Journal of Abnormal Child Psychology
https://doi.org/10.1007/s10802-018-0478-7

Using Early Childhood Behavior Problems to Predict Adult Convictions

Francesca Kassing[1] · Jennifer Godwin[2] · John E. Lochman[1] · John D. Coie[2] · Conduct Problems Prevention Research Group

Abstract
The current study examined whether teacher and parent ratings of externalizing behavior during kindergarten and 1st grade accurately predicted the presence of adult convictions by age 25. Data were collected as part of the Fast Track Project. Schools were identified based on poverty and crime rates in four locations: Durham, NC, Nashville, TN, Seattle, WA, and rural, central PA. Teacher and parent screening measures of externalizing behavior were collected at the end of kindergarten and 1st grade. ROC curves were used to visually depict the tradeoff between sensitivity and specificity and best model fit was determined. Five of the six combinations of screen scores across time points and raters met both the specificity and sensitivity cutoffs for a well-performing screening tool. When data were examined within each site separately, screen scores performed better in sites with high base rates and models including single teacher screens accurately predicted convictions. Similarly, screen scores performed better and could be used more parsimoniously for males, but not females (whose base rates were lower in this sample). Overall, results indicated that early elementary screens for conduct problems perform remarkably well when predicting criminal convictions 20 years later. However, because of variations in base rates, screens operated differently by gender and location. The results indicated that for populations with high base rates, convictions can be accurately predicted with as little as one teacher screen taken during kindergarten or 1st grade, increasing the cost-effectiveness of preventative interventions.

Keywords Screening · Preventative intervention · Convictions · Base rates

Research has consistently identified early behavior problems (e.g., aggression, disruptiveness) as risk factors for persistent, chronic patterns of violent and delinquent behavior later in life (e.g., Okado and Bierman 2014). These early behaviors are also associated with negative outcomes in academic, work, and social contexts, as well as poor physical and mental health outcomes later in life (Kaplow, Curran, Dodge, & Conduct Problems Prevention Research Group [CPPRG] 2002). Children's problem behavior occurs within the context of a variety of risk factors that may be targeted through preventative intervention (e.g., impulse control), but other risk factors may remain post-intervention (e.g., poverty status, neighborhood level of deviant behavior; Matthys and Lochman 2010; Petras et al. 2004b). Preventative interventions targeting early displays of aggression seek to disrupt this trajectory before behavior becomes more severe and persistent into adulthood (CPPRG 1992; Pasalich, Witkiewitz, McMahon, Pinderhughes, & CPPRG 2015). Such interventions are considered targeted prevention (Coie et al. 1993) and may be more cost-effective than interventions delivered to entire classrooms or schools. However, accurate and reliable screening tools are required to correctly identify at-risk children and provide them with early intervention (Jones, Dodge, Foster, Nix, & CPPRG 2002).

The Utility of Early Childhood Screens

Common screening procedures for targeted prevention often include brief surveys administered to caregivers and teachers to identify children who may benefit from the intervention based on their early behavior problems (CPPRG 1999; Petras et al. 2004b). Research has shown that combined parent and teacher screening measures of behavior problems during kindergarten predict higher levels of delinquency (Hill, Lochman, Coie, Greenberg & CPPRG 2004) and lower levels

✉ Francesca Kassing
fkassing@crimson.ua.edu

[1] Department of Psychology, The University of Alabama, Box 870348, Tuscaloosa, AL 35487, USA

[2] Center for Child and Family Policy, Duke University, Box 90545, Durham, NC 27708, USA

Published online: 03 October 2018

of social competence (Lochman & CPPRG 1995) later in childhood. In addition, parent and teacher completed screens of aggression and hyperactivity in preschool have been shown to predict behavior problems 5 years later (Stormont 2000). Based on this research, it appears that parent and teacher screens taken as early as preschool and kindergarten can accurately predict externalizing problems later in childhood.

In considering the predictive quality of early childhood screens, research has begun to focus on objective measures of negative outcomes, such as involvement in the justice system. Assessing criminal involvement is particularly important when considering targeted prevention, given that juvenile crime in urban areas is estimated to cost $89–$110 million (Welsh et al. 2008). Research has demonstrated that teacher and parent measures of children's conduct problems during kindergarten predict utilization of mental health, special education, and juvenile justice systems (Jones et al. 2002; 2015). Furthermore, screens taken during late elementary school have been shown to accurately predict adult violent behavior and convictions (Petras et al. 2004a).

Despite the above-cited research indicating that measures of childhood behavior problems longitudinally predict criminal behavior, no study to date has specifically assessed the utility and cost-effectiveness of early childhood screens in predicting objective measures of adult delinquency (e.g., convictions). Therefore, the first primary aim of the current study was to address this gap by examining whether teacher and parent ratings taken during kindergarten and 1st grade accurately predicted the presence of a conviction in early adulthood and thus represent viable screening procedures for targeted prevention.

Characteristics and Considerations of Effective Screening

Cost-effectiveness and efficiency are key components of targeted preventions for externalizing behavior. If the screening procedures of a targeted intervention can accurately predict future cases of a behavior or disorder, the resulting intervention can be efficient, focused, and effective (Lochman & CPPRG 1995). To assess the accuracy of screening procedures, the following statistical tests are often used: *sensitivity* (the proportion of individuals correctly identified as having the predicted outcome), *specificity* (the proportion of individuals correctly identified as *not* having the predicted outcome), *positive predictive value* or PPV (the proportion of those who were classified as at-risk who develop the predicted outcome) and *negative predictive value* or NPV (the proportion of those who were classified as at-risk who did not develop the predicted outcome). As the cutoff score used to determine inclusion in the intervention decreases, there are fewer false negatives and therefore greater sensitivity; however, there are also more false positives and therefore lower specificity. Greater sensitivity is particularly important when failing to classify a true positive as at-risk has large costs. As discussed earlier, aggression and delinquency carry large costs; therefore, failing to classify someone who needs intervention as at-risk could incur high future costs to society due to that person's future delinquency. Therefore, in the case of the Fast Track intervention (the program from which the current study's sample was collected), sensitivity was prioritized.

Consideration of Base Rates

Base rates of externalizing problems are particularly important when considering the efficiency and effectiveness of early childhood screens (Hill et al. 2004). In fact, the base rate of the predicted outcome in a sample has a significant effect on the PPV and NPV of a screen (Meehl and Rosen 1955). Populations with a *higher* prevalence of the disorder will inevitably have more individuals correctly classified as having the disorder (increased PPV) but fewer individuals correctly classified as not having the disorder (decreased NPV) while populations with *lower* base rates of the disorder will have more individuals incorrectly classified as having the disorder (decreased PPV) and more individuals correctly classified as not having the disorder (increased NPV; Elwood 1993). Therefore, the PPV and NPV will vary depending on the base rate of the population, affecting the overall effectiveness of the screens delivered.

For this study, we were particularly interested in how base rates varied as a function of demographic variables (e.g., gender) and location. Base rates of externalizing problems have been shown to be higher for boys than for girls, which in turn may affect the PPV and NPV in samples that include both genders (Zoccolillo et al. 1996). In fact, screens have been found to be more accurate in predicting later aggressive behavior in boys compared to girls, given their higher base rates of behavior problems (e.g., Hill et al. 2004). Similarly, base rates of externalizing behavior may differ across types of communities. For example, higher rates of conduct problems have been found in urban compared to rural areas (Wichstrøm et al. 1996). Therefore, when considering effective screening, it is important to consider the base rate of the target disorder in the specific community being studied.

Considering base rates in the assessment of psychological disorders is not a new phenomenon and dates back to Paul Meehl's work as early as the 1950s (e.g., Meehl and Rosen 1955). However, despite clear evidence that base rates affect the classification of disorders, clinical research has largely ignored base rates in assessment practices (Elwood 1993). Therefore, the second primary aim of this study was to assess the effects of base rates on screening procedures in the current sample to inform the consideration of sample prevalence in screen development and utilization.

