Published in final edited form as:
J Consult Clin Psychol. 2004 October ; 72(5): 809–820. doi:10.1037/0022-006X.72.5.809.

Effectiveness of Early Screening for Externalizing Problems: Issues of Screening Accuracy and Utility

Laura G. Hill,
Department of Human Development, Washington State University

John E. Lochman,
Department of Psychology, University of Alabama

John D. Coie,
Department of Psychology, Duke University

Mark T. Greenberg, and
Department of Human Development and Family Studies, Pennsylvania State University

The Conduct Problems Prevention Research Group

Abstract

Accurate, early screening is a prerequisite for indicated interventions intended to prevent development of externalizing disorders and delinquent behaviors. Using the Fast Track longitudinal sample of 396 children drawn from high-risk environments, the authors varied assumptions about base rates and examined effects of multiple-time-point and multiple-rater screening procedures. The authors also considered the practical import of various levels of screening accuracy in terms of true and false positive rates and their potential costs and benefits. Additional research is needed to determine true costs and benefits of early screening. However, the results indicate that 1st grade single- and multiple-rater screening models effectively predicted externalizing behavior and delinquent outcomes in 4th and 5th grades and that early screening is justified.

High levels of early aggressive and oppositional behaviors in children can persist and, eventually, develop into chronic patterns of delinquent and psychopathological behaviors (Coie, Terry, Lenox, & Lochman, 1995; Nagin & Tremblay, 1999). These early externalizing behaviors are associated with increased risk for multiple negative outcomes academically, socially, in physical and mental health, and in job performance (Kaplow, Curran, Dodge, & Conduct Problems Prevention Research Group, 2002). Children's externalizing behaviors can also lead to increasingly negative reactions from others. As peers and adults become more rejecting and hostile in reaction to the children's behavior, the children's externalizing problems (EP) can intensify (Conduct Problems Prevention Research Group, 1992; Loeber, 1990). Early intervention aimed at minimizing aggression,

---

Correspondence concerning this article should be addressed to Laura G. Hill, Department of Human Development, Washington State University, Pullman, WA 99164-6236. laurahill@wsu.edu.

The participating investigators of the Conduct Problems Prevention Research Group, listed in alphabetical order, are as follows: Karen L. Bierman, Department of Psychology, Pennsylvania State University; John D. Coie, Department of Psychology, Duke University; Kenneth A. Dodge, Center for Child and Family Policy, Duke University; E. Michael Foster, Department of Health Policy and Administration, Pennsylvania State University; Mark T. Greenberg, Department of Human Development and Family Studies, Pennsylvania State University; John E. Lochman, Department of Psychology, University of Alabama; Robert J. McMahon, Department of Psychology, University of Washington; and Ellen E. Pinderhughes, Department of Psychology and Human Development, Vanderbilt University.

For additional information concerning Fast Track, see http://www.fasttrackproject.org.

oppositionality, and other externalizing behaviors can interrupt this cycle before increasingly negative behavior patterns become entrenched. Such early intervention may be provided either universally or to a targeted group.

Although universal interventions are provided to all members of a given population, targeted interventions are provided only to those identified as showing early signs of developing EP and are intended to prevent development of the full-blown disorder (Mrazek, Haggerty, & National Academy of Sciences Institute of Medicine, Division of Biobehavioral Sciences and Mental Disorders, Committee on Prevention of Mental Disorders, 1994). To the extent that a screening procedure can accurately identify future cases of a disorder, a targeted intervention is preferable for several reasons: It can be more focused, more efficient, and more intensive than a universal intervention (Lochman & Conduct Problems Prevention Research Group, 1995; Offord, 2000).

Targeted interventions for conditions with low base rates in the population have the potential to be especially efficient and cost-effective, because they are delivered to fewer individuals. However, the lower the base rate of a condition, the greater the need for accuracy in screening. Potential savings of time, money, and other resources offered by an indicated intervention represent true savings only to the extent that the intervention is effective, delivered to those who truly need it, and not delivered to those who do not need it.

Characteristics of an Effective Screening

The function of a screening procedure is to classify individuals into two groups: those who are at risk and likely to develop EP and those who are not. The test of screening accuracy is based on the association of this risk classification with classification on another binary outcome: those who have developed EP and those who have not. Statistics that test a screen's accuracy are derived from the matrix of these two binary outcomes and include *sensitivity* (the proportion of true positives [TPs] correctly identified); *specificity* (the proportion of true negatives correctly identified); *positive predictive value*, or PPV (the proportion of those classified as at-risk who did develop the outcome); and *negative predictive value*, or NPV (the proportion classified as not at-risk in whom the outcome is absent). As the proportion of TPs rises, so does the proportion of false positives (FPs). The trade-off between sensitivity and specificity across different cutpoints of a test can be graphed onto a curve known as the receiver operating characteristic, or ROC, of a test (Black, Panzer, Mayewski, & Griner 1991; McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000). ROC curves can be helpful in determining optimal cutpoints or, if multiple tests are being considered, multiple curves can be used to compare the effectiveness of various combinations of tests.

A number of considerations that may affect both accuracy and usefulness of screening for externalizing disorder have been identified (Bennett, Lipman, Racine, & Offord, 1998; Offord, 2000) but not yet systematically explored. Considerations of accuracy include composition of the sample used to validate screening procedures, base rate of EP in different populations and sexes, timing and setting of the screening procedure, and type and timing of outcome measures. Considerations of utility include the practical meaning of a screen in terms of TPs and FPs and determination of the costs and benefits of applying a screening procedure.

Considerations Affecting Accuracy of a Screen

Criterion Sample

A first consideration in designing an effective screen is that the criterion group used to validate a screening procedure should be representative of the population to be screened. Samples with high numbers of severe cases (e.g., clinic samples) result in inflated estimates of screening sensitivity and specificity, simply because individuals who would most likely be misdiagnosed (those with mild and moderate symptoms) are underrepresented (Bennett et al., 1998). Therefore, tests of specificity and sensitivity will not be stable across populations with noncomparable case mixes (Bennett et al., 1999; Black et al., 1991). An effective screening for EP requires validation on a sample of children representing the full range of early behaviors.

Base Rate of Disorder

A second (and often overlooked) consideration is that the base rate of the expected outcome will have a significant effect on the PPV and NPV of a screen (Meehl & Rosen, 1955). Sensitivity and specificity of tests may sound impressive when reported without reference to PPV, NPV, and base rates. For example, a test with sensitivity of.80 and specificity of.95 has a PPV of about 74% if the base rate is 15%, but the PPV is reduced to 46% (lower than chance) if the base rate is 5% (Bennett et al., 1998). Estimates of expected base rates of childhood externalizing disorders in the U.S. population at large range from approximately 3% to 31% (Bennett et al., 1998), with most estimates falling between 5% and 15% (Conduct Problems Prevention Research Group, 2004). Base rates for EP may be higher in samples from a high-risk community, however, resulting in higher PPV.

Another consideration in selecting a criterion sample for EP is that base rates are higher for boys than for girls; thus, the mix of boys and girls in a criterion sample will affect PPV and NPV (Zoccolillo, Tremblay, & Vitaro, 1996). Relatedly, although testing for EP has traditionally focused on overt behaviors, early signs and symptoms of EP may be different for girls than for boys (Flanagan, Bierman, Kam, & Conduct Problems Prevention Research Group, 2002). Some researchers have resolved these problems by screening only boys (Tremblay, Pihl, Vitaro, & Dobkin, 1994) or only girls (Zoccolillo et al., 1996). A comparison of how test characteristics differ between boys and girls on a standard screening procedure has not been documented.

