Teacher, parent, and peer reports of early aggression as screening measures for long-term maladaptive outcomes: Who provides the most useful information?

Katherine H. Clemans, Ph.D.,
Department of Psychology, Amherst College

Rashelle J. Musci, Ph.D,
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

Jeannie-Marie S. Leoutsakos, Ph.D, and
Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine

Nicholas S. Ialongo, Ph.D
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

Abstract

Objective—This study compared the ability of teacher, parent, and peer reports of aggressive behavior in early childhood to accurately classify cases of maladaptive outcomes in late adolescence and early adulthood.

Method—Weighted kappa analyses determined optimal cut points and relative classification accuracy among teacher, parent, and peer reports of aggression assessed for 691 students (54% male; 84% African American, 13% White) in the fall of first grade. Outcomes included antisocial personality, substance use, incarceration history, risky sexual behavior, and failure to graduate from high school on time.

Results—Peer reports were the most accurate classifier of all outcomes in the full sample. For most outcomes, the addition of teacher or parent reports did not improve overall classification accuracy once peer reports were accounted for. Additional gender-specific and adjusted kappa analyses supported the superior classification utility of the peer report measure.

Conclusion—The results suggest that peer reports provided the most useful classification information of the three aggression measures. Implications for targeted intervention efforts which use screening measures to identify at-risk children are discussed.

Keywords

aggression; screening; peers; teachers; parents

Aggressive behavior during childhood is a known risk factor for a wide range of later maladaptive outcomes. Over the past several decades, research has documented links

Correspondence concerning this article should be addressed to Katherine H. Clemans, Department of Psychology, Amherst College, Amherst, MA 01002. kclemans@amherst.edu.

between early aggression and many avenues of problematic behavior later in life, including antisocial behavior and criminal activity, school and employment struggles, risk-taking behavior, and other mental health problems (Fergusson, Horwood, & Ridder, 2005; Huesmann, Eron, & Dubow, 2002). The identification of at-risk children is of considerable interest to psychologists, researchers, school personnel, and implementers of early intervention programs, as recent instances of school-based violence in the United States and other countries have served as grim reminders of the importance of early detection and intervention for troubled youth. Intervening in the development of aggressive or antisocial behavior early in life may help deter future costs to both the individual and society that result from problematic outcomes (Kellam et al., 2008).

An ongoing and unresolved question in the field concerns the best type of informant from whom to collect information about children's behavior. Multiple sources are potentially available, including teachers, parents, peers, trained observers, and the children themselves. Several researchers have suggested that information should be collected from multiple informants whenever possible, given that behavior can be situation-specific and different raters may observe behavior at different times and in different contexts (Achenbach, McConaughy, & Howell, 1987; Renk, 2005). Collecting data from more than one source may require additional time and money, however, and for many programs, limited resources may preclude the use of more than one or two informants. Determining which informant types provide the most useful information for the identification of children at risk for later maladaptive outcomes is an area of research that has important implications for the implementation of targeted early intervention efforts.

Commonly tapped sources of information about aggressive behavior in early childhood are teachers, parents, and, to a lesser extent, children's peers. Self reports of aggression are rarely collected in this age range, as it is thought that young children's limited ability to reflect on their own behavior may affect reliability (Renk, 2005). The use of teacher and/or parent data is particularly common for intervention-related screening purposes (e.g., Lochman & The Conduct Problems Prevention Research Group, 1995; Petras et al., 2004). Loeber, Green, and Mahey (1990) surveyed mental health professionals and found that they believed parents to be better sources of information about children's conduct problems than teachers. Teacher reports, however, are often easier and less expensive for researchers to obtain than parent reports, in part because fewer adult raters need to be recruited and compensated. Peer reports of behavior are sometimes cited as the most valid standard for information about aggression in older children and adolescents (Peets & Kikas, 2006) because much aggressive behavior in this age range is covert and takes place away from the observation of adults. Peers have been less frequently utilized as an informant source in younger age groups due to the limited reading abilities of young children and other logistic difficulties; if using peers for this age group, the use of student photographs for survey materials may be necessary (Petras, Buckley, Leoutsakos, & Ialongo, 2011).

Parent reports, teacher reports, and peer reports have all been used to measure and identify childhood aggression as a risk factor for long-term maladaptive outcomes in late adolescence and early adulthood. Timmermans, van Lier, and Koot (2008), for instance, found that parent-rated aggression at ages 4 and 5 was a unique predictor of late adolescent

drug use. Similarly, Brook and Newcomb (1995) found that parent-rated aggression at ages 5 to 10 was positively associated with drug use in adolescence and early adulthood and negatively associated with academic achievement, including the likelihood of graduating high school. Teacher ratings of aggressive and disruptive behavior assessed in the fall of first grade have been shown to be a significant predictor of late adolescent and early adulthood antisocial personality disorder, substance abuse and dependence, and history of incarceration (Kellam et al., 2008; Petras et al., 2004). Peer nominations of aggressive 8-year-olds have significantly predicted risky sexual behavior in late adolescence (Serbin, Peters, McAffer, & Schwartzman, 1994) as well as criminal justice convictions and the seriousness of criminal acts at age 30 (Huessman, Eron, Lefkowitz, & Walder, 1984). In addition, several studies investigating early childhood predictors of late adolescent and early adulthood maladaptive outcomes have combined aggressive behavior ratings from two or more informant types into a single measure. Fergusson et al. (2005) found that 7- to 9-year-olds' composite scores of parent- and teacher-rated aggression were significantly associated with higher levels of self-reported criminal offending, imprisonment, antisocial personality symptoms, illicit drug use, and risky sexual behavior and lower levels of educational achievement in early adulthood.

Questions remain, however, as to whether particular informants are more useful for some outcomes over others. In Timmermans et al.'s (2008) study, for example, parent-rated aggression was predictive of late adolescent substance use, but not of risky sexual behavior. Reinherz, Giaconia, Carmola, Wasserman, and Paradis (2000) found that teacher-rated aggression at age 9 predicted drug abuse/dependence at age 21, but parent-rated aggression did not. Similarly, parent- and teacher-rated aggression composite scores ceased to be predictive of educational achievement in the study by Fergusson et al. (2005) once other variables were taken into account, but remained a unique predictor of mental health problems, criminal activity, and risky sexual behavior. There is also evidence that the utility of some informant reports may differ by the gender of the child. Ensminger, Juon, & Fothergill (2002), for example, have shown that teacher-rated aggressive behavior in first grade is associated with drug use at age 32 for males but not for females; other ratings of aggression have had stronger relationships with later criminal behavior in males but not females, due in part to females' lower engagement in adult criminal activity (Schaeffer et al., 2006).

The present study builds on this previous literature by investigating the comparative ability of parent, teacher, and peer reports to classify cases of maladaptive outcomes in later adolescence and early adulthood, including antisocial personality disorder, substance abuse, risky sexual behavior, history of incarceration, and delayed high school graduation. As described above, much of the previous research in this field has used population-level regression analyses to investigate how parent, teacher, and peer ratings of aggressive behavior predict late adolescent and early adulthood outcomes. The present study, however, treats parent, teacher, and peer reports as screening measures, which allows for the classification of individuals into high and low risk groups for each outcome. When treated as a screening measure, continuous ratings of early aggressive behavior are cut on a predetermined point into high risk and low risk classification, with the expectation that individuals in the high risk group will be more likely than those in the low risk group to

develop the problematic outcome. Effective screening procedures maximize sensitivity, or the probability that the measure will correctly identify cases that eventually go on to develop the outcome (true positives), and specificity, or the probability that the measure will correctly identify cases that do not go on to develop the outcome (true negatives). Different cut points on each screening measure will have varying sensitivity and specificity, with the most *efficient* cut point defined as that which maximizes the number of correctly classified cases across both risk groups.