The Assessment of Multiple Settings and Time Points

When determining the cost-effectiveness of screening procedures for targeted prevention, it is important to consider whether accurate prediction can be achieved with fewer screens. This is complicated, however, by research consistently demonstrating low-to-moderate cross-informant correspondence when assessing children's mental health (De Los Reyes et al. 2015) and differences among studies assessing the efficacy of early childhood screens across informants. For example, research on the utility of early childhood screens cited above relied on ratings from both parents and teachers to predict later patterns of externalizing behavior problems (Hill et al. 2004; Stormont 2000). However, other research has found that when parent and teacher ratings taken during kindergarten are considered within the context of ten early risk-factors (including demographic variables, recent life stress, and neighborhood quality), parent ratings of conduct problems within the home setting, but not teacher ratings of conduct problems within the school setting, are effective in predicting the number of serious arrests in adulthood (Jones et al. 2015). Furthermore, among a sample of inner-city children, teacher-reported/school-based, but not parent-reported/home-based, oppositionality predicted later conduct disorder symptoms after controlling for initial levels of co-occurring symptoms (Drabick et al. 2011). Despite poor agreement between and differential predictability of teacher and parent ratings, research has also suggested that early reports of oppositionality show similar symptom profiles, risk factors, comorbidities, and outcomes across reporters (McNeilis et al. 2017). It is also possible that the poor agreement between teacher and parent ratings reflects unique displays of behavior problems in home versus school settings (e.g., Achenbach 1982). For example, teachers, more than parents, can observe how children interact with peers and their ability to meet demands at school. Parents, on the other hand, can observe a wider range of behaviors, including behaviors that are more severe (e.g., fire setting, being cruel to animals), that may be more evident in the home setting.

Therefore, while previous research has implied that more parsimonious screening procedures (i.e., single-rater) may be accurate in predicting externalizing behavior longitudinally, it has not directly measured the accuracy of single-rater, single-time point screens in predicting adult crime. Therefore, a secondary goal of the current study was to compare different combinations of screens taken at multiple time points (i.e., kindergarten and 1st grade) and in multiple settings (i.e., parents reporting on the problem behaviors they observe in the home setting and teachers reporting on the problem behaviors they observe the school setting) using measures of sensitivity and specificity to determine the most cost-effective, but accurate screening procedures. Furthermore, to improve the clinical utility of our findings, we also sought to determine how two sets of adult caretakers can predict future serious antisocial behavior (i.e., adult convictions) using measures that are commonly used in the field to identify children with problem behavior.

The Current Study

The current study had two primary aims and one secondary goal. The first aim was to assess the accuracy of teacher and parent screens taken during kindergarten and 1st grade in predicting criminal behavior in adulthood (i.e., having at least one adult conviction by age 25). Given previous research demonstrating the effectiveness of early childhood screens in predicting aggressive behavior later in childhood (e.g., Hill et al. 2004; Stormont 2000) it was hypothesized that these screening procedures would similarly predict criminal outcomes in early adulthood. The second aim was to assess the effects of varying base rates of convictions on the accuracy of the early childhood screens. Specifically, we were interested in how the success of screening measures varied by demographic variables (i.e., gender) and site location. Consistent with research cited above, it was expected that screening procedures would more accurately predict having an adult conviction by age 25 in populations with higher conviction base rates. We expected to find higher conviction base rates among males relative to females and in urban locations relative to rural sites; therefore, we hypothesized that screening procedures would perform better for males and in urban sites. A secondary goal of the study was to determine the most efficient and cost-effective method of early childhood screening by examining several combinations of parent and teacher screens during two subsequent years. There was insufficient research in this area to warrant directional hypotheses; therefore, these analyses were considered more exploratory in nature. Demonstrating effective screening for single-rater/setting or single-time point combinations would support the use of more parsimonious and economical screening procedures.

Method

Sample

Data were collected as part of the Fast Track Project which was designed to prevent the development of externalizing problems and has been extensively described elsewhere (e.g., CPPRG 1992; 2000). Participants were selected based on a three-stage process. First, schools were evaluated in the four locations: Durham, North Carolina; Nashville, Tennessee; Seattle, Washington; and rural, central Pennsylvania. Schools located in areas with high crime rates were chosen based on high poverty rates and low education

rates among parents of school-age children. Thirteen elementary schools in Durham, 10 in Nashville, 15 in Seattle, and 17 in rural Pennsylvania were identified. A slightly larger potential risk-population was identified at the rural Pennsylvania site (i.e., more schools) in anticipation that there would be lower conduct disorder risk rates at this site compared to the other urban sites.

Next, schools were matched based on poverty rates, racial makeup, and size and then randomly assigned to control or intervention conditions. This study uses a normative sample of 387 children recruited from schools, assigned to the control condition, and therefore unaffected by the intervention. All kindergarten children at these control schools were stratified based on race, sex, and decile of the teacher-reported screen of problem behaviors (described in the "Screening Variables" section). Next, participants were randomly chosen from race and sex groups within each decile of teacher screens to create a normative sample that was representative of the school's population. This procedure ensured that the normative sample represented the population in the high-risk schools.

This normative sample received the teacher and parent screens, as well as ongoing Fast Track assessments. The normative sample was 43% African American, 52% European American, and 5% other ethnic backgrounds. Fifty-one percent were male.

Screening Variables

The screening measures used in this study reflect the initial rationale for screen selection during the initial phases of the Fast Track project (e.g., CPPRG 1992; Lochman & CPPRG 1995). More specifically, the measures were chosen to represent commonly used screening measures assessing home- and school-based externalizing behaviors. These measures were designed to consider the perspectives of two different categories of adults who would have good access to children's problem behavior in two unique settings (i.e., at home versus at school).

Screening measures were collected at the end of kindergarten and 1st grade. While the ratings of a child were typically completed by the same primary caretaker in both years, the teacher ratings were completed by different teachers (kindergarten teacher vs. 1st grade teacher).

Teacher Screen Teacher screens were composed of 14 behavioral items from the Teacher Observation of Classroom Adaptation-Revised (TOCA-R) measure evaluating externalizing behavior in the classroom (items can be found at http://fasttrackproject.org/techrept/s/shf/shf1tech.pdf; Werthamer-Larsson et al. 1991). The screen included items, such as "stubborn," "breaks rules," and "harms others," which were rated on a scale of 0 (*almost never)* to 5 (*almost always*). Internal consistency was high in kindergarten ($\propto = 0.91$) and 1st grade ($\propto = 0.93$). Ten items on the teacher screen overlapped with items on the parent screen and sought to assess externalizing behaviors present it the classroom setting (e.g., disobedient, fights, breaks things). The remaining four items assessed other areas of impaired school functioning that teachers may observe in the classroom setting (i.e., on-task behavior, friendliness, self-reliance, and ability to complete assignments).

Parent Screen Parents completed the Child Problem Behavior Scale during the summer after kindergarten (items can be found at http://fasttrackproject.org/techrept/c/cpb/cpb1tech.pdf). The measure was composed of 24 items drawn from the Child Behavior Checklist (CBCL; Achenbach 1991; Nix 2001) and the Revised Problem Behavior Checklist (RPBC), which was also derived from the CBCL. Home-based externalizing and oppositional behaviors, such as "temper tantrums" and "argues" were rated on a scale from 1 (*never*) to 4 (*often*). This scale had high internal consistency ($\propto = 0.87$). Ten of the twenty-four items on the kindergarten parent screen directly corresponded to items on the teacher screen; however, the parent screen had an additional twelve items that assessed a wider range of conduct problems, including some more severe behaviors (e.g., "cruel to animals," "sets fires"). First grade behavior was then rated on 21 items taken from the CBCL, 17 of which were identical to items on the kindergarten parent screen. The remaining 4 items on the 1st grade screen assessed similar behaviors to that of the kindergarten parent screen (i.e., "threatens people," "destroys own things," 'disobedient at home," and "steals at home"). Items on the 1st grade parent screen were rated on a scale from 0 (*never)* to 2 (*often)*. This scale also had high internal consistency ($\propto = 0.88$).