Timing and Setting of Screen

Early intervention is key in preventing the development of major externalizing disorders (Conduct Problems Prevention Research Group, 1992; Tremblay, LeMarquand, & Vitaro, 1999). Interventions with children identified as at risk, some beginning as early as preschool, have been shown to reduce the rate of later violent incidents and felony crimes (Office of the Surgeon General, 2001). Moreover, early intervention may be considerably more cost effective than later intervention (Cohen, 1998). However, early screening poses special problems (Bennett & Offord, 2001). Many children who do not have later EP display problem behaviors such as hitting, biting, and oppositionality at younger ages. An accurate screen must be able to distinguish true signs of a future disorder from behaviors that resemble that disorder but are not indicative of its development. The accuracy of existing screening procedures as a function of their specific timing has been shown to be greater with older children than with younger children (Bennett & Offord, 2001) but has not been systematically explored.

Timing and Type of Outcome Measures

The variability in estimates of base rates reflects in some degree variability in the definition of EP and in timing of their measurement. EP are a matter of degree, but screening procedures call for a categorical outcome, thus necessitating decisions about cutpoints to define that outcome. The nature and severity of EP may also change over the course of development: Outcomes measured in early elementary school focus on oppositionality, aggression, and hyperactivity and do not include the antisocial and psychopathological behaviors measured in adolescence. The type of outcome, the time of measurement, and cutpoints used to determine outcome will all affect estimates of base rates and test characteristics of the screening.

Considerations Affecting Utility of a Screen

Screening procedures for EP use continuous predictors (scale scores) to classify children into two groups: those who need intervention and those who do not. The classification is accomplished by using a diagnostic cutpoint or by determining a decision threshold on an ROC curve. When the scale cutpoint or decision threshold is set liberally, to capture as many potential true cases as possible, the number of FPs increases. Conversely, when the decision threshold is set more conservatively, to minimize FPs, the number of TPs decreases (Swets et al., 2000). Therefore, determining the utility of a screen necessitates an understanding of the costs and benefits of a decision about where to set a cutpoint, given a particular level of screening accuracy (McFall & Treat, 1999; Swets et al., 2000). In the case of screening for EP, such a decision involves consideration of (a) the cost of an intervention relative to its benefit when delivered to those who need it, as well as (b) its cost when delivered to those who do not need it and (c) the cost of not delivering an intervention to those who do need it.

Costs and benefits are both direct and indirect. Direct costs of targeted interventions vary widely both across intervention programs and across individuals within a program, when an intervention is modified for individual need (Bierman, Nix, Maples, & Murphy, 2003). Direct costs increase as the rate of TPs and FPs increases. Some commonly measured direct benefits of interventions include less use of clinical or school resources in childhood (Jones, Dodge, Foster, Nix & Conduct Problems Prevention Research Group, 2002) and reduced taxpayer and victim costs from reduced crime rates of adolescents and young adults (Washington State Institute for Public Policy [WSIPP], 1998). Indirect benefits of an intervention include the potential for higher academic achievement, increased productivity, lower service utilization, and, ultimately, lower welfare costs (Cohen, 1998; Jones et al., 2002; Office of the Surgeon General, 2001). A possible indirect cost of targeted interventions that concerns some advocates of universal interventions (e.g., Offord, 2000) is the potential stigmatization of children identified as having EP who are not, in fact, at risk—that is, children in the FP group.

Effectiveness of Existing Screening Procedures

Potential solutions to measurement problems posed by the developmental nature of EP include multiple-gating procedures (August, Realmuto, Crosby, & MacDonald, 1995; Lochman & Conduct Problems Prevention Research Group, 1995; Loeber, 1990), multiple-rater procedures (Coie et al., 2002; Lochman & Conduct Problems Prevention Research Group, 1995; Offord et al., 1996; Walker et al., 1988), multiple-time-point procedures (Coie et al., 2002), and multiple problem profiles (Flanagan et al., 2002). Approaches may be combined variously and applied to both outcomes and predictors. In the Fast Track experiment (Conduct Problems Prevention Research Group, 1992, 2004), children were screened using a combined multiple-gating and multiple-rater approach. Children were rated for problem behaviors first by kindergarten teachers. Those who scored highest, passing the

first gate, were rated next by parents. The at-risk group was then selected from the combined teacher–parent scores of the gated group (Lochman & Conduct Problems Prevention Research Group, 1995). This screening was effective in predicting to first-grade behavior problems, showing optimal sensitivity and specificity in predicting to a multiple-time-point outcome (problem behaviors either at entry to or at the end of first grade). However, questions remain about screening effectiveness over longer periods and in predicting to important antisocial and psychopathological outcomes that develop in secondary school and later.

Some investigators argue that current screening procedures are inadequate for the task of accurately classifying children at risk for developing EP (Bennett et al., 1998; Offord, 2000). Bennett and colleagues (Bennett et al., 1999; Bennett & Offord, 2001) have set criteria of at least 50% sensitivity and 50% PPV as minimally adequate to justify screening for a targeted intervention. With these arbitrary screening values, at least 50% of children who are identified as having EP will, in fact, develop later EP (PPV). Also, 50% of the children who would have eventually developed EP will have been identified by the screen (sensitivity). Using a nonclinic sample of kindergarten and first-grade children to predict EP 24 and 30 months later, they found sensitivity, specificity, and PPV to be below their preset criteria. They concluded that although there is stability over time in EP, significant misclassification will occur when these behaviors are used to designate high-risk status. Their preset criteria provide a useful starting point for judging screening accuracy and a convenient frame of reference for comparison of accuracy with other samples. However, this approach does not explore the accuracy of current screening procedures in terms of their utility. To date, there has been no published research on screening for EP that addresses the issue of costs and benefits of varying decision thresholds in a screening procedure (i.e., increasing or decreasing both TPs and FPs). Therefore, there is a need to explore further the accuracy of screening for EP, especially for delinquency and psychopathology at substantially later time points, as well as a need to address the question of the practical utility of a screen given its accuracy.

The current study had two sets of goals related to refining the understanding of screening effectiveness: one set exploring issues of screening accuracy and a second set exploring issues of screening utility. In terms of accuracy, a first goal was to determine base rates of EP in high-risk samples and to explore how assumptions about base rates affect test characteristics of screening procedures. Using data from schools in high-risk neighborhoods, we varied assumptions about the base rate of the outcome to examine effects on sensitivity, specificity, and PPV. A second goal was to extend the time frame for outcome measures beyond 1 or 2 years (Bennett et al., 1998). Our outcomes were measured 4 and 5 years after the initial screening, at the end of fourth and fifth grades, when students were initiating delinquent behaviors. A third goal was to document how test characteristics of various screening procedures differ by sex. A final goal was to compare performance on similar screens from two community samples. We replicated analyses reported by Bennett et al. (1999), including a broad array of screening tests and comparing results across the two community samples. Throughout these analyses, we used the preset criteria for predictive value reported by Bennett et al. (1999) as a reference point.

In terms of utility, our primary goal was to set forth practical implications of the various screening procedures described above. We looked at how predicting to three different outcomes was reflected in numbers of TP and FP classifications. We also looked at the changes in TP and FP rates created by moving the decision threshold for a single outcome on a ROC curve and at how these changes affect societal costs of program delivery.

Method

Study Sample

Sample description—Schools were identified in high-risk areas of four communities: Durham, North Carolina; Nashville, Tennessee; Seattle, Washington; and rural central Pennsylvania. Risk was determined on the basis of high poverty rates and low education rates among parents of schoolchildren and on schools being located in catchment areas with high crime rates. Schools at each site were matched according to size, racial composition, and poverty levels, then randomly assigned to control or intervention conditions. The intervention, designed to prevent the development of EP, has been extensively described elsewhere (Conduct Problems Prevention Research Group, 1992).

A normative sample of 396 children was selected. This sample was composed of children who did not receive the intervention because they were from control schools where intervention services were not provided. Participants were first stratified to represent the school population according to race, sex, and decile of the teacher screen problem behavior scores. The normative sample was then randomly selected from the race and sex groups within each teacher screen decile, thus ensuring that the normative sample was representative of the larger school population. The sample used for this study included all normative participants. Of these participants, 44% were African American, 52% were European American, and 4% were of other ethnic backgrounds. Fifty percent of the participants were male.