Importantly, targeted intervention programs have used ratings of aggressive behavior in early elementary school to screen and identify children at risk for later conduct problems and antisocial behavior (e.g., Lochman et al., 1995). Particularly when the outcome in question has low population prevalence, targeted interventions that use screening measures to identify subsamples with elevated risk may be a more effective utilization of resources than universal interventions, which provide intervention services regardless of risk (Mrazek & Haggerty, 1994). The ability of a screening measure to accurately predict future cases of an outcome can affect the efficiency and cost-effectiveness of targeted intervention efforts (Hill, Lochman, Coie, Greenberg, & The Conduct Problems Prevention Research Group, 2004; Lochman et al., 1995; Salkever et al., 2008).

Several studies conducted by The Conduct Problems Prevention Research Group (Hill et al., 2004; Lochman et al., 1995) have addressed the question of whether screening measures based on ratings from particular informants are *uniquely* useful; in other words, whether they will provide additional information about outcome probability once other informants' ratings are taken into account. Using data from the Fast Track intervention program, in which children were rated by both teachers and parents on aggressive, disruptive, and oppositional behavior in first grade, the studies found that the use of both teacher and parent reports improved classification accuracy over the use of teacher reports alone for a range of externalizing outcomes from 30 months to 4 years later. The authors recommended that targeted intervention efforts could conserve resources by identifying at-risk youth using a multiple gating approach, in which a universal, low-cost teacher screen is followed up by parent screening only for those identified as high risk by teachers (Lochman et al., 1995).

Thus, comparing the ability of informant reports to correctly predict future cases of maladaptive outcomes offers an alternative approach to basic population-level regression analyses for determining whether particular raters of early childhood aggressive behavior provide more or less useful information than other informants. Furthermore, the results of this approach may help to inform resource allocation for data collection in future research efforts and preventive intervention trials.

Present study

The primary goal of this research was to compare the relative utility of aggression ratings from three different informant sources as screening measures for the same set of distal outcomes. We used the Kappa Tree statistical program (Leoutsakos, 2007), which allowed us to evaluate weighted kappa coefficients from all available informant ratings to determine (1) the most efficient cut points on teacher, parent, and peer ratings for predicting cases of

the distal outcomes, (2) which informant report (teacher, parent, or peer) provided the best classification utility for each distal outcome; and (3) whether additional informant reports provided added improvement to the overall classification utility for each outcome once information from the optimal informant was taken into account. Further information on the Kappa Tree program and relevant analyses is provided in the Analytic Plan (below).

Because evidence has shown that boys and girls have varying levels of risk for maladaptive outcomes, we also investigated whether patterns of optimal informant reports and cut points would vary by gender. In addition, we performed a sensitivity analysis to determine whether patterns of optimal informant reports and cut points would change as a function of the emphasis placed on the screening measures' ability to correctly predict true cases of the outcome.

The present study offers several contributions to the current literature. The use of the KappaTree program allowed the identification of optimal cut points on each screening measure, avoiding the need to impose arbitrary a priori cutoffs on continuous ratings. This approach was employed by Hill et al. (2004) to investigate the screening utility of teacher and parent reports in first grade for externalizing behavior 3 to 4 years later, but to our knowledge has not thus far been used to investigate longer-term outcomes in late adolescence and early adulthood. Hill and colleagues found that utilizing information from both parents and teachers improved classification accuracy over the use of teacher ratings alone. In this study, we sought to extend Hill and colleagues' findings by evaluating peers as informants along with teachers and parents as well as investigating the classification utility of these informants for a range of long-term maladaptive outcomes in late adolescence and early adulthood.

Methods

Participants and Procedure

Data were taken from a longitudinal preventive intervention trial conducted by the Johns Hopkins Prevention Intervention Research Center. All procedures had IRB approval. At the initial assessment in 1993, 799 first graders from 27 classrooms in 9 elementary schools in Baltimore were randomized to a control group or to one of two universal preventive interventions, the goals of which were to improve behavior and school performance proximally and reduce later drug use, antisocial behavior, and school failure distally. A complete description of the intervention trial procedures is available in Ialongo et al. (1999). In this study, participants in both the intervention and control groups of the trial were utilized in order to maximize the number of positive cases for each of the late adolescent/ early adult outcomes. The parent, teacher, and peer ratings of aggressive behavior in the present study were baseline assessments taken in the fall of first grade before the intervention trial began, and thus were not affected by intervention status. Participants were contacted for follow-up interviews each year between 2006 and 2009, when the participants were age 19 to 23. The initial sample was 54.4% male, 84.9% African American, 13% White, and 1% Asian and Hispanic American, and the average age at the first assessment in the fall of first grade was 6.2 years (SD ±0.37). The portion of the sample receiving free or reduced lunch in first grade was 62.3%. Within all eligible classrooms in the intervention

trial, 691 students had parental consent to participate in data collection and received aggressive behavior ratings from at least one informant source in first grade. Of this initial sample, teacher ratings of aggression were available for 678 (98%); peer nominations of aggression were available for 594 (86%), and parent ratings of aggression were available for 593 (86%).

Follow-up assessments in late adolescence/early adulthood were completed via a combination of in-person and telephone interviews. For sensitive questions, such as drug use and sexual behavior, computer-assisted telephone interviewing was used. A tracking database with parent information and alternate contacts was maintained and updated after each interview wave, and participants received monetary compensation for each interview. LexisNexis, a computer-assisted legal research service, was used to locate participants who had moved or changed contact information. Data was also collected from participants' high school records and from the Maryland Criminal Justice Information System. From these assessments, late adolescent/early adulthood follow-up data was obtained for 641 (93%) on antisocial personality disorder (ASPD) diagnostic criteria, substance dependence diagnostic criteria, and risky sexual behavior; for 635 (91%) on high school graduation, and for 687 (99%) on incarceration history.

Missing data and attrition—Missing data comparisons on the first grade aggression ratings revealed no significant differences between participants with missing and non-missing aggression ratings from any rater on gender, intervention status, ASPD, drug dependence, or incarceration history. Participants with missing peer nominations of aggression were somewhat less likely to graduate from high school on time ($B = -.51$, $p < .05$) and to engage in risky sexual behavior ($B = -.67$, $p < .01$). Attrition analyses revealed that participants with missing data on ASPD, substance dependence, and risky sexual behavior had slightly lower levels of teacher-rated aggression in first grade, $t(67.66) = -2.56$, $p < .05$. No significant differences were found for peer-rated aggression, parent-rated aggression, gender, or intervention status between participants with present and missing outcome data.

Measures

Aggression ratings: Teacher—Teacher ratings of aggressive behavior in the fall of 1st grade were obtained using items from the Teacher Observation of Classroom Adaptation-Revised (TOCA-R; Werthhamer-Larsson, Kellam, & Wheeler, 1991). The Accepting Authority subscale of this measure consisted of 14 items which asked teachers about students' aggressive, delinquent, and disruptive behavior. This scale, along with other items in the TOCA-R, has been successfully used in past research as a screening test to identify young children at risk of later externalizing problems for targeted interventions (Hill et al., 2004; Lochman et al., 1995). For this study, we used a subset of 7 items which specifically assessed aggressive behavior ($\alpha = .92$) so that the wording of items in the teacher, parent, and peer scales would be as similar as possible; analyses run with the full subscale did not significantly change the pattern of results. Teachers rated each child's behavior over the previous three weeks on a frequency scale from 1 to 6 (1 = not at all, 6 = always); items were combined to produce an average score. Items included "fights," "yells at others,"

"loses temper," "starts fights with classmates," "harms others," "teases classmates," and "breaks rules."