Outcome Variable

Any Conviction by Age 25 Adult conviction data were collected from local court records and national databases based on name, date of birth, and social security number. Conviction was coded as missing if arrest records provided insufficient information to determine the outcome of the arrest or whether formal charges were pursued. Convictions for status offenses (i.e., minors in possession of alcohol/tobacco, runaway, truancy), probation violations, and minor traffic violations were excluded. All other convictions were included, such as public order violations (i.e., public consumption, loitering, and gambling; 19% of convictions), driving under the influence of alcohol/drugs (4% of convictions), drug possession and sale (21% of convictions), property crimes (28% of convictions), and violent crimes (28% of convictions). Based on data collected over the entire study period, 55% of the current sample were identified as "abstainers" (indicating that they did not have any convictions through age 25), 8% of the current

sample were identified as "adolescent-limited" (indicating that they only had convictions prior to age 19), 17% of the current sample were identified as "adult starters" (indicating that they only had convictions after age 18) and 20% of the current sample were identified as "life course persistent" (indicating that they had at least one conviction prior to age 19 and at least one conviction after age 18; Goulter, Godwin, & CPPRG 2018).

Any Conviction was examined as a dichotomous variable (0 indicating no convictions and 1 indicating one or more adult convictions by age 25). Follow-up analyses also examined a dichotomous variable capturing Multiple Convictions (0 indicating none or 1 conviction and 1 indicating two or more adult convictions by age 25).

Attrition and Missing Data

Participants who were missing any screen score or outcome variable were removed from the sample so that a consistent sample was used to evaluate the effectiveness of different combinations of screen scores across time and raters. Forty-five potential participants (11.6% of the original 387 sample) were missing 1 or more screen scores and eight of these were missing conviction data, yielding a sample of 334 for this study. Chi-square analyses revealed that Seattle was missing significantly more screen scores compared to the other sites; however, missing and non-missing samples did not differ by race or gender.

Analytic Strategy

Sensitivity and specificity provide information regarding whether the screening instrument is making useful predictions about future outcomes (in this case, criminal behavior). Cutoffs for screen scores must be set in order to assign at-risk status to potential participants; however, there is a tradeoff between sensitivity and specificity when choosing a cutoff. Based on a review of longitudinal studies examining sensitivity, it is unlikely that more than 50% of children exhibiting externalizing behavior at age 4/5 will develop persistent conduct problems later (Bennett et al. 1998). Fifty percent of conduct problem youth will develop antisocial behavior in adolescence and adulthood, leading to a rule-of-thumb for minimal sensitivity of 50% in screening research (Bennett and Offord 2001). Bennett, Offord and colleagues (1999; 2001) have recommended that Positive Predictive Value (PPV) be set at 50% as well as 50% for sensitivity in assessments of screening tools for targeted prevention programs. With these preset criteria, at least half of those children needing intervention would receive it (50% sensitivity), and at least half of the children designated as "at-risk" will in fact need the intervention (50% PPV). PPV is linked to different levels of specificity depending on the base rate of the outcome; for example, when the base rate is low, then specificity can be quite high, and specificity declines as the base rate increases. To determine the required specificity level, the values of sensitivity and PPV (both equal to 0.5) are substituted into the equation defining PPV: specificity = 1–1.5*prevalence.

Bennett and Offord (2001) have noted that screening may not be justified in normative samples with low base rates (e.g., 5–10%), because screening tools will not meet the preset criteria. However, with higher base rates (e.g., 20%) in normative elementary school populations from high-risk environments, early school screening has met the preset criteria of 50% sensitivity and 50% PPV (Hill et al. 2004). The normative school sample for the present study, representative of the recruited schools, is the same as the one examined in Hill et al. (2004), but looking at later criminal outcomes. The 50% rule-of-thumb for sensitivity assessments of screening tools has continued to be used for studies of screening accuracy (e.g., Sawyer et al. 2014). An additional benefit for using these preset criteria of 50% for sensitivity and PPV is that they provide a convenient frame of reference for comparing accuracy across other samples (Hill et al. 2004), as were done in the site comparisons in the present study.

To assess the accuracy of different combinations of screening measures, six logistic regressions predicting the probability of a conviction were estimated. All models included dummy variables capturing site, given that statistically significant differences were found by site among those with and without missing data for the screening variables. The models varied based on which screen scores were included: (1) teacher screen from kindergarten only, (2) teacher screen from 1st grade only, (3) teacher screen from kindergarten and 1st grade, (4) teacher and parent screen from kindergarten, (5) teacher and parent screen from 1st grade, and (6) all four screen scores. The different sensitivity and specificity values were then calculated from the results. ROC curves were used to visually depict the tradeoff between sensitivity (proportion of true positives captured by the screen) and 1 minus specificity (proportion of false positives captured by the screen) for each screening cutoff point. The diagonal line represents combinations of sensitivity and (1-specificity) for which the screening measures did not distinguish between those with and without the outcome. ROC curves that are further away from the diagonal line (and therefore those with greater area under the curve) represent more accurate screening measures. ROC curves visually display the screening measure cutoffs that meet preset sensitivity and PPV requirements.

Following Hill et al. (2004), comparisons of model fit across different specifications were facilitated by the Akaike's information criterion (AIC; Burnham and Anderson 2004). AIC measures goodness of fit using the likelihood of the model adjusted for the number of parameters included in the model. The best model fit is defined by the lowest AIC.

Delta is calculated for each comparison model by subtracting the minimum AIC among the 6 models from the comparison model AIC. Models with delta less than or equal to 2 are considered to have equivalent fit. Models with delta greater than or equal to 4 have substandard fit (Bennett et al. 1999). Nagelkerke's R^2 (Bennett et al. 1999; Nagelkerke 1991) is also included to allow comparisons with the results presented in Hill et al. (2004).

Results

Descriptive Statistics

Table 1 provides the means and standard deviations for the different screen scores for the full sample as well as the subsamples by site and gender. The table also provides the proportion of each sample with an adult conviction.

Predicting Any Conviction

Full Sample Table 2 provides the results when predicting Any Conviction from different combinations of screen scores, controlling for site. When specificity cutoffs were set to reflect the base rate for the sample (0.398), in this case, specificity equals 0.669, all combinations of parent and teacher screen scores met the sensitivity cutoff for well-performing screening tools. When sensitivity cutoffs were set to 0.5, all except one model met the specificity cutoff of 0.669 for a well-performing screening tool, corresponding to PPV of 0.5 at a base rate of 0.398. The model with teacher and parent screens from 1st grade did not meet the requirement. The model using both teacher screens yielded the lowest AIC. The models using all four screens, teacher screen from kindergarten only, and teacher and parent screen from kindergarten also had acceptable model fit with delta values less than 4.

Figure 1 provides the corresponding ROC curves from the six combinations of screen scores. All cutoffs on the ROC curve above the horizontal 0.5 sensitivity line represent cutoffs for which sensitivity was 0.5 or greater. The intersections of the ROC curves with the 0.5 sensitivity line are the points described in the "Sensitivity held at 0.5" column of Table 2. All cutoffs on the ROC curve to the left of the 0.331 vertical line represent cutoffs for which PPV was 0.5 or greater (given the sample base rate). The intersections of the ROC curves with the 0.331 1-specificity line are the points described in the "Specificity held at 0.669" column of Table 2. Therefore, the cutoffs on the ROC curve in the upper left quadrant of the figure reflect cutoffs that met both requirements of a good screening measure. The point furthest away from the diagonal line represents the most accurate cutoff.

While 4 screen score combinations met sensitivity and PPV preset criteria and had acceptable fit (two teacher screen scores representing the best model fit [yellow line in the figure], kindergarten teacher only [red], both kindergarten screen scores [green], and all four scores [brown]), they each had different ideal cutoffs. Each ideal cutoff represents a different trade off between capturing true positives and false positives.