Attrition—Data reported in this study are from the end of the school year in kindergarten, first, fourth, and fifth grades. In fifth grade, 48 students (12%) did not have teacher data; 50 (13%) lacked parent data; and 70 (18%) were missing data from both sources. The sample size for analyses that combined all years and both parent and teacher data was 309, or 80% of the original sample, in fifth grade. There were no significant differences in attrition by race, sex, or status on the initial teacher and parent screens.

Screening Variables

We compared effectiveness of single-rater versus multiple-rater and single-time-point versus multiple-time-point screening strategies. Teacher and parent screening instruments, collected at the end of kindergarten and the end of first grade, were entered as separate predictors in logistic regression analyses.

Teacher screen—The teacher screen was composed of 14 items from the Teacher Observation of Classroom Adaptation—Revised (TOCA–R) developed by Werthamer-Larsson, Kellam, and Wheeler (1991). Items rated on a scale ranging from 0 (*almost never*) to 5 (*almost always*) measure externalizing behaviors relevant to a classroom situation (e.g., “breaks rules,” “harms others,” and “takes others' property”). Internal consistency of the screen was high in both Year 1 (α =.87) and Year 2 (α =.92).

Parent screen—In the summer after kindergarten, parents responded by telephone or in person to the Child Problem Behavior Scale, composed of 24 items about child externalizing behaviors. Items were drawn from the Child Behavior Checklist (CBCL; Achenbach, 1991; Nix, 2001) and the Revised Problem Behavior Checklist (RPBC), which itself was derived from the CBCL. Parents rated kindergarten behaviors such as “whining,” “bullying,” and “disobedient” on a scale ranging from 1 (*never*) to 4 (*often*). Internal consistency was high (α =.87). For the first grade sample, parents responded to 21 identical or similar items (on the RPBC) about externalizing behaviors drawn from the CBCL. Items were rated from 0 (*almost never*) to 2 (*often*). The resulting scale had high internal consistency (α =.88).

Outcome Variables

We used outcome variables from three domains and multiple raters: externalizing behaviors (parent and teacher report), delinquency (youth self-report), and diagnosed psychopathology (parent and youth report).

Externalizing: Parent, teacher, and combined outcomes—We used the 33-item Externalizing subscale of the CBCL (parent version) for the parent-rated outcome and the Authority Acceptance subscale (AAC) of the TOCA–R for the teacher-rated outcome. Both measures were administered in both fourth and fifth grades. Both the CBCL and the AAC, a 10-item subset of the teacher screen described above, had high internal consistency (α =.92 for both measures in both Grades 4 and 5).

We standardized and summed parent and teacher scale scores and then used different cutpoints to form several dichotomous outcome variables, each defining a problem outcome group (that is, students considered to have the EP outcome) and a no-outcome group (all other students). These outcome variables included *single time point* EP groups (externalizing in fourth grade, externalizing in fifth grade) with a presumed 20% annual base rate in each grade (i.e., with the cutpoint set to capture the top 20% on the combined parent–teacher score). We also created a *multiple time point* EP group (externalizing across both years). For this latter outcome, the base rate was determined by how many children fell into the annual EP group in both fourth and fifth grades. When we assumed an annual base rate of 20%, the multiple time point base rate was 11%.

To explore more conservative assumptions about base rates and to compare our data with the data presented by Bennett et al. (1999), we also created annual outcome groups by taking the top 15% (rather than those in the top 20%) of students in each year. Those students who were in the top 15% in both fourth and fifth grade then made up the second multiple time point outcome group; the base rate for this group was 8%.

Self Reported Delinquency (SRD): Child self report—In Grades 4 and 5, students reported the frequency with which they had committed any of 27 delinquent acts over the course of the previous year (e.g., “carried a weapon,” “ran away from home,” “attacked someone”). Items were taken directly or derived from the National Youth Survey (Elliott, Huizinga, & Morse, 1986). All 27 items were classified into the following nonoverlapping categories of delinquent behavior: crimes against persons, theft, vandalism, school delinquency, alcohol use, and drug use.

The EP group, 22% of the sample, was composed of those students who reported committing at least one delinquent act in each of two or more categories in both Grades 4 and 5.

DISC: Parent and child report of conduct and oppositional disorders—Parents and children were given the Diagnostic Interview Schedule for Children (DISC) the summer after fifth grade. The DISC problem outcome group (base rate of 18%) consisted of children who received at least one externalizing diagnosis, according to both parent and child reports.

Case Mix of Boys and Girls

We split the externalizing outcome group by sex and looked at the resulting base rate within sex. As expected, boys made up the larger portion of the problem outcome groups: Base rate for boys in the 20% whole-sample outcome group was 27%; for girls, it was 12%.

Results

Descriptive Results

Mean scores, cutoff scores for 15% and 20% base rates, and correlations for continuous screening and outcome variables are presented in Table 1.

Missing Data

Missingness in Grades 4 and 5 was unrelated to sex, race, or values on the kindergarten and first-grade screening variables. Nevertheless, we conducted analyses using multiply imputed data; results did not differ notably from those analyses conducted on participants with complete data.

Order of Outcome Analyses

We looked first at the effects of varying characteristics of the screening itself: using single versus multiple time points and single versus multiple raters to predict to the parent–teacher externalizing outcome. We then added multiple measures to the screen, equivalent to those reported by Bennett et al. (1999), and looked at whether these additional variables increased accuracy of prediction to externalizing, delinquency, and psychopathology outcomes. Next, holding the screening variables constant and predicting to the parent–teacher outcome, we explored sex differences in screening accuracy as well as differences when assumptions about base rates are changed. Third, we used ROC curves to compare test characteristics in predicting to externalizing, delinquency, and psychopathology outcomes. Finally, we explored costs and benefits of increasing TP and FP rates by varying decision thresholds.

Analysis Strategy

Logistic regression—We used logistic regression for all analyses to accommodate multiple continuous scale scores in predicting to dichotomous outcomes. Logistic regression allows for quantitative comparison of different predictive models and produces statistics for the full range of potential screening variable cutpoints.

Model comparison—We include Nagelkerke's R^2 (Bennett et al., 1999; Nagelkerke, 1991) for ease of model comparison with the data presented by Bennett et al. (1999). However, for selection of the best approximating model from the a priori set of model candidates within tables, we use Akaike's information criterion (AIC) and its corresponding delta statistic (Burnham & Anderson, 1998).

Effects of Varying the Screening Variables

Comparison of single versus multiple time points and raters—We tested six screening models, all predicting to the fifth grade parent–teacher EP outcome with a presumed base rate of 20%. The single-rater models included (a) teacher rating from kindergarten only, (b) teacher rating from first grade only, and (c) teacher rating from both kindergarten and first grade. Multiple-rater models were (d) teacher and parent ratings from kindergarten only, (e) teacher and parent ratings from first grade only, and (f) teacher and parent ratings from both kindergarten and first grade.

In Table 2, we present test characteristics from all six models. To facilitate comparison across studies, we used a table format similar to that used by Bennett et al. (1999). We explore whether these models meet the preset criteria for accuracy of at least.50 sensitivity and at least.50 PPV, also used by Bennett et al. (1999). Assuming that their base rate was 15% and that their TP rate equaled their screened positive rate, they held specificity constant at.91, the point at which PPV equals 50% when base rate equals 15%. Recall that the

relation of PPV to sensitivity and specificity changes as the base rate changes: The higher the base rate, the lower the level of specificity required to achieve a PPV of 50%. This relation can be seen on the ROC curve presented in Figure 1. When we use the same assumptions as Bennett et al. (1999) did, at our 20% assumed base rate, the specificity necessary to achieve PPV greater than or equal to.50 when sensitivity is held constant at.50 is approximately.875 (see Table 2). As noted earlier, logistic regression produces test characteristics for all possible cutpoints along the range of predictor values. Thus, in the column under the heading PPV >.50, we record the sensitivity of the test when the cutpoint is set so that specificity equals.875. To meet the criteria for accuracy set forth by Bennett et al. (1999), we want to see a sensitivity of at least.50 in this column. Similarly, in the column under the heading Sensitivity >.50, we hold the sensitivity of the test constant at.50 and record the specificity, looking to see whether specificity at this cutpoint is.875 or greater. We can see that Models 2, 3, 5, and 6 exceed these criteria.

