanish Journal of Psychology*, *5*, 45-53.

Guerra, N. G., Huesmann, L. R., Tolan, P. H., VanAcker, R., & Eron, L. D. (1995). Stressful events and individual beliefs as correlates of economic disadvantage and aggression among urban children. *Journal of Consulting and Clinical Psychology*, *63*, 518-528.

Henry, D., Cartland, J., Ruch-Ross, H., & Monahan, K. (2004). A return potential model of setting norms for aggression. *American Journal of Community Psychology*, *33*, 131-149.

Henry, D., Guerra, N. G., Huesmann, L. R., Tolan, P. H., VanAcker, R., & Eron, L. D. (2000) Normative influences on aggression in urban elementary school classrooms. *American Journal of Community Psychology*, *28*, 59-81.

Henry, D. B., Miller-Johnson, S., Simon, T., Schoeny, M., & The Multisite Violence Prevention Project. (in press). Validity of teacher nominations and ratings for selecting influential high risk students for a targeted intervention. *Prevention Science*.

Herzberg, P. Y. (2004). Validity of the German Questionnaire on Aggressive Traffic Behavior (AViS). *Zeitschrift fur Differentielle und Diagnostische Psychologie*, *25*, 153-164.

Hilton, N. Z., Harris, G. T., & Rice, M. E. (2003). Correspondence between self-report measures of interpersonal aggression. *Journal of Interpersonal Violence*, *18*, 223-239.

Huesmann, L. R., & Eron, L. D. (1986). *Television and the aggressive child*. Hillsdale, NJ: Lawrence Erlbaum.

Huesmann, L. R., Eron, L. D., Guerra, N. G., & Crawshaw, V. B. (1994). Measuring children's aggression with teachers' predictions of peer nominations. *Psychological Assessment*, *6*, 329-336.

Huesmann, L. R., Eron, L. D., Lefkowitz, M. M., & Walder, L. O. (1984). Stability of aggression over time and generations. *Developmental Psychology*, *20*, 1120-1134.

Huesmann, L. R., Lefkowitz, M. M., & Eron, L. D. (1978). Sum of MMPI Scales F, 4, and 9 as a measure of aggression. *Journal of Consulting & Clinical Psychology*, *46*, 1071-1078.

Huesmann, L. R., Maxwell, C., Eron, L., Dahlberg, L. L., Guerra, N. G., Tolan, P. H., et al. (1996). Evaluating a cognitive/ecological program for the prevention of aggression in urban children. *Preventive Medicine*, *12*, 120-128.

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, *10*, 93-98.

Lefkowitz, M. M., Eron, L. D., Walder, L. O., & Huesmann, L. R. (1977). *Growing up to be violent: A longitudinal study of the development of aggression*. New York: Pergamon.

MacLean, W. E., Tapp, J. T., & Johnson, W. L. (1985). Alternative methods and software for calculating interobserver agreement for continuous observation data. *Journal of Psychopathology and Behavioral Assessment*, *7*, 65-73.

McIntyre, A. (1972). Sex differences in children's aggression. *Proceedings of the Annual Convention of the American Psychological Association*, *7*, 93-94.

Metropolitan Area Child Study Research Group. (2002). A cognitive-ecological approach to preventing aggression in urban settings: Initial outcomes for high risk children. *Journal of Consulting and Clinical Psychology*, 70, 179-194.

Minturn, L. (1967). The dimensions of aggression: A descriptive scaling study of the characteristics of aggressive pictures. *Journal of Experimental Research in Personality*, *2*, 86-99.

Nugent, W. R. (1994). A differential validity study of the Self-Esteem Rating Scale. *Journal of Social Service Research*, *19*, 71-86.

O'Connor, D. B., Archer, J., & Wu, F. (2001). Measuring aggression: Self-reports, partner reports, and responses to provoking scenarios. *Aggressive Behavior*, *27*, 79-101.

Orpinas, P., & Frankowski, R. (2001). The Aggression Scale: A self-report measure of aggressive behavior for young adolescents. *Journal of Early Adolescence*, *21*, 50-67.

Osterman, K., Bjorkqvist, K., Lagerspetz, K. M. J., & Kaukiainen, A. (1994). Peer and self-estimated aggression and victimization in 8-year-old children from five ethnic groups. *Aggressive Behavior*, *20*, 411-428.

Palme, E. J., & Thakordas, V. (2005). Relationship between bullying and scores on the Buss-Perry Aggression Questionnaire among imprisoned male offenders. *Aggressive Behavior*, *31*, 56-66.

Pellegrini, A. D. (2003). Perceptions and functions of play and real fighting in early adolescence. *Child Development*, *74*, 1522-1533.

Pepler, D. J., Craig, W. M., & Roberts, W. L. (1998). Observations of aggressive and nonaggressive children on the school playground. *Merrill-Palmer Quarterly*, *44*, 55-67.

Repp, A. C., Karsh, K. G., VanAcker, R., Felce, D., & Harman, M. (1989). A computer-based system for collecting and analyzing observational data. *Journal of Special Education Technology*, *9*, 207-21.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children: Manual.* Circle Pines, MN: American Guidance Service.

Shields, J., Konold, T. R., & Glutting, J. J. (2004). Validity of the Wide Range Intelligence Test: Differential effects across race/ethnicity, gender, and educational level. *Journal of Psychoeducational Assessment*, *22*, 287-303.

VanAcker, R., Bush, J., Grant, S. H., & Getty, J. E. (1992). *A software package for the intermittent time sampling of behavior.* Chicago: Stoelting.

Wang, E. W., Rogers, R., Giles, C. L., Diamond, P. M., Herrington-Wang, L. E., & Taylor, E. R. (1997). A pilot study of the Personality Assessment Inventory (PAI) in corrections: Assessment of malingering, suicide risk, and aggression in male inmates. *Behavioral Sciences and the Law*, *15*, 469-482.

Waschbusch, D. A., Willoughby, M. T., & Pelham, W. E., Jr. (1998). Criterion validity and the utility of reactive and proactive aggression: Comparisons to attention deficit hyperactivity disorder, oppositional defiant disorder, conduct disorder, and other measures of functioning. *Journal of Clinical Child Psychology*, *27*, 396-405.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-059). Beverly Hills, CA: Sage.

David B. Henry is associate professor of psychology at the Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago. With his colleagues in the Families and Communities Research Group, he has conducted prevention trials and developmental risk studies, including the Metropolitan Area Child Study, the SAFE Children Project, the CDC Multisite Violence Prevention Project, and the Chicago Youth Development Study. He is author or coauthor of more than 60 published works, including articles on prevention, risk, development, statistical methods, peer relations, and attitudes.