Table 1 Descriptive statistics

	N	Mean	SD
Full Sample			
Teacher K Screen	342	16.77	11.62
Parent K Screen	342	50.77	10.74
Teacher 1st Screen	342	15.49	12.84
Parent 1st Screen	342	7.53	5.79
Any Conviction	334	0.40	
Durham, NC			
Teacher K Screen	95	18.35	11.00
Parent K Screen	95	49.16	10.89
Teacher 1st Screen	95	17.75	13.19
Parent 1st Screen	95	6.13	5.29
Any Conviction	93	0.48	
Nashville, TN			
Teacher K Screen	94	21.15	12.00
Parent K Screen	94	52.50	10.84
Teacher 1st Screen	94	19.37	12.81
Parent 1st Screen	94	9.73	7.15
Any Conviction	94	0.48	
Seattle, WA			
Teacher K Screen	61	14.10	10.85
Parent K Screen	61	50.77	10.03
Teacher 1st Screen	61	15.54	12.67
Parent 1st Screen	61	6.89	5.00
Any Conviction	55	0.27	
Rural Pennsylvania			
Teacher K Screen	92	12.45	10.50
Parent K Screen	92	50.65	10.83
Teacher 1st Screen	92	9.14	10.15
Parent 1st Screen	92	7.15	4.54
Any Conviction	92	0.30	
Males			
Teacher K Screen	172	19.51	11.75
Parent K Screen	172	53.49	10.80
Teacher 1st Screen	172	18.83	13.59
Parent 1st Screen	172	8.97	6.22
Any Conviction	166	0.56	
Females			
Teacher K Screen	170	14.00	10.83
Parent K Screen	170	48.01	9.97
Teacher 1st Screen	170	12.11	11.10
Parent 1st Screen	170	6.08	4.93
Any Conviction	168	0.24	

Table 2 Analyses of Any Conviction by age 25 (all models control for site)

Screen scores included in the model	R2	AIC	Delta	Specificity held at 0.669 (PPV = 0.5; Base Rate = 0.398)		Sensitivity held at 0.5	
				Sens*	Spec	Sens	Spec**
Teacher K (red)	0.18	411.52	3.00	**0.586**	0.692	0.526	**0.746**
Teacher 1st (orange)	0.14	422.43	13.91	**0.526**	0.687	0.504	**0.706**
Teacher K and **Teacher 1st** (yellow)	0.20	408.52		**0.624**	0.682	0.511	**0.726**
Teacher K and Parent K (green)	0.18	412.16	3.64	**0.579**	0.682	0.511	**0.776**
Teacher 1st and Parent 1st (blue)	0.15	421.35	12.83	**0.511**	0.692	0.511	0.692
Teacher K & 1st and Parent K & 1st (brown)	0.20	410.84	2.32	**0.617**	0.677	0.504	**0.766**

Screen score names in bold in column 1 indicate coefficients significant at the 5% level. Colors in parentheses indicate line color in corresponding Fig. 1.

* Bold values indicate models meeting sensitivity requirement for a good screen score system.

** Bold values indicate models meeting specificity requirement (corresponding to PPV equal to 0.5) for a good screen score system

By Site When Any Conviction was examined within each site separately, a different pattern emerged (Table 3). For Durham and Nashville, the base rates for Any Conviction were relatively high, 0.484 and 0.479, respectively, and all the screen score combinations performed well in terms of PPV and sensitivity. In Durham, the kindergarten teacher only model had the lowest AIC, although the models with both teacher scores and both parent and teacher scores from kindergarten had equivalent fit levels with delta values less than or equal to 2. In Nashville, the model with both teacher scores had the lowest AIC; however, all other models had delta scores less than 4 indicating sufficiently good fit.

In contrast, the base rates in Pennsylvania and Seattle were considerably lower, 0.304 and 0.273 respectively, thus, the screen scores did not perform as well. In Pennsylvania, the requirements for PPV and sensitivity were only met when kindergarten teacher and parent screen scores were used. This model also had the lowest AIC. In Seattle, three models met the PPV and sensitivity preset criteria: 1st grade teacher only, both teacher screen scores, and all four screen scores. The model with 1st grade teacher only had the lowest AIC; however, the model with both teacher scores had equivalently good fit with delta less than 2. The model with all four screens had decidedly poorer fit with delta greater than 4.

Figure 2 provides the ROC curves for the individual site models. For Durham and Nashville, the ROC curves for all models fell in the upper left quadrant of the figure, indicating that all combinations of screen scores had cutoffs that met the preset PPV and sensitivity criteria. For Pennsylvania, the ROC curves only passed through the upper left quadrant for two models: kindergarten scores only (green) and all four scores (brown). For Seattle, only two ROC curves passed through the upper left quadrant: teacher scores only (yellow) and all four scores (brown).

By Gender For males (base rate = 0.560), all combinations of screen scores performed well in terms of PPV and sensitivity (Table 4) in predicting Any Conviction. The model with both teacher scores had the lowest AIC. Three other models also had sufficiently good fit with delta values less than 4: kindergarten teacher only, teacher and parent from kindergarten, and all four scores. For females, however, the base rate was low (0.238) and, regardless of which screen scores were used, the screening scores did not perform well in terms of PPV and sensitivity. Figure 3 provides the ROC curves by gender. All ROC curves for males fell within the upper left quadrant, whereas none of the ROC curves fell in the upper left quadrant for females.

Predicting Multiple Convictions

Follow-up analyses examined Multiple Convictions (defined as having two or more adult convictions) by age 25 relative to having zero or one conviction. Similar to Any Conviction, results indicated that the model using both teacher screens had the lowest AIC and met the sensitivity and PPV requirements. The model using all four screens had equivalent model fit with a delta value less than 2 and met the sensitivity and PPV requirements. The models using kindergarten teacher only and both teacher and parent screens from kindergarten met the sensitivity and PPV requirements but did not have adequate fit relative to the best fitting model.

Discussion

Overall, these results indicate that screen scores for conduct problems in early elementary school perform remarkably well