Turning to comparison of the models with one another, we note that Model 6, including all predictors, has the lowest AIC and, therefore, a delta of 0. However, Model 5, with teacher and parent ratings from first grade only, has a delta of 2, indicating that it is essentially no different from Model 6. In contrast, Models 1–4, using as predictors only teacher screens or only kindergarten parent and teacher screens, all have deltas of 21 or greater. In sum, the first four models have markedly less predictive value than the last two do, and the last two are equivalent. The test characteristics also indicate that although Models 5 and 6 are equivalent in terms of predictive value, Model 5 is more parsimonious (i.e., it needs ratings from only 1 year). Therefore, we use the Model 5 screening from this point forward to predict to different outcomes.

Effects of including additional variables in the screen—We tested the effects of including in the screen a number of additional variables, similar to those included in analyses reported by Bennett et al. (1999): family socioeconomic status, maternal depression scores, parenting practices, parent-rated child attention-deficit/hyperactivity disorder symptoms, and parent satisfaction. In Table 3, we present comparisons of test characteristics for Model 5 only versus Model 5 plus the parent and demographic variables. In addition to the fifth grade parent–teacher externalizing outcome, we looked at prediction of these models to fourth- and fifth-grade delinquency and fifth-grade psychopathology outcomes. We found that the addition of numerous variables from multiple domains did not appreciably increase predictive value, regardless of the outcome. In predicting to the parent–teacher externalizing outcome, Model 5 alone was markedly superior to Model 5 plus additional variables, as indicated by the multiple-variable model's delta value of 11. In predicting to the delinquency and psychopathology outcomes, the delta values are small, indicating that Model 5 did not differ significantly from the multiple-variable model. Also, the sensitivity and specificity values do not increase appreciably with the addition of multiple predictors. These findings replicate those reported by Bennett et al. (1999), albeit with a high-risk sample instead of a community sample and with a presumed base rate of 20% instead of their 15%.

Different Outcome Groups: Exploring Sex Differences and Varying Base Rates

In this section of results, we hold the screening variables constant and explore outcome groups created by using different cutpoints—that is, by varying assumptions about base rates. We use the Model 5 screen (parent and teacher ratings from first grade only) and explore its accuracy in predicting membership in outcome groups (a) divided by sex and (b) created by changing assumptions about base rates. We report results only for the parent–teacher externalizing outcome because of space limitations. However, the main findings

from these analyses (Models 5 and 6 being markedly superior to other models) apply to all three outcomes examined.

Sex differences—We divided the fifth grade externalizing outcome group (presented in Table 2) into boys (27%) and girls (12%). With PPV held constant at.50 (i.e., at a specificity of.80 for boys and.93 for girls), sensitivity for boys was.68 and for girls was.44. With sensitivity held constant at.50, specificity for boys was.93 and for girls was.89. In other words, test characteristics for boys but not for girls exceeded preset criteria. In Figure 1, we present a visual comparison of these test statistics using ROC curves. It can be seen that sensitivity and specificity for boys are higher than for girls and that preset criteria for accuracy (represented by points on the curve falling into the upper left quadrant) are met only for boys. The interaction of sex with risk status is significant for the teacher predictor of externalizing, $\chi^2 = 10.72$, $p < .001$, but not for the parent rating, $\chi^2 = 0.07$, *ns.*

Different base rates—We explored the effects of assuming a lower base rate in two ways: First, we created an outcome group of students who scored in the top 20% in both fourth grade and fifth grade (11% of the sample). Prediction to this persistent externalizing group exceeded preset criteria (sensitivity of.53 and specificity of.94 when PPV was held constant at.50, and specificity of.95 when sensitivity was held constant at.50). Next, we took a more conservative approach by assuming an annual externalizing base rate of only 15% and creating an outcome group who scored in that top 15% in both fourth and fifth grade (8% of the sample). Prediction to this more restricted persistent group also exceeded preset criteria (sensitivity of.60 and specificity of.96 with PPV held constant; specificity of.98 with sensitivity held constant at .50).

Comparing Prediction With Externalizing, Delinquency, and Psychopathology Outcomes

In Table 3, we present test characteristics for the Model 5 screen predicting to outcomes in three different domains. An alternative method of comparing models is to plot them onto an ROC curve for visual inspection. In Figure 2, we present a comparison of screening prediction to the fifth grade parent-teacher EP, the fourth and fifth grade self-reported delinquency outcome, and the fifth grade parent-child reported psychopathology outcome (DISC).

The practical significance of differences in screening accuracy using these different outcome measures is presented in the inset of Figure 2. The inset indicates the number of TP and FP classifications obtained when the decision threshold is set at a sensitivity of.50 (the line) for each outcome. Assuming the group consists of 1,000 children and the base rate is 20% for each outcome, 50% of those who eventually developed an externalizing outcome, or 100 children, would have been correctly identified (for each outcome) using this cutpoint. If our outcome of interest is the parent–teacher EP group, this decision threshold would result in 40 children being incorrectly classified by the screen as likely to develop EP. If we are predicting to SRD outcomes, we would have 80 FPs, and we would have 112 FPs in predicting to the DISC outcome with this screening procedure.

Effects of Setting Different Decision Thresholds on a ROC Curve

A further step in assessing accuracy involves calculating the effects of different decision thresholds, given a particular case mix of boys and girls and definition of outcome. In other words, criteria for accuracy should be anchored by consideration of the practical outcomes of using a screen (McFall & Treat, 1999; Swets et al., 2000). In Figure 3, we present the ROC curve for the fifth grade externalizing outcome (the same as Curve A from Figure 2). Four different points on the curve illustrate the implications of changing the TP rate by changing the cutpoint used to create the EP outcome group. The letter *B* represents the point

at which our test characteristics meet one of the criteria presented by Bennett et al. (1999), the cutpoint at which sensitivity equals 50%. The letter *C* represents the point on the curve at which test characteristics meet Bennett et al.'s (1999) other criterion: PPV of 50% when base rate is 20%. The letters *A* and *D* fall outside the range of the preset criteria and represent decision thresholds of 30% and 90% sensitivity. The inset presents numbers of TP and FP classifications (out of 1,000) for each of these decision points. As we move up the curve, both TP and FP classifications increase.

Relative Costs and Benefits of Different Decision Thresholds

Clearly, higher numbers of TP classifications and lower numbers of FPs are both desirable. How many TPs are enough, though, and how many FPs are too many to justify the expense of a targeted intervention? These questions can only be answered in the context of a consideration of the costs and benefits of such interventions. An in-depth discussion of this topic is beyond the scope of our article. However, for purposes of illustration only, we use program cost data presented by the WSIPP (1998; ranging from $2,000 to $14,000) and information on the societal cost of lifetime crime, substance abuse, and school failure from Cohen (1998; estimated at $1.7 to $2.3 million). In Table 4, we have increased the estimated program cost per child to $20,000 and set the societal cost per true case at $2 million. To examine effects of intervention on social costs of crime, we assume that “true noncases” (true negatives and FPs) have a 5% chance of school failure, lifetime crime, and substance abuse and that our measure of underlying risk is closely linked to these long-term outcomes; without intervention, the high-risk youth have a 30% chance of school failure, substance abuse, and crime. Finally, we assume that our intervention is quite effective and reduces the risk of these outcomes from 30% to 15% (Cohen, 1998). To determine overall intervention cost, we first calculated cost per child as the sum of the program costs and expected social costs for each category (TP, FP, true negative, and false negative), multiplied by the percentage of the sample falling into each category, at a sensitivity of.50 (see Table 4). Using the same assumptions, we also computed costs per child at each of the four decision thresholds presented in Figure 3. Increasing the number of FPs ultimately decreases the cost to society under model assumptions. Decision Point D, which falls outside the range of criteria for accuracy set by Bennett et al. (1999), results in a lower societal cost ($160,800) than Decision Points B ($173,600) and C ($164,160), which fall inside that range.