Aggression ratings: Parent—Parent ratings of aggressive behavior were obtained from the POCA (Parent Observation of Classroom Adaptation; Ialongo, Werthamer, & Kellam, 1999). Items on the POCA mirrored those of the TOCA-R. A subset of 7 items from the accepting authority subscale that specifically assessed aggressive behavior was used in the current analyses (α = .73). Parents rated each child's behavior over the previous three weeks on a frequency scale from 1 to 4 (1 = not at all, 4 = always); items were combined to produce an average score. Instructions did not limit parents to report only on home behavior, so scores could reflect parents' knowledge of childrens' behavior both at home and at school. Items included "fights," "yells at others," "loses temper," "starts physical fights," "hurts others physically," "teases other children," and "breaks rules."

Aggression ratings: Peer—Peer nominations of aggressive behavior were taken from the Peer Assessment Inventory (PAI; Dolan et al., 1993; Pekarik, Prinz, Leibert, Weintraub, & Neale, 1976). Students completed in-class survey sessions in which a researcher read each item description aloud and students nominated classmates by circling photos of classmates on a sheet corresponding to that item. Students were not limited in their number of nominations. The survey contained 3 items which assessed aggressive behavior (e.g., "Which children start fights?"; "Which children are bullies?" "Which children get into trouble a lot?"; α = .89). Raw nomination totals for the aggressive behavior items were summed and then divided by the number of participating nominators within a child's classroom.

Antisocial personality disorder—ASPD diagnosis was assessed using items from the Diagnostic Interview Schedule (Robins et al., 2000). This measure is consistent with criteria set forth by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV, American Psychiatric Association, 1994). At each assessment, participants responded to 23 items representing six categories of symptom criteria, which included irritability/aggressiveness, reckless disregard for the safety of self or others, failure to conform to social norms, deceitfulness, impulsivity, and consistent irresponsibility. ASPD was coded as present if the individual participated in at least one follow-up interview and endorsed items in three or more symptom categories at any time point between age 19 and 23.

Drug dependence—Drug dependence diagnosis was assessed using DSM-IV criteria from the National Survey on Drug Use & Health (SAMHSA, 2001). This measure is consistent with criteria set forth by the DSM-IV. Participants responded to items assessing seven categories of symptom criteria for a range of illicit drugs. Drug dependence was coded as present if the individual participated in at least one follow-up interview and endorsed items in three or more symptom categories at any time point between age 19 and 23.

Risky sexual behavior—Participants reported on their frequency of engaging in unprotected sexual activity. Risky sexual behavior was coded as present if the individual

participated in at least one follow-up interview and indicated that they had engaged in this behavior in the past year.

Incarceration history—Participants' records were gathered from the Maryland Criminal Justice Information System in 2009 and reflected adult incarceration history in the state. This variable was coded as 1 = ever incarcerated and 0 = never incarcerated.

Late or no high school graduation—Data on the timing of high school graduation was available for the majority of individuals from school records and teacher reports obtained from participants' high schools. In the cases in which school data was not available, the timing of graduation was coded from participants' self reports of educational history, which were assessed at each follow-up interview. This variable was coded as 0 if the participant had graduated from high school or started college or vocational school by the first follow-up assessment in 2006 and 1 if the participant never graduated or graduated/obtained a GED after 2006.

Analytic plan

Analyses were completed using KappaTree (Leoutsakos, 2007), an R adaptation of the ROC4 program (Yesavage, 2008). The KappaTree program cuts continuous test variables on a range of different points and calculates kappa (κ) coefficients for each point, which are used to determine the optimal cut points for each test and optimal overall test. These κ coefficients can be assigned a weight from 0 (maximizing specificity) to 1 (maximizing sensitivity); a κ with a weight of .5 maximizes overall classification efficiency and is equivalent to Cohen's κ. The use of the weighted κ statistic has been recommended for analyses of screening measure utility (Kraemer et al., 1999), as it has the flexibility to emphasize sensitivity or specificity based on the considerations and goals of the intervention or research program in question.

A test variable with a significant 2×2 χ^2 table according to its kappa-determined optimal cut point is considered a legitimate classifier of the outcome. After the initial cut, the program runs again within the high risk and low risk groups and determines optimal cuts within each risk group on remaining test variables, creating a branching series of results (Yesavage, 2008). Additional branches are created only when there is a remaining legitimate classifier variable within that group. In other words, additional branches provide information on whether remaining classifiers improve outcome classification once the sample has been cut on the optimal classifier. For each outcome, the program calculates the "ragged" κ, which is a weighted κ calculated across the terminal points of the final branching pattern (see Petras et al., 2011), gives the overall sensitivity, specificity and efficiency when information from all branches is used, and provides confidence intervals for differences between κs at each branching level and the ragged κ.

The KappaTree adaptation of the original ROC4 program gives the user enhanced control over the specifications of the program, including the ordering and usage of test variables, the κ weight, and the χ^2 α criterion. For these analyses, we specified the following: The use of test variables to create a maximum of one branch (i.e., once the program cut the sample on a particular test variable, that variable was not evaluated again for cuts on subsequent

branches); a κ weight of .5; an α criterion less than or equal to .05 for chi-squared tests of the legitimacy of test variables; and calculation of nonparametric bootstrapped confidence intervals for κs and differences between κ values due to the positive skewness of aggression ratings. We constrained the program to evaluate cut points at increments of .05. To aid in the comparison and interpretation of cut points, variables were transformed into z-scores for kappa tree analyses; thus, cut points were evaluated beginning at the lowest z-score value for each measure and then incrementally adding .05 SD for each evaluation until the maximum value was reached. To evaluate whether a ragged κ value represented a significant improvement over the κ value for the first cut, we looked at 95% bootstrapped confidence intervals for the difference between the first-cut and ragged κs.

Kappa tree analyses for each of the distal outcomes were run separately using teacher, parent, and peer reports of aggressive behavior in first grade as predictor variables. The KappaTree program starts with the *N* for the outcome variable and then runs individual analyses for each predictor, resulting in the use of all available data for each analysis. The order of analysis was as follows: First, we investigated results for the full sample. We then split the sample by gender and re-ran each analysis to determine whether branching patterns varied for boys and girls. Finally, we manipulated the kappa weight in a sensitivity analysis to investigate whether changes in the pattern of results occurred when greater emphasis was placed on identifying true positives and avoiding false negatives. Petras et al. (2011) suggested that a kappa weight of .5 is most applicable to targeted interventions under real-world conditions, in which the intervention is both relatively costly (a reason to avoid false positives) and shows only moderate effects on the desired outcome (a reason to avoid false negatives; Kraemer et al., 1999). However, in cases in which a targeted intervention screening process is relatively inexpensive and the outcome has a low population base rate, it may be better to employ a screening approach that emphasizes the minimization of false negatives, thus ensuring that a greater proportion of individuals who are truly at risk will receive the intervention. The appendix presents initial test statistics and a description of branching patterns when the kappa weight is specified at .75, which places increased emphasis on avoiding false negatives and reduces emphasis on avoiding false positives.