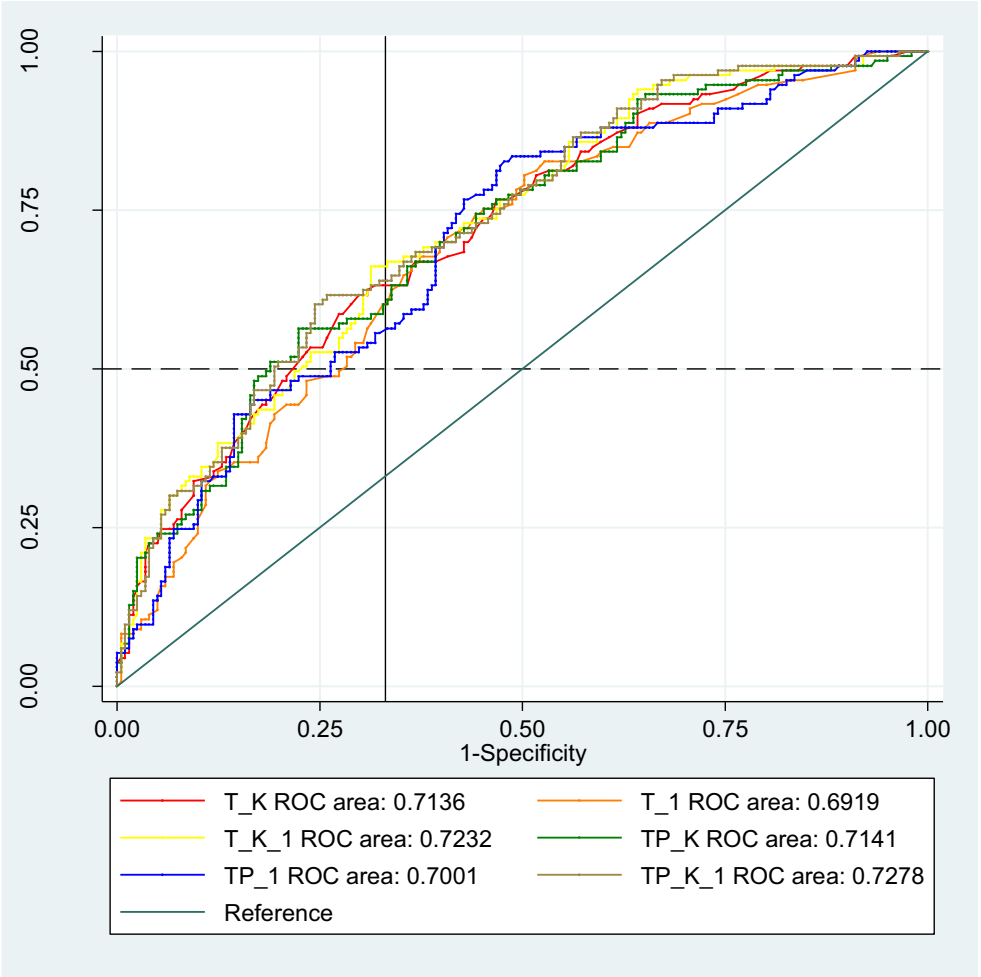


Fig. 1 Full Sample – Predicting Any Conviction by Age 25 (Prevalence = 39%). Note: T_K refers to model including kindergarten teacher screen score. T_1 refers to model including 1st grade teacher screen score. T_K_1 refers to model including kindergarten and 1st grade teacher screen scores. TP_K refers to model including kindergarten teacher and parent screen scores. TP_1 refers to model including 1st grade teacher and parent screen scores. TP_K_1 refers to model including kindergarten and 1st grade teacher and parent screen scores. All points above the dotted horizontal 0.5 sensitivity line represent cutoffs for which sensitivity was 0.5 or greater. All cutoffs to the left of the solid vertical line represent cutoffs for which PPV is 0.5 or greater (given the sample base rate). Therefore, the cutoffs represented by the ROC curve in the upper left quadrant of the figure reflect cutoffs that met both requirements of a good screening measure

when predicting criminal convictions 20 years later. These findings have important implications, suggesting that prevention programs can be meaningfully provided to young children and their families to prevent adult criminal behavior. However, there is variation in screening success by site and gender because of differences in base rates, emphasizing the importance of considering these factors in the use of screening measures.

Combined teacher screens from kindergarten and 1st grade provided the most accurate and parsimonious prediction of Any Conviction for the full sample. Previous research has shown that teacher reports of externalizing behavior during preschool account for up to 45% of the variance in the same measures given during 1st grade (Heller et al. 1996). Therefore, the increased effectiveness of using teacher screens taken at both kindergarten and 1st grade may reflect the relative stability of externalizing behavior in the school setting given that these behaviors were observed during both years. Using teacher screens from both years also helps avoid any reporting bias that may occur from relying on ratings from only one teacher or one year. For example, teachers have been shown to over- or under-estimate students' academic abilities, school readiness, peer acceptance, and teacher-child conflict based on factors such as race, social skills, and presence of inattention symptoms (Baker et al. 2015; Yates and Marcelo 2014). Therefore, relying on reports from multiple teachers across years may help buffer against any teacher bias observed during a single time-point.

In Durham and Nashville, where prevalence rates were high (0.484 and 0.479, respectively), risk for Any Conviction was accurately predicted with as little as one teacher screen of classroom behaviors during kindergarten. This greatly impacts the cost-efficiency of screening for adult criminal behavior for these sites, eliminating the need for multiple-year and multiple-source screens. Therefore, sites with higher base rates may be ideal for more minimalist screening procedures because they allow for higher true positive rates, thereby affecting screen sensitivity. In contrast, the prevalence rates were lower in Pennsylvania and Seattle (0.304 and 0.273, respectively) and therefore, more extensive screening procedures (i.e., using multiple screens) may be necessary to accurately predict adult criminal behavior for low prevalence sites.

The success of the screens also varied by gender. All screening combinations performed well predicting Any

Table 3 Analyses of Any Conviction by age 25, by site

Durham, NC							
Screen scores included in the model	**R2**	**AIC**	**Delta**	**Specificity held at 0.531 (PPV = 0.5; Base Rate = 0.484)**		**Sensitivity held at 0.5**	
				Sens*	**Spec**	**Sens**	**Spec****
Teacher K (red)	0.17	119.8		**0.756**	0.542	0.556	**0.792**
Teacher 1st (orange)	0.06	128.6	8.80	**0.600**	0.604	0.556	**0.646**
Teacher K and Teacher 1st (yellow)	0.17	121.8	2.00	**0.756**	0.542	0.533	**0.729**
Teacher K and Parent K (green)	0.18	121.7	1.90	**0.733**	0.542	0.511	**0.750**
Teacher 1st and Parent 1st (blue)	0.08	129.3	9.50	**0.600**	0.542	0.511	**0.667**
Teacher K & 1st and Parent K & 1st (brown)	0.20	123.8	4.00	**0.711**	0.542	0.511	**0.729**
Nashville, TN							
Screen scores included in the model	**R2**	**AIC**	**Delta**	**Specificity held at 0.54 (PPV = 0.5; Base Rate = 0.479)**		**Sensitivity held at 0.5**	
				Sens*	**Spec**	**Sens**	**Spec****
Teacher K (red)	0.11	125.8	0.20	**0.622**	0.551	0.511	**0.633**
Teacher 1st (orange)	0.09	127.9	2.30	**0.533**	0.551	0.511	**0.571**
Teacher K and Teacher 1st (yellow)	0.14	125.6		**0.556**	0.551	0.511	**0.633**
Teacher K and Parent K (green)	0.11	127.8	2.20	**0.578**	0.551	0.511	**0.633**
Teacher 1st and Parent 1st (blue)	0.12	127.5	1.90	**0.622**	0.551	0.533	**0.592**
Teacher K & 1st and Parent K & 1st (brown)	0.16	128.5	2.90	**0.578**	0.551	0.511	**0.633**
Rural Pennsylvania							
Screen scores included in the model	**R2**	**AIC**	**Delta**	**Specificity held at 0.782 (PPV = 0.5; Base Rate = 0.304)**		**Sensitivity held at 0.5**	
				Sens*	**Spec**	**Sens**	**Spec****
Teacher K (red)	0.14	107.2	2.10	0.393	0.828	0.500	0.766
Teacher 1st (orange)	0.12	109.1	4.00	0.464	0.797	0.500	0.781
Teacher K and Teacher 1st (yellow)	0.18	106.5	1.40	0.464	0.797	0.500	0.766
Teacher K and **Parent K** (green)	0.20	105.1		**0.536**	0.797	0.500	**0.828**
Teacher 1st and Parent 1st (blue)	0.12	111.1	6.00	0.464	0.797	0.500	0.656
Teacher K & 1st and Parent K & 1st (brown)	0.23	106.9	1.80	0.464	0.813	0.500	0.781
Seattle, WA							
Screen scores included in the model	**R2**	**AIC**	**Delta**	**Specificity held at 0.812 (PPV = 0.5; Base Rate = 0.273)**		**Sensitivity held at 0.5**	
				Sens*	**Spec**	**Sens**	**Spec****
Teacher K (red)	0.11	64	3.20	0.267	0.825	0.533	0.675
Teacher 1st (orange)	0.19	60.8		**0.533**	0.825	0.533	**0.850**
Teacher K and Teacher 1st (yellow)	0.19	62.7	1.90	**0.533**	0.825	0.533	**0.825**
Teacher K and Parent K (green)	0.11	66	5.20	0.267	0.825	0.533	0.650
Teacher 1st and Parent 1st (blue)	0.19	62.5	1.70	0.467	0.825	0.533	0.800
Teacher K & 1st and **Parent K** & 1st (brown)	0.20	66.4	5.60	**0.533**	0.825	0.533	**0.825**

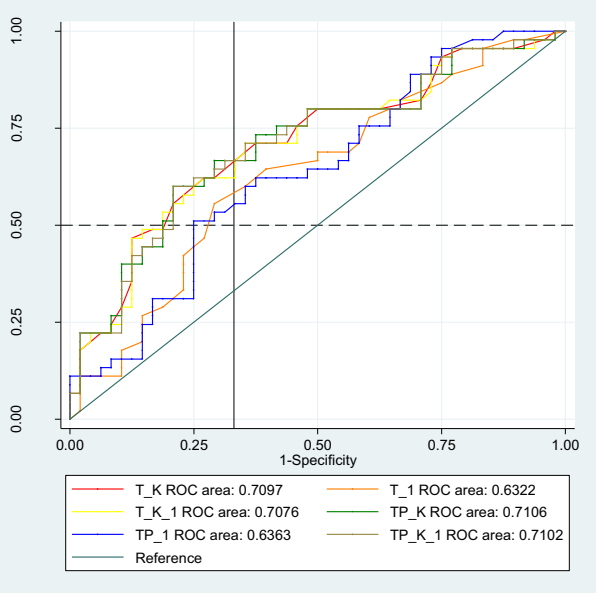
Screen score names in bold in column 1 indicate coefficients significant at the 5% level. Colors in parentheses indicate line color in corresponding Fig. 2.