As this example shows, preset criteria for accuracy may not map onto the actual costs and benefits of delivering a targeted intervention. We emphasize that this exercise is intended only to illustrate the need for considering a screen's accuracy in the context of its utility and, although it is based on published data, does not represent a formal cost-effectiveness analysis. More sophisticated and accurate methods for determining costs and benefits of interventions for EP and of using different decision thresholds given information about base rates are presented in detail elsewhere (Cohen, 1998; Glascoe, Foster, & Wolraich, 1997; McFall & Treat, 1999; Swets et al., 2000).

Discussion

Our goal in this article was to examine factors that affect the accuracy and usefulness of screening for EP. Specifically, we examined the effects of different base rates, different case mixes, different times and predictors for the screen, and different outcomes. Second, we examined the practical import of using different outcomes and setting different decision thresholds for a single outcome.

The Question of Base Rates

We explored the relation of base rate to test characteristics by varying assumptions about base rate and by comparing various screening procedures. Data from multiple raters and multiple outcome measures in both fourth grade and fifth grade suggest that an annual base rate of approximately 20%, a higher base rate than has generally been reported, is a reasonable estimate for a normative elementary school population from high-risk environments. The base rate of those engaging in parent–teacher-rated EP persistently (for 2 years) was approximately 11%. Even assuming more conservative base rates on the parent–teacher outcome (15% annual and 8% persistent), screening procedures adequately predicted externalizing and delinquent behaviors up to 5 years after screening.

Timing of Screening

This study provides evidence that single-time-point, multiple-rater screening procedures may be the most effective and efficient in predicting externalizing outcomes. First-grade screening data (Model 5) were notably more effective than were kindergarten screens (Model 4) and effectively predicted EP and delinquent outcomes. This finding is consistent with results reported by Bennett and Offord (2001), who found higher screening accuracy for their 8- and 9-year-old cohort than for their 5- and 6-year-old cohort. Although intervention strategies have focused on identifying at-risk children as early as possible, first grade may be a better point for screening than kindergarten simply because the task demands of first grade resemble later school experience more closely. Malone, Coie, and Conduct Problems Prevention Research Group (2003), in an examination of trajectories of problem behavior over time, identified a group of children whose EP diminished rapidly after first grade and speculated that these children are simply slower to master the behaviors needed for effective adjustment to school. However, it is not clear whether it is age maturation or situational demands that most contribute to the greater predictive power of first-grade predictors.

Sample Composition and Generalizability

The base rate on all outcomes was between 2 and 4 times higher for boys than for girls: For the parent–teacher outcome, the base rate for boys was 2.3 times the rate for girls (16% vs. 7%); for diagnosed externalizing disorders, the rate was almost 3 times higher for boys (26% vs. 9%); and for self-reported delinquency, the rate was 4 times higher for boys (37% vs. 9%). Thus, the predictive value of the whole-sample models is attributable primarily to the boys. This may be because girls' overt externalizing behaviors, which tend to be less severe, are correspondingly less stable. Thus, girls “fall out” of a high-risk category at a greater frequency and are more easily misclassified. Different types of screening tests and procedures may be necessary for girls (Cote, Zoccolillo, Tremblay, Nagin, & Vitaro, 2001; Flanagan et al., 2002). Determination of whether a screening procedure is accurate enough for use with girls requires assessment of the costs and benefits of TPs and FPs at various decision thresholds.

Although values computed on a single sample are unlikely to generalize, the sample in this study was similar to that studied by Bennett and colleagues (Bennett et al., 1999; Bennett & Offord, 2001), and similar test characteristics for screening procedures were found across the two studies. Although both were normative samples and represented a full range of child behaviors, ours was drawn from higher risk environments, and this may explain why our procedures had higher sensitivity and specificity. Thus, it is feasible that the screening procedure reported here will provide relatively stable test values across other high-risk samples, given similar outcomes and base rates of externalizing disorders. Comparison with data reported by Bennett et al. (1999) indicates that including multiple child and family

measures does not add significantly to the predictive value of the screening and so represents an unnecessary cost.

Type and Timing of Outcome

We examined the accuracy of a screen in predicting to EP as rated by parents and teachers; to self-reported delinquent behaviors in multiple categories; and to diagnosed externalizing psychopathology reported by child, parent, or both 4–5 years after the initial screening. Patterns of outcomes across these different measures provide some guidance for estimating base rates, converging at about 20% for an annual rate, lower for a persistent 2-year rate. Test characteristics are higher for externalizing and self-reported delinquency behaviors than for diagnoses of psychopathology, as might be expected. Test characteristics are highest for ratings with shared method variance (i.e., teacher screen predicting to parent and teacher outcomes). However, teacher ratings alone have high predictive value for both parent–teacher externalizing and self-reported delinquency outcomes. This finding is consistent with that of Jones et al. (2002), who reported that parent ratings did not add significantly to prediction of service utilization. However, the parent report is important in predicting to parent–child reported psychopathogy.

Screening Utility

We used ROC curves to demonstrate the practical significance of using a first-grade parent–teacher screen to identify children with later EP in school, self-reported delinquent behaviors, and diagnosed psychopathology. We looked also at the effects of varying the decision threshold on the ROC curve for a single outcome. We recommended that screening accuracy be considered not in isolation but rather in light of these practical outcomes (i.e., TP and FP rates and their relative costs and benefits). To illustrate this point, we presented a rough calculation of the potential effects of setting different decision thresholds. The ratio of TP to FP rates of these different thresholds ranged from 60 TP and 8 FP to 180 TP and 280 FP. In our sample calculation, there were benefits to setting liberal decision thresholds, despite the increase in FPs. However, this calculation was for purposes of illustration only. It does not take into account important cost and benefit considerations of identifying (and misidentifying) children for a targeted intervention.

The indirect costs of the FP rate are of particular importance, and there is currently little evidence with which to calculate them. For example, we do not know the effects of stigmatization on those children mistakenly identified as having EP and receiving intervention. There is little research about the scope of stigmatization (e.g., whether teachers and peers respond differently to children in the FP group or whether participation in an intervention is noted in school records and whether that affects later opportunities) or about its individual-level effects (e.g., do children in the FP group have lower self-esteem or achievement than expected?). Another possible indirect cost of FPs is the potential for iatrogenic effects of intervention: There is some evidence that EP may increase, especially in those with lower levels of EP, when children are placed in intervention groups (Poulin, Dishion, & Burraston, 2001). However, in a long-term intervention where services are modified to fit individual need, FPs may be dropped from the intervention. In addition, FP rates may be inflated by the fact that some children with early EP will not show continuity but may nevertheless benefit in early years from receiving intervention. Finally, an indirect cost of high FP rates may include a dilution of the intensive nature of a targeted intervention.

Another important cost consideration is whether to use teacher-only or parent–teacher measures for screening. Here, parent– teacher models were statistically superior to teacher-only models. However, teacher-only models did have good predictive value for both

externalizing and delinquency outcomes. Benefit– cost analyses could verify whether the added rater is necessary, given the additional cost and time needed to collect parent data.

Conclusion

The screening procedures reported here, as well as those in other studies of EP, show much greater specificity than sensitivity. This indicates that the procedures may be especially useful for ruling out future cases of externalizing disorders. Also, the procedure is clearly more useful for screening boys, and the FP rate at more liberal cutpoints may include more FPs than TPs. These limitations should be weighed against the potential advantages of screening for an indicated intervention: The high sensitivity rates reported indicate that a majority of children who will develop serious EP can be identified at an early age. We believe the savings to society gained by preventing the cascading effects of long-term EP (Coie et al., 1995; Nagin & Tremblay, 1999) are greater than the costs of delivering a targeted intervention to those who do not need it. At this time, no evidence suggests that the costs of FPs in a targeted preventive intervention outweigh the benefits. Furthermore, existing cost-effectiveness data indicate that early intervention produces a net benefit in terms of court and victim costs (WSIPP, 1998). Research is needed to determine the effectiveness of interventions targeted to those factors related to specific outcomes for the range of children identified in screening procedures. The current analyses suggest that under fairly stringent assumptions, simple screening in first grade is effective in identifying children who will be identified by parents, teachers, mental health professionals, and themselves 3 to 4 years later as having EP. We recommend that those considering targeted interventions use a teacher or parent–teacher screen for EP in first grade or in both kindergarten and first grade to achieve maximum sensitivity, specificity, and PPV.