Results

Preliminary analyses

Descriptive statistics—Sample means, standard deviations, correlations, and measures of skewness were calculated for each reporter of aggressive behavior (see table 1). Peer and teacher reports of aggressive behavior were highly correlated with one another, whereas parent report was weakly but significantly correlated with peer and teacher reports. Prevalence rates of the distal outcomes and valid percentages of positive outcome cases are given in Table 2.

Because the measures used to assess teacher- and parent-reported aggression (TOCA-R and POCA) were variations of the same scale, we tested the configural invariance of these measures using exploratory factor analysis with promax rotation (Cole, Hoffman, Tram, & Maxwell, 2000). A separate factor analysis was completed for each 7-item scale, and the two were compared to evaluate the similarity of factor structure and factor loadings across

informants. Both analyses produced a single factor solution (eigenvalue > 1; Kaiser, 1960), suggesting configural invariance in the structure of the scales. These factors accounted for, respectively, 67.2% of variance in the TOCA-R and 41.0% of variance in the POCA. Primary factor loadings for the 7 items ranged from .79 to .88 for the TOCA-R and from .49 to .73 for the POCA.

Intervention status—We explored the potential role of intervention status as a confounding variable for the distal outcomes measured in this study by first running all analyses with intervention status included as a test variable along with teacher, parent, and peer reports of aggression in first grade. Intervention status did not produce significant χ^2 values for any of the five outcomes, nor did the program choose to branch on intervention status in any of the analyses. Because the incorporation of intervention status did not alter the test statistics or kappa tree branching patterns of teacher, parent, and peer reports, the final analyses presented in Table 1 and Figure 1 do not include it as a test variable.

Kappa tree analyses for each outcome

We ran a kappa tree analysis for each of the distal outcomes of interest using teacher, parent, and peer reports of aggression in first grade as test variables. Table 3 presents initial test characteristic statistics for each informant, including optimal cut point, χ^2 value, κ, efficiency (eff), sensitivity (sen), specificity (spec), and positive predictive value (PPV). Figure 1 presents a graphical representation of the kappa tree branching patterns, with optimal cut points and outcome *Ns* for each branch.

Sixteen percent of the sample met self-reported diagnostic criteria for ASPD. Both peer and teacher reports of early aggression were legitimate classifiers of later ASPD. The best classifier was peer reports, which produced the largest κ (.16; 95% CI [.08, .24]) and an efficiency of .73 at a cut point of .41. This suggests that individuals with a peer reported aggression score higher than .41 SD were at higher risk for future development of ASPD. No further branching occurred after this cut.

For drug dependence, 11% of the sample met self-reported diagnostic criteria. At the initial level of analysis, only peer reports of early aggression were a legitimate classifier of later drug dependence diagnosis. The optimal cut point for peer reports was .61, which offered a κ of .12 (95% CI [.04, .20]) and an efficiency of .77. After the sample was branched according to the optimal peer cut point, teacher reports became a legitimate classifier for participants in the high risk branch only. Test characteristics for teacher reports at this level were as follows: Optimal cut point = .13; $\chi^2 = 6.90$, $p < .01$; $\kappa = .20$; eff = .69; sen = .54; spec = .73; PPV = .29. Unexpectedly, within the peer high risk branch, having scores *below* the optimal teacher cut point indicated higher risk. No further branching occurred after this cut. The overall ragged κ value across all branches was .19 (95% CI [.07, .31]) and the overall efficiency was .86. The difference between the first cut and ragged κ values was .07 (95% CI [−.03,.17]), suggesting that inclusion of information from teacher reports did not significantly improve classification of drug dependency over the use of peer reports alone.

A history of incarceration was present for 10% of the sample. Both peer and teacher reports of early aggression were legitimate classifiers of later incarceration. The best classifier was

peer reports, which produced the largest κ (.21, 95% CI [.12, .31]) and highest efficiency (. 82) at a cut point of .81. After the sample was branched according to the optimal peer cut point, teacher reports did not remain a legitimate classifier of later incarceration within either branch; thus, no further branching occurred.

Recent engagement in high risk sexual behavior was reported by 58% of the sample. At the initial level of analysis, only peer reports of early aggression were a legitimate classifier of later high risk sex. The optimal cut point for peer reports was .06, which offered a κ of .09 (95% CI [.01, .17]) and an efficiency of .57. After the sample was branched according to the optimal peer cut point, teacher reports became a legitimate classifier for participants in the high risk branch only. Test characteristics for teacher reports at this level were as follows: Optimal cut point = −.62; χ^2 = 5.49, $p < .05$; κ = .12; eff = .57; sen = .57; spec = .56; PPV = .69. One final branching on parent scores occurred within the low-risk teacher group. Test characteristics for parent reports at this level were as follows: Optimal cut point = .15; χ^2 = 5.45, $p < .05$; κ = .19; eff = .61; sen = .67; spec = .52; PPV = .66. The overall ragged κ value across all branches was .16 (95% CI [.08, .24]) and the overall efficiency was .86. The differences between the first cut and ragged κs was .07 (95% CI [.01, .13]), suggesting that improvement in classification due to the inclusion of teacher and parent reports was just significant at the $p < .05$ level.

Finally, 46% of the sample failed to graduate from high school on time. All three informant reports of early aggressive behavior were legitimate classifiers of late/no high school graduation. Peer reports were the best classifier at a cut point of .21, producing the largest κ (.22, 95% CI [.14, .28]) and highest efficiency (.63). After the sample was branched according to the optimal peer cut point, both teacher and parent reports remained legitimate classifiers of high school graduation for participants in the peer low risk branch. Of these, parent reports were determined to be the best classifier, with test characteristics as follows: Optimal cut point = .50; χ^2 = 6.84, $p < .01$; κ = .14; eff = .62; sens = .76; spec = .37; PPV = . 67. In the peer high risk branch, only teacher reports remained a legitimate classifier, with test characteristics as follows: Optimal cut point = 1.63; χ^2 = 5.44, $p < .05$; κ = .14; eff = . 52; sens = .85; spec = .32; PPV = .42. No further branching occurred. The overall ragged kappa value was .13 (95% CI [.09, .18] and the overall efficiency was .58, suggesting that inclusion of information from teacher and parent reports did not improve classification of ASPD cases over the use of peer reports alone. In fact, a comparison of κ values suggests that the inclusion of parent and peer reports at the second level significantly decreased classification utility; however, this comparison should be viewed with caution as the reduction in κ value between the first cut and the ragged κ may have been due to a 6% reduction in the sample size at the second branching level as a result of missing data on parent reports.

Investigation of gender patterns—We next ran Kappa Tree analyses separately by gender to investigate whether patterns of optimal classifiers and cut points would vary among girls and boys. For ASPD, peer reports and parent reports were legitimate classifiers for boys. The best classifier was peer reports (optimal cut point = .41; χ^2 = 8.36, $p < .01$; κ = .16; eff = .64; sens = .53; spec = .67; PPV = .30), and no further branching occurred. None of the informant reports were legitimate classifiers for girls' cases of ASPD.

For drug dependence, peers reports as well as teacher reports were legitimate classifiers for boys. Peer reports were the best classifier of the two (optimal cut point = 2.41; χ^2 = 5.92, p < .05; κ = .14; eff = .85; sens = .20; spec = .93; PPV = .25); no further branching occurred. In addition, peer reports as well as parent reports had significant χ^2 tests for girls. Of these, peer reports were the best classifier (optimal cut point = .91; χ^2 = 10.53, p < .01; κ = .20; eff = .86; sens = .26; spec = .93; PPV = .29), and no further branching occurred.