* Bold values indicate models meeting sensitivity requirement for a good screen score system.

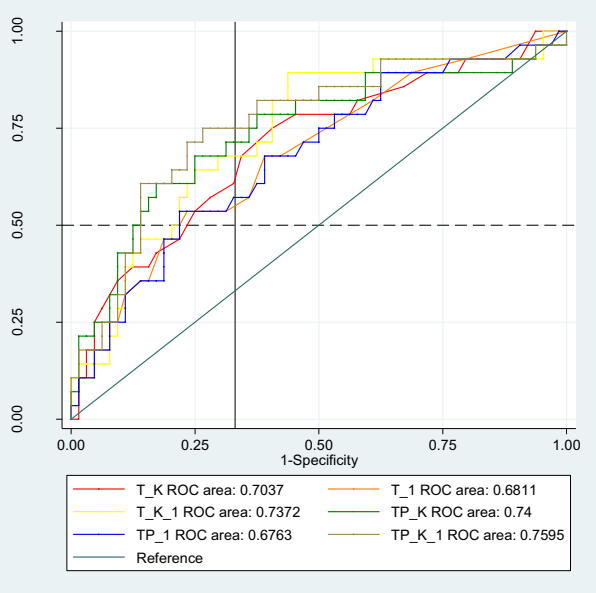
** Bold values indicate models meeting specificity requirement (corresponding to PPV equal to .5) for a good screen score system

Conviction among males (base rate of 0.560), but none performed well among females (base rate of 0.238). These results indicate that for males (like higher base rate sites), more minimalist screening procedures at the time of a child's entry into school (such as a single teacher screen during kindergarten) can accurately predict early adult criminal behavior. For females, on the other hand, other methods of screening may be required to accurately predict later convictions. The higher

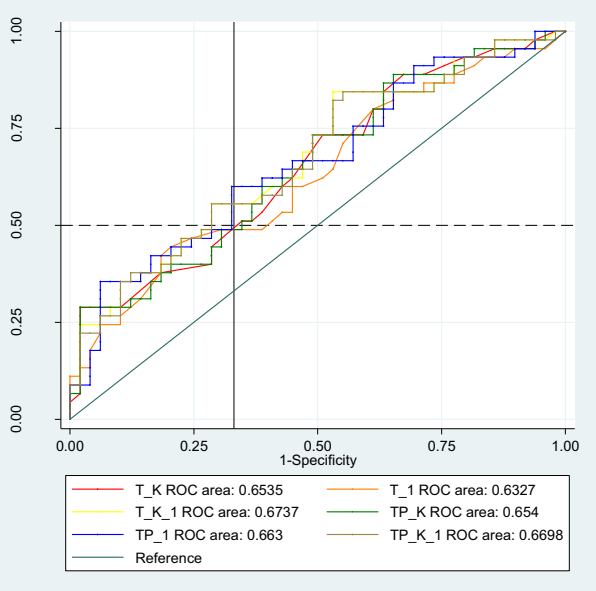
DURH (Prevalence = 48.4%)



PENN (Prevalence = 30.4%)



NASH (Prevalence = 47.9%)



WASH (Prevalence = 27.3%)

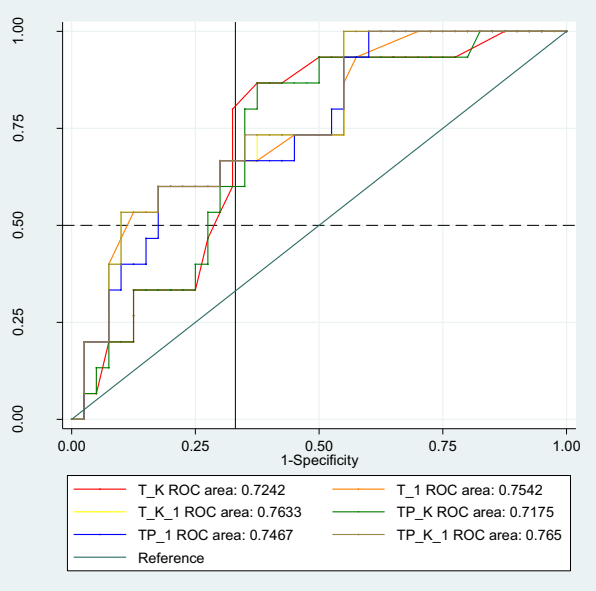


Fig. 2 Predicting Any Conviction by Age 25, by Site. Note: Abbreviations refer to screen scores included in the model: T_K = kindergarten teacher screen score only, T_1 = 1st grade teacher screen score only, T_K_1 = kindergarten and 1st grade teacher screen scores, TP_K = kindergarten teacher and parent screen scores, TP_1 = 1st grade teacher and parent screen scores, and TP_K_1 = kindergarten and 1st grade teacher and parent screen scores. All points above the dotted horizontal 0.5 sensitivity line represent cutoffs for which sensitivity was 0.5 or greater. All cutoffs to the left of the solid vertical line represent cutoffs for which PPV is 0.5 or greater (given the sample base rate). Therefore, the cutoffs represented by the ROC curve in the upper left quadrant of the figure reflect cutoffs that met both requirements of a good screening measure

base rates among boys found in this sample support previous research showing more elevated rates of externalizing behavior in boys than girls (e.g., Martel 2013; Zoccolillo et al. 1996). Thus, it is important to consider gender, as well as site, when estimating base rates and identifying accurate prediction of future criminal behavior by screens. These results are also consistent with research demonstrating more oppositional defiant disorder (ODD) diagnoses among boys than girls based

Table 4 Analyses of Any Conviction by age 25, by gender

Males

Screen scores included in the model	R2	AIC	Delta	Specificity held at 0.364 (PPV = 0.5; Base Rate = 0.560)		Sensitivity held at 0.5	
				Sens*	Spec	Sens	Spec**
Teacher K (red)	0.20	211.49	0.54	**0.817**	0.384	0.505	**0.753**
Teacher 1st (orange)	0.16	216.00	5.05	**0.860**	0.370	0.527	**0.753**
Teacher K and Teacher 1st (yellow)	0.21	210.95		**0.849**	0.370	0.505	**0.767**
Teacher K and Parent K (green)	0.20	213.41	2.46	**0.796**	0.384	0.505	**0.726**
Teacher 1st and Parent 1st (blue)	0.17	217.19	6.24	**0.882**	0.370	0.527	**0.753**
Teacher K & 1st and Parent K & 1st (brown)	0.22	214.41	3.46	**0.849**	0.370	0.505	**0.753**

Females

Screen scores included in the model	R2	AIC	Delta	Specificity held at 0.844 (PPV = 0.5; Base Rate = 0.238)		Sensitivity held at 0.5	
				Sens*	Spec	Sens	Spec**
Teacher K (red)	0.11	182.18		0.250	0.852	0.500	0.633
Teacher 1st (orange)	0.06	187.59	5.41	0.250	0.852	0.575	0.516
Teacher K and Teacher 1st (yellow)	0.11	184.17	2.00	0.250	0.852	0.525	0.617
Teacher K and Parent K (green)	0.11	184.13	1.95	0.250	0.852	0.525	0.617
Teacher 1st and Parent 1st (blue)	0.06	189.59	7.41	0.250	0.852	0.550	0.508
Teacher K & 1st and Parent K & 1st (brown)	0.11	188.11	5.93	0.250	0.852	0.525	0.578

Screen score names in bold in column 1 indicate coefficients significant at the 5% level. Colors in parentheses indicate line color in corresponding Fig. 3.