Acknowledgments

This work was supported by National Institute of Mental Health Grants R18 MH48043, R18 MH50951, R18 MH50952, and R18 MH50953. The Center for Substance Abuse Prevention and the National Institute on Drug Abuse also have provided support for Fast Track through a memorandum of agreement with the National Institute of Mental Health. This work was also supported in part by Department of Education Grant S184U30002 and National Institute of Mental Health Grants K05MH00797 and K05MH01027.

We are grateful for the close collaboration of the Durham public schools, the Metropolitan Nashville public schools, the Bellefonte area schools, the Tyrone area schools, the Mifflin County schools, the Highline public schools, and the Seattle public schools. We greatly appreciate the hard work and dedication of the many staff members who implemented the project, collected the evaluation data, and assisted with data management and analyses. Thanks also to Paul F. Griner and Thomas G. Power for helpful comments on early drafts of this article.

References

Achenbach, TM. Manual for the Child Behavior Checklist 4–18 and 1991 profile. Burlington: University of Vermont, Department of Psychiatry; 1991.

August GJ, Realmuto GM, Crosby RD, MacDonald AW III. Community-based multiple-gate screening of children at risk for conduct disorder. Journal of Abnormal Child Psychology. 1995; 23:521–544. [PubMed: 7560560]

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]

Bennett KJ, Lipman EL, Racine Y, Offord DR. Do measures of externalising behaviour in normal populations predict later outcome? Implications for targeted interventions to prevent conduct disorder. Journal of Child Psychology and Psychiatry. 1998; 39:1059–1070. [PubMed: 9844977]

Bennett KJ, Offord DR. Screening for conduct disorder: Does the predictive accuracy of conduct disorder symptoms improve with age? Journal of the American Academy of Child & Adolescent Psychiatry. 2001; 40:1418–1425. [PubMed: 11765287]

Bierman KL, Nix RL, Maples JJ, Murphy SA. Examining the use of clinical judgment in the context of an adaptive intervention design: The Fast Track program. 2003 Manuscript submitted for publication.

Black, E.; Panzer, RJ.; Mayewski, RJ.; Griner, PF. Characteristics of diagnostic tests and principles for their use in quantitative decision making. In: Panzer, R.; Black, ER.; Griner, PF., editors. Diagnostic strategies for common medical problems. Washington, DC: American College of Physicians; 1991. p. 1-16.

Burnham, KP.; Anderson, DR. Model selection and inference: A practical information-theoretic approach. New York: Springer-Verlag; 1998.

Cohen MA. The monetary value of saving a high-risk youth. Journal of Quantitative Criminology. 1998; 4:5–33.

Coie J, Miller-Johnson S, Maumary-Gremaud A, Lochman JE, Terry R, Hyman C. The impact of stable childhood rejection and aggressiveness on sustained adolescent disorder: The significance of multi-year screening. 2002 Manuscript submitted for publication.

Coie J, Terry R, Lenox K, Lochman J. Childhood peer rejection and aggression as predictors of stable patterns of adolescent disorder. Development and Psychopathology. 1995; 7:697–713.

Conduct Problems Prevention Research Group. A developmental and clinical model for the prevention of conduct disorder: The FAST Track program. Development and Psychopathology. 1992; 4:509–527.

Conduct Problems Prevention Research Group. The Fast Track experiment: Translating the developmental model into a prevention design. In: Kupersmidt, JB.; Dodge, KA., editors. Children's peer relations: From development to intervention. Washington DC: American Psychological Association; 2004. p. 181-208.

Cote S, Zoccolillo M, Tremblay RE, Nagin D, Vitaro F. Predicting girls' conduct disorder in adolescence from childhood trajectories of disruptive behaviors. Journal of the American Academy of Child & Adolescent Psychiatry. 2001; 40:678–684. [PubMed: 11392346]

Elliott DS, Huizinga D, Morse B. Self-reported violent offending: A descriptive analysis of juvenile violent offenders and their offending careers. Journal of Interpersonal Violence. 1986; 1:472–514.

Flanagan K, Bierman KL, Kam CM. Conduct Problems Prevention Research Group. Identifying at-risk children at school entry: The usefulness of multibehavioral problem profiles. 2002 Manuscript submitted for publication.

Glascoe FP, Foster EM, Wolraich ML. An economic analysis of developmental detection methods. Pediatrics. 1997; 99:830–837. [PubMed: 9164778]

Jones D, Dodge KA, Foster EM, Nix R. Conduct Problems Prevention Research Group. Early identification of children at risk for costly mental health service use. Prevention Science. 2002; 3:247–256. [PubMed: 12458763]

Kaplow JB, Curran PJ, Dodge KA. Conduct Problems Prevention Research Group. Child, parent, and peer predictors of early-onset substance use: A multisite longitudinal study. Journal of Abnormal Child Psychology. 2002; 30:199–216. [PubMed: 12041707]

Lochman JE. Conduct Problems Prevention Research Group. Screening of child behavior problems for prevention programs at school entry. Journal of Consulting and Clinical Psychology. 1995; 63:549–559. [PubMed: 7673532]

Loeber R. Development and risk factors of juvenile antisocial behavior and delinquency. Clinical Psychology Review. 1990; 10:1–41.

Malone PS, Coie JD. Conduct Problems Prevention Research Group. Are there multiple pathways in the emergence of serious conduct problems?. 2003 Manuscript in preparation.

McFall RM, Treat TA. Quantifying the information value of clinical assessments with signal detection theory. Annual Review of Psychology. 1999; 50:215–241.

Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin. 1955; 52:194–216. [PubMed: 14371890]

Mrazek, PJ.; Haggerty, RJ. National Academy of Sciences Institute of Medicine, Division of Biobehavioral Sciences and Mental Disorders, Committee on Prevention of Mental Disorders. Reducing risks for mental disorders: Frontiers for preventive intervention research. Washington, DC: National Academy Press; 1994.

Nagelkerke NJD. A note on the general definition of the coefficient of determination. Biometrika. 1991; 78:691–692.

Nagin D, Tremblay RE. Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. Child Development. 1999; 70:1181–1196. [PubMed: 10546339]

Nix, RL. Child Behavior Checklist (Technical report). 2001. Retrieved December 8, 2002 from http://www.fasttrackproject.org/techrept/c/cbc/

Office of the Surgeon General. Youth violence: A report of the Surgeon General. 2001. Retrieved May 28, 2003 from the U.S. Surgeon General's Web site: http://www.surgeongeneral.gov/library/youthviolence/youvioreport.htm

Offord DR. Selection levels of prevention. Addictive Behaviors. 2000; 25:833–842. [PubMed: 11125774]

Offord DR, Boyle MH, Racine Y, Szatmari P, Fleming JE, Sanford M, Lipman EL. Integrating assessment data from multiple informants. Journal of the American Academy of Child & Adolescent Psychiatry. 1996; 35:1078–1085. [PubMed: 8755805]

Poulin F, Dishion TJ, Burraston B. 3-year iatrogenic effects associated with aggregating high-risk adolescents in cognitive–behavioral preventive interventions. Applied Developmental Science. 2001; 5:214–224.

Swets JA, Dawes RM, Monahan J. Psychological science can improve diagnostic decisions. Psychological Science in the Public Interest. 2000; 1:1–26.