For incarceration history, none of the informant reports were legitimate classifiers for girls' cases of incarceration; however, peer reports as well as teacher reports remained significant χ^2 tests for boys. Of these, peer reports were the best classifier (optimal cut point = .81; χ^2= 12.64, p < .001; κ = .19; eff = .73; sens = .47; spec = .78; PPV = .28), and no further branching occurred.

For risky sexual behavior, the boys' analysis did not result in a legitimate classifier. For girls only, however, teacher reports had a significant χ^2 test (optimal cut point = −.32; χ^2 = 4.39, p < .05; κ = .10; eff = .51; sens = .40; spec = .72; PPV = .72), though the overall efficiency of teacher reports was poor. No further branching occurred for girls.

For late or no high school graduation, all three informants remained legitimate classifiers for girls, whereas only peer and parent reports were legitimate classifiers for boys. Peer reports were determined to be the best classifier for both genders (optimal cut point for girls = −.34; χ^2 = 8.35, p < .01; κ = .18; eff = .63; sens = .73; spec = .42; PPV = .69; optimal cut point for boys = .76; χ^2 = 19.14, p < .001; κ = .22; eff = .60; sens = .86; spec = .36; PPV = .55). Further branching at the second level occurred on teacher reports for girls and on parent reports for boys. Ragged κs were .24 [.12, .29] for girls and .27 [.16,.38] for boys. Differences between first cut and ragged κs were .06 [.02, .11] for girls and .05 [−.05,.16] for boys, suggesting that the inclusion of teacher reports significantly improved classification of graduation status for girls, whereas the inclusion of parent reports did not significantly improve classification of graduation status for boys.

Modification of kappa weight—As a sensitivity analysis, the appendix presents first cut test statistics when the kappa weight was specified at .75, which placed increased emphasis on avoiding false negatives and reduces emphasis on avoiding false positives. In these analyses, branching patterns were unchanged and peer reports remained the optimal classifier for all of the outcomes under investigation.

Discussion

This study compared the ability of teacher, parent, and peer reports of early aggressive behavior to correctly classify maladaptive outcomes in late adolescence and early adulthood, including diagnostic criteria of ASPD and drug dependence, incarceration history, risky sexual behavior, and failure to graduate from high school on time. The overall results suggested that the peer report measure of aggressive behavior used in this study (PAI) had more utility than the teacher report (TOCA-R) or parent report (POCA) measures to correctly classify distal outcomes. Peer reports were determined to be the best classifier of outcome status for all five outcomes. Teacher reports were legitimate classifiers at outset for

ASPD, drug dependence, incarceration history, and high school graduation, and after initial branching were legitimate classifiers for drug dependence, high school graduation, and risky sexual behavior for at least one peer report-determined risk group. However, teacher reports were always a less useful classifier than peer reports according to weighted κ values of optimal cut points. Overall ragged κ and efficiency values for outcomes with multiple branches suggested that, when compared to the use of peer reports alone, the incorporation of teacher reports did not significantly improve classification utility for any outcome in the full sample, although inclusion of both parent and teacher reports offered significant improvement over peer reports for risky sexual behavior. Parent reports of aggressive behavior offered the lowest classification utility of the three informant measures, significantly classifying only high school graduation at outset.

We also investigated the classification utility of peer, teacher, and parent reports when the sample was split by gender and when the kappa weight was adjusted to place more emphasis on avoiding false negatives and less emphasis on avoiding false positives. When results were investigated separately by gender, peer reports remained the optimal classifier for girls' drug dependence and high school graduation status and for boys' cases of ASPD, incarceration history, and high school graduation status. In addition, teacher reports became an optimal classifier for girls' high risk sexual behavior only, and teacher reports improved classification of girls' high school graduation after peer reports were taken into account. In sensitivity analyses utilizing an upward-adjusted kappa weight, peers remained the optimal classifier for all outcomes in the full sample. Thus, although teacher reports did provide optimal or improved classification for two of the girls' outcomes, both sets of additional analyses further supported findings of the superior classification utility of the peer report measure.

There are several possible explanations for why peers reports consistently provided better information with respect to predicting distal outcomes. The first is that peers may be reporting children's aggressive behavior more accurately than teachers or parents. At older ages, peers are considered to be the most valid informants of negative behaviors like aggression because they are privy to group-related behaviors and contextual information that parents and teachers are not (Peets & Kikas, 2006); this may be true even at younger ages. Another possible and related explanation might be that some of the behaviors that peers see and report are the behaviors that matter most for later adjustment. In other words, the way in which children behave in unstructured situations among friends and classmates may be more salient to behavior in later life than interactions with or observations by family members or school authority figures.

It is important to note, however, that the results of this study do not *necessarily* shed light on whether the peer, teacher, or parent reports provide the most accurate account of children's early aggressive behavior. All informant reports are perceptions of another's behavior, which are based in part on social schemas of aggression, reputational history of the individual in question, and situational contexts, and thus are subject to bias of varying kind and intensity. Regardless of accuracy, what the results of this study do suggest is that peer reports may provide the most *useful* information about early aggressive behavior and may be most relevant to subsequent maladaptive outcomes. Perceptions of behavior within the peer

group are related to peer acceptance and rejection; children who are perceived as aggressive by peers are more disliked by others and receive fewer friendship nominations (Crick, Murray-Close, Marks, & Mohajeri-Nelson, 2009). Problematic peer relations in elementary school can have a cascading effect in which rejection by peers leads to internalizing problems, externalizing behaviors, and association with deviant friends down the line (Lynne-Landsman, Bradshaw, & Ialongo, 2010). Parents', teachers', and peers' perceptions of a child's behavior will all play a role in the way the child is treated by those individuals, but it is possible that the perceptions of peers may have the most salient and long-lasting ramifications for later adjustment.

The treatment of informant reports as screening measures and the determination of optimal predictors and cut points with weighted κ analyses provided an alternative method of assessing the ability of informant reports to predict future outcomes as compared to the population-level regressions employed by most prior studies in this area. Though the primary objective of this study was to compare the relative ability of peer, teacher, and parent reports to predict future outcomes, the analyses also allowed us to evaluate the independent utility of these informant reports to function as effective screening measures for maladaptive outcome risk. Comparison of κ, χ^2, and efficiency statistics suggested that the peer report measure was most successful at classifying late/no high school graduation and incarceration and least successful at classifying risky sexual behavior. For outcomes with lower base rates (i.e., ASPD, drug dependence, and incarceration), specificity and efficiency were generally high for all legitimate classifiers whereas sensitivity and PPV were relatively low, suggesting that when informant reports of early aggression are treated as screening tests, their strengths lie more in correctly identifying individuals who are not at risk of these outcomes. Bennett et al. (1999) suggested that for screening measures for later externalizing behavior to be considered adequate, sensitivity and PPV should reach levels of at least .5. However, Hill et al. (2004) calculated that using cut points with resulted in sensitivity and PPV values below .5 may result in the most cost-effective utilization of resources in a selective intervention. Thus, evaluations of intervention cost, potential iatrogenic effects of the intervention, and cost to society of untreated cases should all be considered when determining the appropriate cut point for selective intervention screening.