* Bold values indicate models meeting sensitivity requirement for a good screen score system.

** Bold values indicate models meeting specificity requirement (corresponding to PPV equal to 0.5) for a good screen score system

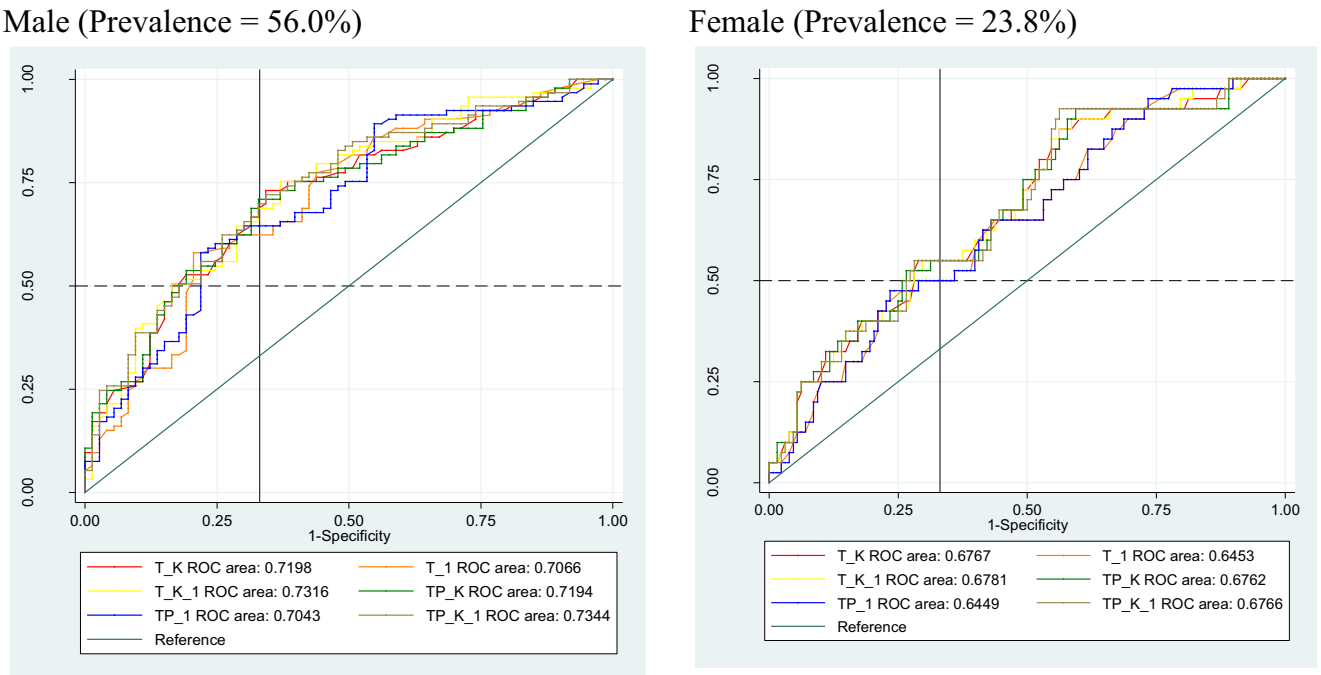


Fig. 3 Predicting Any Conviction by Age 25, by Gender. Note: Abbreviations refer to screen scores included in the model: T_K = kindergarten teacher screen score only, T_1 = 1st grade teacher screen score only, T_K_1 = kindergarten and 1st grade teacher screen scores, TP_K = kindergarten teacher and parent screen scores, TP_1 = 1st grade teacher and parent screen scores, and TP_K_1 = kindergarten and 1st grade teacher and parent screen scores. All points above the dotted horizontal 0.5 sensitivity line represent cutoffs for which sensitivity was 0.5 or greater. All cutoffs to the left of the solid vertical line represent cutoffs for which PPV is 0.5 or greater (given the sample base rate). Therefore, the cutoffs represented by the ROC curve in the upper left quadrant of the figure reflect cutoffs that met both requirements of a good screening measure

on teacher reports, suggesting that teachers may be less sensitive to ODD symptoms in girls, leading to an under-estimation of ODD in this population (McNeilis et al. 2017). This further suggests that factors other than the base rate of a behavior could affect the predictive utility of early behavior problems among girls; therefore, further research is needed on the best practices for predicting later externalizing outcomes among girls.

Combined teacher and parent screens were effective in predicting Any Conviction, but the effectiveness of the year (kindergarten versus 1st grade) varied by site. In Durham and Pennsylvania, combined teacher and parent screens taken during kindergarten performed well, while in Nashville and Seattle, combined teacher and parent screens taken during 1st grade performed well. Overall, however, results suggest that screens assessing home-based behavior are not necessary for the accurate prediction of adult criminal behavior and that adequate screening can be accomplished with classroom-based screening alone. This is consistent with previous research conducted over shorter time periods in elementary school which has indicated that parent-screening adds little incremental validity when teacher screens are available (e.g., Hill et al. 2004; Lochman & CPPRG 1995). Further, as noted, in populations with higher base rates, a single measure of classroom-based behavior taken during kindergarten can accurately predict adult criminal behavior. It is also important to note that mean levels of teacher screen scores varied across site in ways that paralleled the differential rates of future crime. In contrast, parent screen scores were very similar across the sites. It is possible that parents have less basis for comparison with other children in the home setting than teachers do in the school setting, making it difficult for them to determine normative levels of these behaviors. In addition, aggressive behavior exhibited at home may reflect parent-child conflict which, alone, may signal less risk for future crime than aggressive behavior exhibited in school settings, which may represent more pervasive self-regulation difficulties (Stormshak, Bierman, & CPPRG 1998).

Limitations

Despite the strengths of the current study (such as the use of a large, nationally-representative sample), there are several limitations that should be considered. While the current study drew its sample from four unique sites, future research should work to replicate these results in samples from other locations. In addition, given the success of screening procedures in predicting convictions effectively only for boys, it is important that future research determine better ways to predict outcomes for at-risk girls. It is possible that girls' overt externalizing behavior is less severe and therefore less stable across time, leading them to "fall out" of these predictive trajectories (Hill et al. 2004). In the future, it will be important to continue to assess the utility and cost-effectiveness of screening procedures specifically for low base rate populations (e.g., girls). Finally, some participants from the original sample did not have complete data and therefore could not be included in analyses. However, 86% of participants from the original study were still able to be used in the present study.

Implications and Conclusions

Successful screening for later externalizing behavior is a prerequisite for targeted prevention for at-risk children. Findings from this study confirm that such screening procedures exist and can accurately predict externalizing behavior outcomes, such as adult convictions, almost 20 years later. In addition, this study demonstrated that screening procedures can be particularly effective in populations in which base rates of externalizing behaviors are high. Understanding the utility, but also the limitations, of screening procedures is essential to providing the most effective, targeted preventative interventions. Therefore, these findings can inform the development and use of cost-effective and accurate screening tools in high-risk communities to prevent externalizing outcomes, such as convictions, through early intervention.