Tremblay, RE.; LeMarquand, D.; Vitaro, F. The prevention of oppositional defiant disorder and conduct disorder. In: Quay, HC.; Hogan, AE., editors. Handbook of disruptive behavior disorders. Dordrecht, the Netherlands: Kluwer Academic Publishers; 1999. p. 525-555.

Tremblay RE, Pihl RO, Vitaro F, Dobkin PL. Predicting early onset of male antisocial behavior from preschool behavior. Archives of General Psychiatry. 1994; 51:732–739. [PubMed: 8080350]

Walker HM, Severson HH, Stiller B, Williams G, Haring NG, Shim MR, Todis B. Systematic screening of pupils in the elementary age range at risk for behavior disorders: Development and trail testing of a multiple gating model. Remedial and Special Education. 1988; 9(3):8–14.

Washington State Institute for Public Policy. Watching the bottom line: Cost effective interventions for reducing crime in Washington. 1998. Retrieved July 23, 2003 from http://www.wa.gov/wsipp/crime/pdf/bline.pdf

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]

Zoccolillo M, Tremblay R, Vitaro F. *DSM–III–R* and *DSM– III* criteria for conduct disorder in preadolescent girls: Specific but insensitive. Journal of the American Academy of Child & Adolescent Psychiatry. 1996; 35:461–470. [PubMed: 8919708]

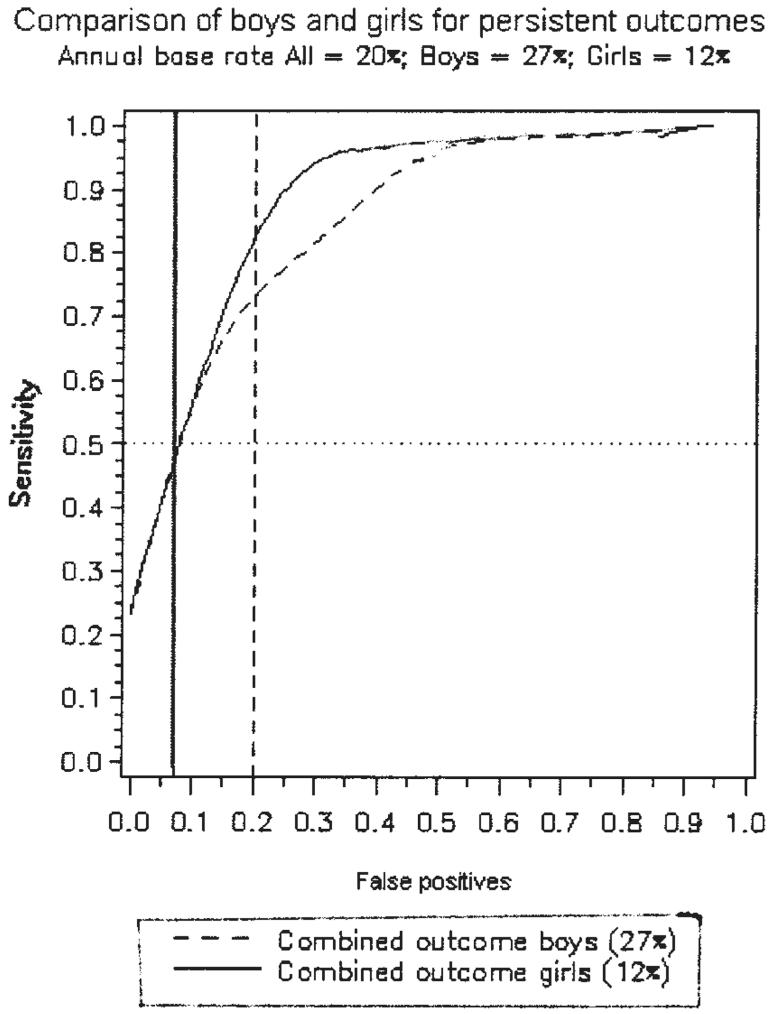

Figure 1.

Test characteristics for boys and girls at 20% overall base rate (12% of girls and 27% of boys classified as at risk). The horizontal line indicates the point at which sensitivity is.50. The dotted vertical line indicates a specificity for boys of.80 (the point at which positive predictive value is equal to or greater than.50 when the base rate is 27%). The solid vertical line indicates a specificity for girls of.93 (the point at which positive predictive value is equal to or greater than.50 when the base rate is 12%). The portion of the curve in the upper left quadrant (for boys) represents decision thresholds that meet the criteria for accuracy set by Bennett et al. (1998). No portion of the girls' curve is in the girls' upper left quadrant.

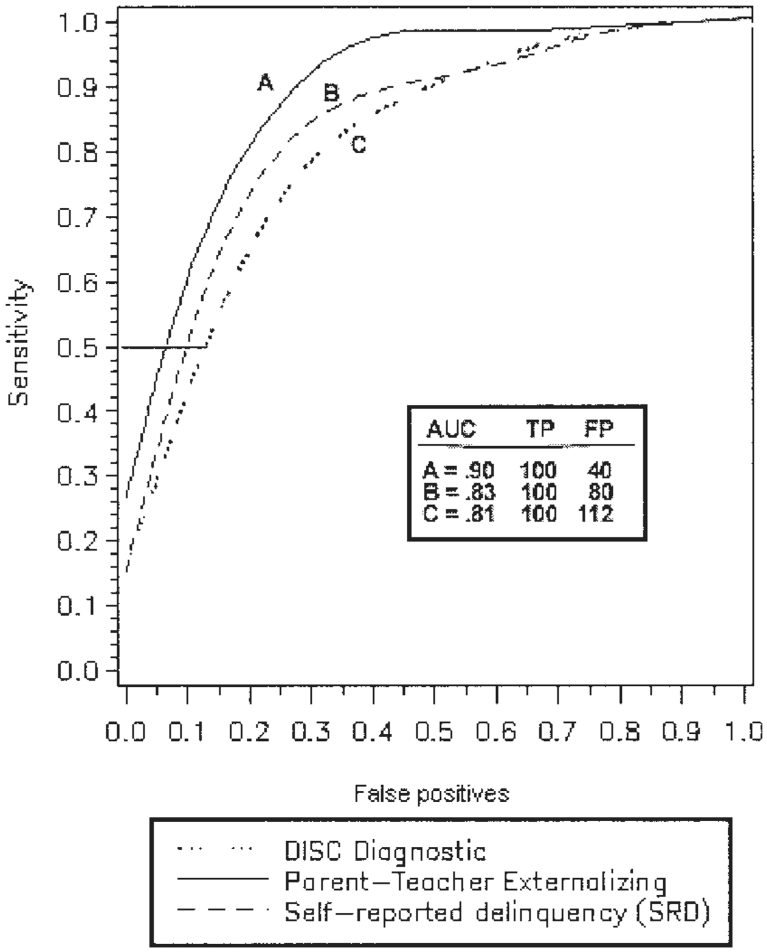

Figure 2.

True positive (TP) and false positive (FP) rates and area under the curve (AUC) at 50% sensitivity for three outcomes: (A) Parent-Teacher Externalizing, (B) Self-Reported Delinquency, and (C) Diagnostic Interview Schedule for Children (DISC). An AUC of.50 indicates classification accuracy equal to chance; an AUC of 1 signifies perfect classification. Data in the figure inset indicate numbers of false positives for each measure when the number of true positives is 100.