In addition, two strengths of our study were the long interval of time between assessment of predictors and distal outcomes (up to 18 years) and the fact that all distal outcomes were measured using informant sources which were different than those of our initial aggression ratings, thus minimizing effects due to shared variance. Because both of these study characteristics would be expected to decrease the efficiency of early screening measures, the fact that peer reports remained legitimate classifiers for all five of the distal outcomes, with sensitivity close to or above Bennett et al.'s (1999) recommendations when a κ weight of .75 was used to select optimal cut points, is notable and suggests that this informant measure should be considered when designing screening measures for future research.

The relative lack of usefulness of parent ratings of early aggressive behavior was interesting for two reasons. First, surveys of mental health professionals have suggested that parents are typically perceived to be the most useful informant about young children's externalizing behavior (Loeber et al., 1990). However, our findings suggest that this perception may be

incorrect, particularly with respect to long-term distal outcomes. Aggressive and oppositional behavior displayed in peer settings is thought by many to be more problematic, from a clinical standpoint, than the same behaviors displayed at home, and this may have contributed to the reason that teacher reports in this study provided better classification utility than parent reports for long-term outcomes. Second, parent ratings of behavior often require individual home visits and participant compensation, which can make them more time-consuming and expensive to collect than school-based measures using teachers and peers. If parent reports indeed offer reduced utility as a screening measure when compared to peer and teacher reports, focusing only on in-school data collection efforts may allow intervention implementers to conserve resources with little detriment to screening accuracy.

The most significant limitation of the present research was that the number of items, wording of items, and instructions differed to various degrees across the PAI and the TOCA-R/POCA, and the varying structures among these measures may have confounded the results. It is important, therefore, to view our findings with some caution, as it is possible that the observed differences between peer, teacher, and parent reports were a function of instrument characteristics rather than qualities specific to the informants themselves. We would like to stress that the results of the study need replication with other, more similar scales before the influences of informant context and measurement structure can be satisfactorily parsed apart.

It should be noted, however, that the three aggression scales from which our items were taken represent common formats for parent, teacher, and peer report measures and have been used in prior literature, particularly in studies of prevention program screening and assessment (e.g., Dolan et al., 1993; Hill et al., 2005; Ialongo et al., 1999; Lochman et al., 1995). Peer nomination measures completed by children often contain fewer items than single-informant adult measures for several reasons, including restrictions on the classroom time allotted for assessments as well as the fact that the reliability of multiple-rater scales is more robust to decreases in item numbers than the reliability of single-rater scales. In addition, it is often not developmentally appropriate to administer the same items to adults and young children due to children's limited literacy and attention spans. Thus, although the aggression scales were not completely equivalent in structure, we believe the findings of this study will be useful to many researchers of early aggressive behavior due to the fact that the differences among these measures represent common practices in current prevention research.

A few additional limitations should be mentioned. Several of the distal outcomes, including ASPD, drug dependence, high school graduation, and high risk sex, contained components that relied on self reports of socially undesirable outcomes, and inaccurate reporting of behaviors could affect the validity of these measures. Future work may benefit from the inclusion of other reporters of distal outcomes to substantiate the self-report nature of some of the measures. The risky sexual behavior outcome, defined as engaging in sexual activity without a condom, was endorsed by a majority of the sample and thus may have been too broadly defined; it is possible that the use of a more stringent definition incorporating negative consequences of unprotected sex, such as STDs or unwanted pregnancy, would produce better classification or a different pattern of results. Additionally, the population

utilized for this study was urban, low SES, and primarily African American, which is an understudied population highly relevant to prevention research. These results, however, may not be generalizable to the population as a whole; in particular, the interpretation of aggressive behaviors by parents, teachers and peers may have different implications in other samples.

The results from this study offer insight into the usefulness of peer perceptions for predicting a multitude of outcomes in late adolescence and adulthood and present an alternative methodology for comparing the utility of multiple informant reports of the same behavior. In addition to the needed replication discussed above, these results also suggest several areas for future research. For instance, an evaluation of peers' classification utility as compared to existing parent and teacher screens for more proximal outcomes, such as conduct disorder and other externalizing behavior in later childhood and early adolescence, would be helpful for school-based targeted interventions which use multiple gating procedures to identify children at risk. Additionally, other types of distal outcomes should be explored: For example, would peers have the highest efficiency for outcomes like internalizing disorders or employment status? Finally, replication of optimal cut points for the measures used in this study will be needed before guidelines can be constructed for the use of these measures as screens in future work.

Often researchers, whether because of monetary or time constraints, are forced to choose between informants of particular behaviors. If researchers are interested in using these measures to predict distal outcomes, it is essential to pick an informant screen that provides the most useful information. This study illustrates a way to explore the best potential informant for existing data and suggests that peer reports of early aggressive behavior may be a particularly salient predictor of long-term maladaptive outcomes.

Acknowledgments

References

Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin. 1987; 101:213–232. [PubMed: 3562706]

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed.. Washington, DC: Author; 2000.

Bennett KJ, Lipman EL, Brown S, Racine Y, Boyle MH, Offord DR. Predicting conduct problems: Can high-risk children be identified in kindergarten and Grade 1? Journal of Consulting and Clinical Psychology. 1999; 67:470–480. [PubMed: 10450617]

Brook JS, Newcomb MD. Childhood aggression and unconventionality: Impact on later academic achievement, drug use, and workforce involvement. Journal of Genetic Psychology. 1995; 156:393–410. [PubMed: 8543928]

Cole DA, Hoffman K, Tram JM, Maxwell SE. Structural differences in parent and child reports of children's symptoms of depression and anxiety. Psychological Assessment. 2000; 12:174–185. [PubMed: 10887763]

Crick, NR.; Murray-Close, D.; Marks, ELP.; Mohajeri-Nelson, N. Aggression and peer relationships in school-age children: Relational and physical aggression in group and dyadic contexts. In: Rubin, KH.; Bukowski, WM.; Laursen, B., editors. Handbook of peer interactions, relationships, and groups. New York: Guilford Press; 2009. p. 287-302.

Dolan LJ, Kellam SG, Hendricks Brown C, Werthamer-Larsson L, Rebok GW, Mayer LS, Laudolff J, et al. The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. Journal of Applied Developmental Pychology. 1993; 14:317–345.

Ensminger ME, Juon HS, Fothergill KE. Childhood and adolescent antecedents of substance use in adulthood. Addiction. 2002; 97:833–844. [PubMed: 12133122]

Fergusson DM, Horwood JL, Ridder EM. Show me the child at seven: The consequences of conduct problems in childhood for psychosocial functioning in adulthood. Journal of Child Psychology and Psychiatry. 2005; 46:837–849. [PubMed: 16033632]

Hill LG, Lochman JE, Coie JD, Greenberg MT. The Conduct Problems Prevention Research Group. Effectiveness of early screening for externalizing problems: Issues with screening accuracy and utility. Journal of Consulting and Clinical Psychology. 2004; 72:809–820. [PubMed: 15482039]

Huesmann LR, Eron LD, Dubow EF. Childhood predictors of adult criminality: Are all risk factors reflected in childhood aggressiveness? Criminal Behavior and Mental Health. 2002; 12:185–208.

Huesmann LR, Eron LD, Lefkowitz MM, Walder LO. Stability of aggression over time and generations. Developmental Psychology. 1984; 20:1120–1134.

Ialongo NS, Werthamer L, Kellam SG, Brown CH, Wang S, Lin Y. Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. American Journal of Community Psychology. 1999; 27:1999.