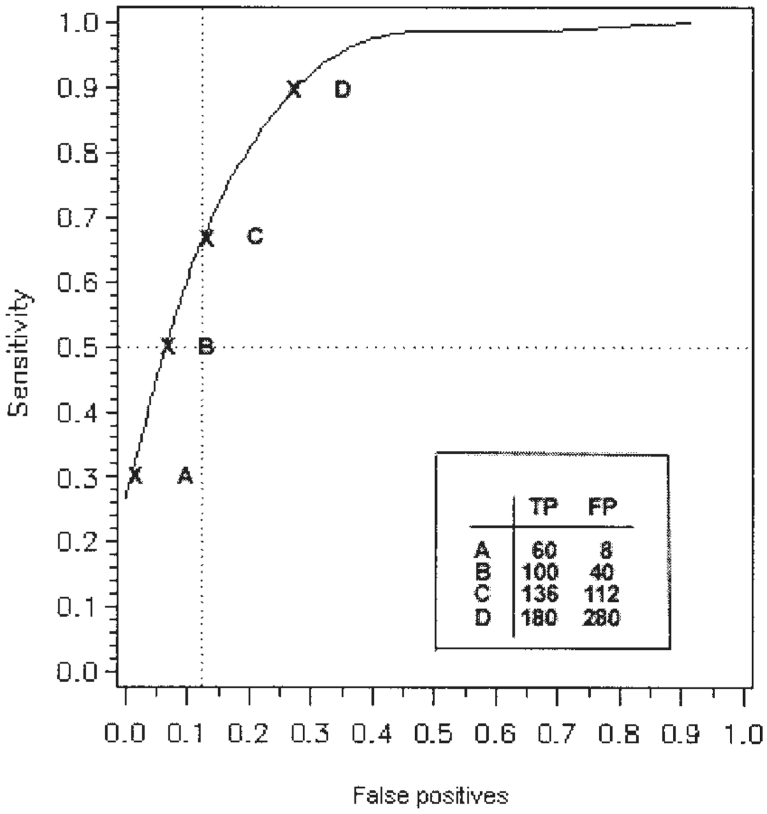

Figure 3.

Comparison of true positive (TP) and false positive (FP) rates at four different sensitivity thresholds for the fifth grade parent–teacher externalizing outcome. Four different points on the curve illustrate the implications of changing the true positive rate by changing the cutpoint used to create the externalizing problems outcome group. The letter B represents the point at which our test characteristics meet Bennett et al.'s (1999) criterion of the cutpoint at which sensitivity equals 50%. The letter C represents the point on the curve at which test characteristics meet Bennett et al.'s (1999) other criterion of PPV of 50% when base rate is 20%. The letters A and D fall outside the range of the preset criteria and represent decision thresholds of 30% and 90% sensitivity, respectively. The inset presents numbers of TP and FP classifications (out of 1,000) for each of these decision points. Moving up the curve, both TP and FP classifications increase. The dotted horizontal line indicates sensitivity of.50. The dotted vertical line indicates specificity of.875 for a PPV of 50% when the base rate = 20%.

Table 1

Means, Cutoff Scores, and Relations Among Predictors and Outcomes

Predictor	1	2	3	4	5	6	7	8	*M*	*SD*
1. TOCA K	—	.57	.21	.31	.47	.50	.27	.28	16.65	11.61
2. TOCA 1st		—	.23	.32	.58	.55	.37	.35	15.72	12.77
3. CPB K[a]			—	.46	.48	.48	.44	.42	2.13	0.44
4. RPBC 1st[a]				—	.36	.35	.70	.67	7.45	5.74
5. SHP 4th					—	.59	.38	.39	1.09	0.97
6. SHP 5th						—	.40	.40	1.08	0.98
7. CBCL 4th							—	.78	50.88	11.44
8. CBCL 5th								—	50.07	11.55
Cutoff 20%	26	26	60[b]	12	1.9	1.8	59[b]	59[b]		
Girls	22	20	57[b]	10	1.4	1.5	58[b]	57[b]		
Boys	28	29	63[b]	14	2.2	2.2	63[b]	62[b]		
Cutoff 15%	28	28	62[b]	14	2.1	2.2	63[b]	62[b]		
Girls	25	24	59[b]	11	1.7	1.6	59[b]	59[b]		
Boys	30	33	65[b]	16	2.5	2.6	66[b]	64[b]		

Note. All correlations are significant at $p \leq .001$. TOCA = Teacher Observation of Classroom Adaptation— Revised; K = kindergarten; 1st = first grade; CPB = Child Problem Behavior Scale; RPBC = Revised Problem Behavior Checklist; SHP = Social Health Profile; 4th = fourth grade; 5th = fifth grade; CBCL = Child Behavior Checklist.

[a] Parent-rated externalizing problems was scaled differently in kindergarten and first grade.

[b] Scores reported are *t* scores.

Table 2

Different Screening Models Predicting to Parent-Teacher Fifth Grade Externalizing Outcome: Base Rate = 20%

						PPV >.50		Sensitivity >.50	
Model no.	**Predictor**	**Likelihood ratio χ^2**	**R^2**	**AIC**	**Δ**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
1	**Teacher K***	47.1	.15	240	47	.45	.88	.52	.86
2	**Teacher 1st***	66.4	.21	221	28	**.50**	**.88**	**.50**	**.88**
3	**Teacher K*** & **Teacher 1st***	74.7	.23	214	21	**.57**	**.88**	**.50**	**.91**
4	**Teacher K*** & **Parent K***	59.2	.19	230	37	.48	.88	.50	.87
5	**Teacher 1st*** & **Parent 1st***	93.6	.28	195	2	**.64**	**.88**	**.52**	**.95**
6	Teacher K & **Teacher 1st*** and Parent K & **Parent 1st**	99.9	.30	193	0	**.68**	**.88**	**.50**	**.94**

Note. Predictors in bold had significant individual χ^2 values. Test values in bold surpass preset criteria. Models with a delta greater than 4 are notably inferior to the best model (Δ = 0); models with a delta greater than 10 fail to explain a substantial amount of variation in the model. PPV = positive predictive value; AIC = Akaike's information criterion; K = kindergarten; 1st = first grade.

*p <.001.

J Consult Clin Psychol. Author manuscript; available in PMC 2012 July 13.

Table 3

Comparison of Screening Model 5 With Multiple Predictor Screening Models for Three Outcomes

			PPV >.50		Sensitivity >.50	
Outcome and predictor	R^2	**Δ**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
Parent–teacher fifth grade externalizing (20%)						
Model 5 (teacher & parent first grade externalizing)	.32	0	**.64**	**.88**	**.52**	**.95**
Model 5 plus Bennett et al. (1999) variables[a]	.33	11	**.62**	**.88**	**.50**	**.93**
Self-reported delinquency (22%)						
Model 5	.23	5	**.60**	**.86**	**.50**	**.90**
Model 5 plus Bennett et al. (1999) variables	.31	0	**.58**	**.86**	**.50**	**.90**
DISC externalizing diagnosis (18%)						
Model 5	.18	4	.41	.89	.51	.86
Model 5 plus Bennett et al. (1999) variables	.23	0	**.53**	**.89**	**.51**	**.91**

Note. Test values in bold surpass preset criteria. Models with a delta greater than 4 are notably inferior to the best model (Δ = 0); models with a delta greater than 10 fail to explain a substantial amount of variation in the model. PPV = positive predictive value. DISC = Diagnostic Interview Schedule for Children.

[a]The variables were socioeconomic status; maternal education, maternal depression, parenting practices, attention-deficit/hyperactivity disorder, and parent satisfaction.

Table 4

Social and Program Costs per Child When Sensitivity Threshold for Externalizing Problems Is Set at.50

True condition (%)	Screened as (%)		% of sample	Program cost ($)	Chance of becoming lifetime criminal (%)	Costs of a lifetime of crime ($)	Expected social costs ($)	Expected social costs + program cost ($)
Positive (20)	TP	50	10	20,000	15	2,000,000	300,000	320,000
	FN	50	10	0	30	2,000,000	600,000	600,000
Negative (80)	TN	90	72	0	5	2,000,000	100,000	100,000
	FP	10	8	20,000	5	2,000,000	100,000	120,000

Note. Program cost is based on data from Washington State Institute for Public Policy (1998). Chance of becoming a lifetime criminal and costs of a lifetime of crime is based on data from Cohen (1998). Expected social costs = chance of becoming lifetime criminal × costs of a lifetime of crime. The expected cost of society per child (Expected social and program costs × % of sample, summed across all four groups) is $173,600. TP = true positive; FN = false negative; TN = true negative; FP = false positive.