Kaiser HF. The application of electronic computers to factor analysis. Educational and Psychological Measurement. 1960; 20:141–151.

Kellam SG, Brown CH, Poduska JM, Ialongo NS, Wang W, Toyinbo P, et al. Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. Drug and Alcohol Dependence. 2008; 95S:S5–S28. [PubMed: 18343607]

Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of risk factors for clinical or policy significance. Psychological Methods. 1999; 3:257–271.

Leoutsakos, JS. Kappa Tree User's Manual. 2007. Retrieved from www.jhsph.edu/prevention/publications/index

Lochman JE. The Conduct Problems Prevention Research Group. Screening of child behavior problems for prevention programs at school entry. Journal of Consulting and Clinical Psychology. 1995:63.

Loeber R, Green SM, Lahey BB. Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. Journal of Clinical Child Psychology. 1990; 19:136–143.

Lynne-Landsman SD, Bradshaw CP, Ialongo NS. Testing a developmental cascade model of adolescent substance use trajectories and young adult adjustment. Development and Psychopathology. 2010; 22:933–948. [PubMed: 20883591]

Mrazek, P.; Haggerty, R. Appendix A: Summary, Committee on Prevention of Mental Disorders. In: Mrazek, P.; Haggerty, R., editors. Reducing risks for mental disorders: Frontiers for preventive intervention research. Washington, D.C: National Academy Press; 1994. p. 487-553.

Peets K, Kikas E. Aggressive strategies and victimization during adolescence: Grade and gender differences, and cross-informant agreement. Aggressive Behavior. 2006; 32:68–79.

Pekarik E, Prinz R, Leibert C, Weintraub S, Neale J. The Pupil Evaluation Inventory: A sociometric technique for assessing children's social behavior. Journal of Abnormal Child Psychology. 1976; 4

Petras H, Buckley JA, Leotsakos JS, Ialongo NS. The use of multiple versus single assessment time points to improve screening accuracy in identifying children at risk for later serious antisocial behavior. 2011 Manuscript submitted for publication.

Petras H, Schaeffer CM, Ialongo NS, Hubbard S, Muthén B, Lambert SF, et al. When the course of aggressive behavior in childhood does not predict antisocial outcomes in adolescence and young adulthood: An examination of potential explanatory variables. Development and Psychopathology. 2004; 16:919–941. [PubMed: 15704821]

Reinherz HZ, Giaconia RM, Carmola AM, Wasserman MS, Paradis AD. General and specific childhood risk factors for depression and drug disorders by early adulthood. Journal of the American Acadamy of Child and Adolescent Psychiatry. 2000; 39:223–231.

Renk K. Cross-informant ratings of the behavior of children and adolescents: The "gold standard". Journal of Child and Family Studies. 2005; 14:457–468.

Robins, LN.; Cottler, LB.; Bucholz, KK.; Compton, WM.; North, CS.; Rourke, KM. Diagnostic Interview Schedule for the DSM-IV (DIS-IV). St. Louis, MO: Washington University, Department of Psychiatry; 2000.

Salkever DS, Johnston S, Karakus MC, Ialongo NS, Slade EP, Stuart EA. Enhancing the net benefits of disseminating efficacious prevention programs: A note on target efficiency with illustrative examples. Administration and Policy in Mental Health and Mental Health Services Research. 2008; 35:261–269. [PubMed: 18343990]

Schaeffer CM, Petras H, Ialongo NS, Masyn KE, Hubbard S, Poduska J, Kellam S. A comparison of girls' and boys' aggressive-disruptive behavior trajectories across elementary school: Prediction to young adult antisocial outcomes. Journal of Consulting and Clinical Psychology. 2006; 74:500–510. [PubMed: 16822107]

Serbin LA, Peters PL, McAffer VJ, Schwartzman AE. Childhood aggression and withdrawal as predictors of adolescent pregnancy, early parenthood, and environmental risk for the next generation. Canadian Journal of Behavioral Science. 1991; 23:318–331.

Shaffer D, Fisher P, Lucas C, Dulcan M, Schwab-Stone M. NIMH diagnostic interview schedule for children version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. Journal of the American Academy of Child and Adolescent Psychiatry. 2000; 39:28–38. [PubMed: 10638065]

Substance Abuse and Mental Health Services Administration. Results from the 2010 National Survey on Drug Use and Health: Summary of National Findings. Rockville, MD: Author; 2011. NSDUH Series H-41, HHS Publication No. (SMA) 11-4658

Timmermans M, van Lier PAC, Koot HM. Which forms of child/adolescent externalizing behaviors account for late adolescent risky sexual behavior and substance use? Journal of Child Psychology and Psychiatry. 2008; 49:386–394. [PubMed: 17979959]

Werthamer-Larsson L, Kellam SG, Wheeler L. Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. American Journal of Community Psychology. 1991; 19:585–602. [PubMed: 1755437]

Yesavage J. Introduction to ROC5. 2008 Retrieved from http://www.stanford.edu/~yesavage/ROC.html.

APPENDIX

Initial test characteristics and description of Kappa Tree branching pattern for full sample when κ weight = .75 (emphasis placed on avoiding false negatives)

First cut test characteristics for teacher, parent, and peer reports with kappa weight of .75

	best cut point	χ^2	κ	eff	sen	spec	PPV
ASPD							
peers	.41	15.26[***]	.19	.73	.42	.78	.26
teacher	.13	11.62[**]	.16	.68	.45	.72	.23
parent	.15	.86	.05	.55	.49	.56	.18

	best cut point	χ^2	κ	eff	sen	spec	PPV
Drug dependence							
peers	.31	11.06[**]	.16	.72	.44	.75	.17
teacher	−.18	3.24	.08	.60	.50	.62	.13
parent	.50	4.05[*]	.10	.68	.41	.72	.15
Ever incarcerated							
peers	.41	32.07[**]	.27	.76	.56	.78	.21
teacher	.43	21.81[**]	.21	.77	.45	.80	.20
parent	.54	1.61	.05	.55	.54	.55	.11
High risk sex							
peers	.06	4.46[*]	.10	.57	.69	.40	.64
teacher	.88	2.96	.08	.58	.86	.19	.60
parent	.14	1.68	.05	.54	.57	.48	.61
Late or no h.s. graduation							
peers	.76	29.65[**]	.26	.62	.91	.26	.60
teacher	.28	14.98[***]	.18	.59	.81	.33	.58
parent	.50	8.00[**]	.14	.58	.75	.36	.59

[***] $p < .001$.

[**] $p < .01$.

[*] $p < .05$.

Outcome	First Cut	Second Cut	Third Cut

ASPD *N*=641
Outcome *N*=100 (16%)

Peer > .41 *N*=137
Outcome *N*=35 (26%)

Peer ≤ .41 *N*=417
Outcome *N*=49 (12%)

Drug dep. *N*=641
Outcome *N*=68 (11%)

Peer > .61 *N*=150
Outcome *N*=26 (17%)

Teacher > .13 *N*=102
Outcome *N*=12 (12%)

Peer ≤ .61 *N*=404
Outcome *N*=32 (8%)

Teacher ≤ .13 *N*=48
Outcome *N*=14 (29%)

Incarcerated *N*=687
Outcome *N*=69 (10%)

Peer > .81 *N*=97
Outcome *N*=23 (24%)

Peer ≤ .81 *N*=493
Outcome *N*=32 (6%)

High risk sex *N*=641
Outcome *N*=372 (58%)

Peer > .06 *N*=364
Outcome *N*=231 (63%)

Teacher > -.63 *N*=189
Outcome *N*=131 (69%)

Parent > .15 *N*=64
Outcome *N*=30 (47%)

Peer ≤ .06 *N*=190
Outcome *N*=103 (54%)

Teacher ≤ -.63 *N*=174
Outcome *N*=100 (57%)

Parent ≤ .15 *N*=93
Outcome *N*=61 (66%)

Late/no HS grad *N*=635
Outcome *N*=295 (46%)

Peer > .21 *N*=164
Outcome *N*=118 (59%)

Teacher > 1.63 *N*=41
Outcome *N*=49 (78%)

Teacher ≤ 1.63 *N*=41
Outcome *N*=52 (57%)

Peer ≤ .21 *N*=391
Outcome *N*=131 (37%)

Parent > .50 *N*=101
Outcome *N*=48 (48%)

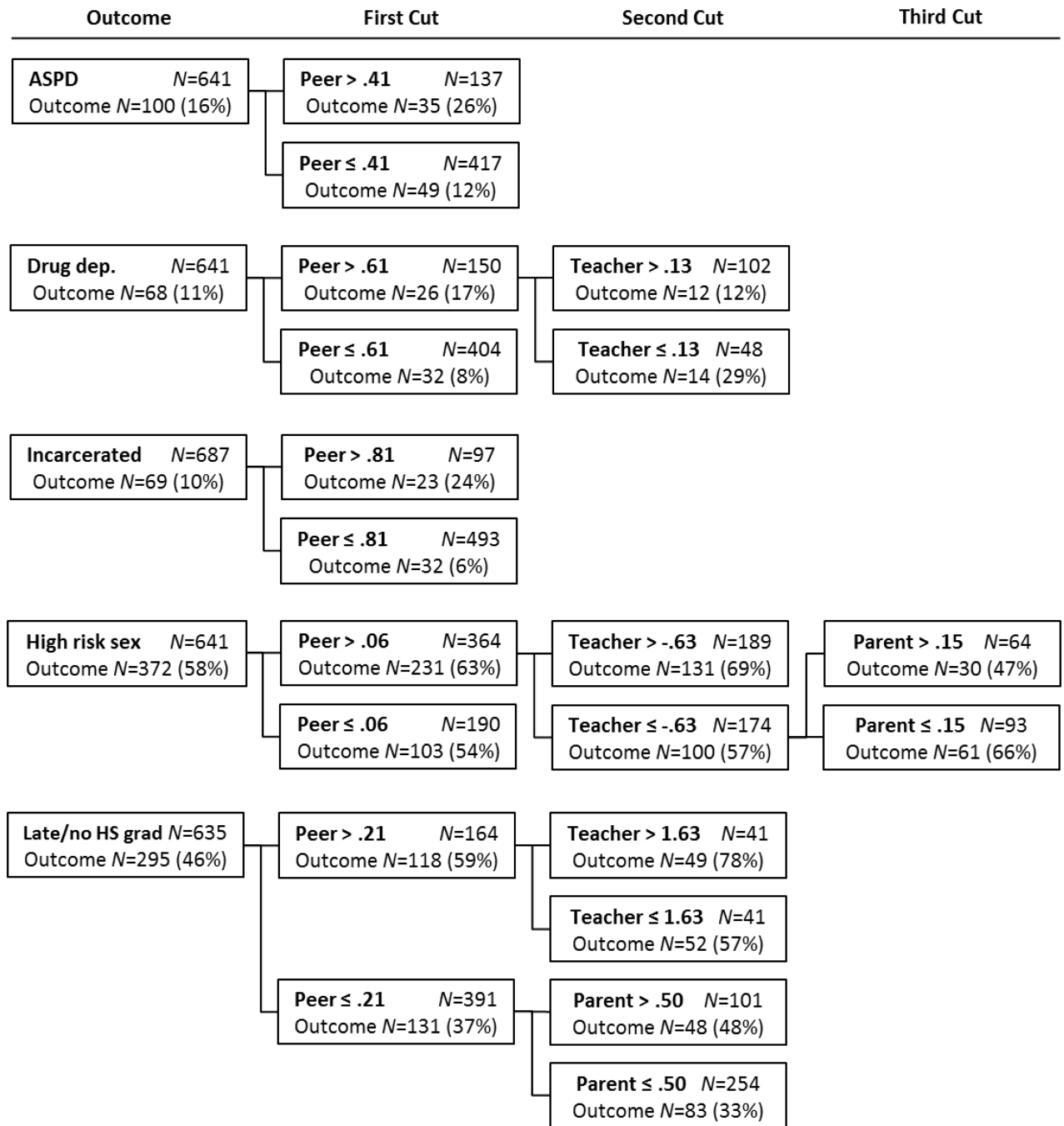**Parent ≤ .50** *N*=254
Outcome *N*=83 (33%)

Figure 1.
Kappa Tree branching pattern for full sample.

Table 1

Descriptive statistics for first grade ratings of aggressive behavior

Informant	Peer (r)	Teacher (r)	Mean	SD	Range	Skew	Kurtosis
Peer			.15	.14	.00–.79	1.60	2.62
Teacher	.64***		1.69	.96	1.00–6.00	1.91	3.64
Parent	.20**	.19**	1.51	.41	1.00–3.71	1.33	3.08

***
  p < .001.

**
  p < .01.

Table 2

Prevalence of distal outcomes in late adolescence/early adulthood

Outcome	Full sample	Girls	Boys
ASPD	100 (15.6%)	29 (9.7%)	71 (20.8%)
Drug dep.	68 (10.6%)	32 (10.7%)	36 (10.6%)
Incarcerated	69 (10.0%)	8 (2.5%)	61 (16.5%)
Late or no HS grad	295 (46.5%)	182 (38.9%)	158 (53.1%)
High risk sex	372 (58.0%)	193 (64.3%)	179 (52.5%)

Note. Percentage values represent valid percentages among participants with nonmissing data. *N* values for nonmissing data were as follows: ASPD, drug dep., and high risk sex = 641; ever incarcerated = 687; late/no HS grad. = 635.

Table 3

First cut test characteristics for teacher, parent, and peer reports

	best cut point	χ^2	κ	eff	sen	spec	PPV
ASPD							
peers	.41	15.26 ***	.16	.73	.42	.78	.26
teacher	.13	11.62 **	.12	.68	.45	.72	.23
parent	85	.72	.04	.73	.21	.82	.18
Drug dependence							
peers	.61	9.26 **	.12	.77	.35	.82	.18
teacher	2.98	4.28*	.06	.88	.06	.98	.27
parent	1.55	3.31	.08	.85	.13	.93	.20
Ever incarcerated							
peers	.81	28.43 ***	.21	.82	.42	.86	.24
teacher	1.33	17.73 ***	.16	.83	.28	.89	.23
parent	.15	1.61	.03	.55	.54	.55	.11
High risk sex							
peers	.06	4.62*	.09	.56	.69	.40	.64
teacher	.88	2.96	.06	.58	.86	.19	.60
parent	.15	1.68	.06	.54	.57	.48	.61
Late or no h.s. graduation							
peers	.21	30.29 ***	.22	.63	.80	.41	.62
teacher	−.33	17.83 ***	.17	.59	.63	.54	.61
parent	.50	8.00 **	.12	.58	.75	.36	.59

*** $p < .001$.

** $p < .01$.