Acknowledgments This work was supported by National Institute of Mental Health (NIMH) grants R18 MH48043, R18 MH50951, R18 MH50952, R18 MH50953, K05MH00797 and K05MH01027; Department of Education grant S184 U30002; and NIDA grants DA16903, DA017589, K05DA015226, and P30DA023026. The Center for Substance Abuse Prevention and the National Institute on Drug Abuse also provided support through a memorandum of agreement with the NIMH. Additional support for the preparation of this work was provided by a LEEF B.C. Leadership Chair award, Child & Family Research Institute Investigator Salary and Investigator Establishment Awards, and a Canada Foundation for Innovation award to Robert J. McMahon. Funders provided financial support, but responsibility for the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, and approval of the manuscript rests solely with the authors. 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.

The Conduct Problems Prevention Research Group includes (alphabetically): Karen L. Bierman, Ph.D., Pennsylvania State University; John D. Coie, Ph.D., Duke University; Kenneth A. Dodge, Ph.D., Duke University; Mark T. Greenberg, Ph.D., Pennsylvania State University; John E. Lochman, Ph.D., University of Alabama; Robert J. McMahon, Ph.D., Simon Fraser University and B.C. Children's Hospital; and Ellen E. Pinderhughes, Ph.D., Tufts University.

Compliance with Ethical Standards

Conflict of Interest Drs. Bierman, Coie, Dodge, Greenberg, Lochman, McMahon, and Pinderhughes are the principal investigators on the Fast Track Project and have a publishing agreement with Guilford Publications, Inc. Royalties from that agreement will be donated to a

professional organization. They are also authors of the PATHS curriculum and donate all royalties from Channing-Bete, Inc. to a professional organization. Dr. Greenberg is a developer of the PATHS curriculum and has a separate royalty agreement with Channing-Bete, Inc. Bierman, Coie, Dodge, Greenberg, Lochman, and McMahon are the developers of the Fast Track curriculum and have publishing and royalty agreements with Guilford Publications, Inc. Dr. McMahon is a coauthor of Helping the Noncompliant Child and has a royalty agreement with Guilford Publications, Inc.

Ethical Approval All procedures were approved by the Institutional Review Boards of participating universities (i.e., Duke University, University of Washington, Vanderbilt University, and Penn State University).

Informed Consent Written informed consent from parents and oral assent from children were obtained for the collection of demographic and screening variables. Additional informed consent was not required for the collection of adult conviction data, given the public accessibility of these data.

References

Achenbach, T. M. (1982). *Developmental psychopathology* (2nd ed.). New York: Wiley.

Achenbach, T. M. (1991). *Manual for the child behavior checklist/4–18 and 1991 profile*. Department of Psychiatry: University of Vermont.

Baker, C. N., Tichovolsky, M. H., Kupersmidt, J. B., Voegler-Lee, M. E., & Arnold, D. H. (2015). Teacher (mis)perceptions of preschoolers' academic skills: Predictors and associations with longitudinal outcomes. *Journal of Educational Psychology, 107*, 805–820. https://doi.org/10.1037/edu0000008.

Bennett, K. J., & Offord, D. R. (2001). Screening for conduct problems: Does the predictive accuracy of conduct disorder symptoms improve with age? *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1418–1425. https://doi.org/10.1097/00004583-200112000-00012.

Bennett, K. J., Lipman, E. L., Racine, Y., & Offord, D. (1998). Do measures of externalizing behavior in normal populations predict later outcome? Implications for targeted interventions to prevent conduct disorder. *Journal of Child Psychology and Psychiatry, 39*, 1059–1070.

Bennett, K. J., Lipman, E. L., Brown, S., Racine, Y., Boyle, M. H., & Offord, D. R. (1999). Predicting conduct problems: Can high-risk children be identified in kindergarten and grade 1? *Journal of Consulting and Clinical Psychology, 67*, 470–480. https://doi.org/10.1037/0022-006X.67.4.470.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304.

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., et al. (1993). The science of prevention: A conceptual framework and some directions for a national research program. *The American Psychologist, 48*, 1013–1022.

Conduct Problems Prevention Research Group. (1992). A developmental and clinical model for the prevention of conduct disorder: The FAST track program. *Development and Psychopathology, 4*, 509–527. https://doi.org/10.1017/S0954579400004855.

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

Conduct Problems Prevention Research Group. (2000). Merging universal and indicated prevention programs: The Fast track model. *Addictive Behaviors, 25*, 913–927.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*, 858–900. https://doi.org/10.1037/a0038498.

Drabick, D. A. G., Bubier, J., Chen, D., Price, J., & Lanza, H. I. (2011). Source-specific oppositional defiant disorder among inner-city children: Prospective prediction and moderation. *Journal of Clinical Child & Adolescent Psychology, 40*, 23–35. https://doi.org/10.1080/15374416.2011.533401.

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review, 13*, 409–419. https://doi.org/10.1016/0272-7358(93)90012-B.

Goulter, N., Godwin, J., & Conduct Problems Prevention Research Group. (2018). *Person-oriented analyses of Fast Track effects: Typologies of adult criminal convictions.* Poster presented at the Banff international conference on behavioral science, Banff, BC, Canada.

Heller, T. L., Baker, B. L., Henker, B., & Hinshaw, S. P. (1996). Externalizing behavior and cognitive functioning from preschool to first grade: Stability and predictors. *Journal of Clinical Child Psychology, 25*, 376–387. https://doi.org/10.1207/s15374424jccp2504_3.

Hill, L. G., Coie, J. D., Lochman, J. E., & Greenberg, M. T. (2004). Effectiveness of early screening for externalizing problems: Issues of screening accuracy and utility. *Journal of Consulting and Clinical Psychology, 72*, 809–820. https://doi.org/10.1037/0022-006X.72.5.809.

Jones, D., Dodge, K. A., Foster, E. M., Nix, R., & Conduct Problems Prevention Research Group. (2002). Early identification of children at risk for costly mental health service use. *Prevention Science: The Official Journal of the Society for Prevention Research, 3*, 247–256.

Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health, 105*, 2283–2290. https://doi.org/10.2105/AJPH.2015.302630.

Kaplow, J. B., Curran, P. J., Dodge, K. A., & Conduct Problems Prevention Research Group. (2002). Child, parent, and peer predictors of early-onset substance use: A multisite longitudinal study. *Journal of Abnormal Child Psychology, 30*, 199–216.

Lochman, J. E., & Conduct Problems Prevention Research Group. (1995). Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology, 63*, 549–559.

Martel, M. M. (2013). Sexual selection and sex differences in the prevalence of childhood externalizing and adolescent internalizing disorders. *Psychological Bulletin, 139*, 1221–1259. https://doi.org/10.1037/a0032247.

Matthys, W., & Lochman, J. E. (2010). *Oppositional defiant disorder and conduct disorder in childhood*. Oxford: Wiley-Blackwell.

McNeilis, J., Maughan, B., Goodman, R., & Rowe, R. (2017). Comparing the characteristics and outcomes of parent- and teacher-reported oppositional defiant disorder: Findings from a national sample. *Journal of Child Psychology and Psychiatry, 59*, 659–666. https://doi.org/10.1111/jcpp.12845.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

Nix, R. L. (2001). Child Behavior Checklist (Technical report). https://doi.org/http://www.fasttrackproject.org/technical-reports.php.

Okado, Y., & Bierman, K. L. (2014). Differential risk for late adolescent conduct problems and mood dysregulation among children with early externalizing behavior problems. *Journal of Abnormal Child Psychology, 43*, 735–747. https://doi.org/10.1007/s10802-014-9931-4.

Pasalich, D. S., Witkiewitz, K., McMahon, R. J., Pinderhughes, E. E., & Conduct Problems Prevention Research Group. (2015). Indirect effects of the Fast track intervention on conduct disorder symptoms and callous-unemotional traits: Distinct pathways involving discipline and warmth. *Journal of Abnormal Child Psychology, 44*, 587–597. https://doi.org/10.1007/s10802-015-0059-y.

Petras, H., Chilcoat, H. D., Leaf, P. J., Ialongo, N. S., & Kellam, S. G. (2004a). Utility of TOCA-R scores during the elementary school years in identifying later violence among adolescent males. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*, 88–96. https://doi.org/10.1097/00004583-200401000-00018.

Petras, H., Schaeffer, C. M., Ialongo, N., Hubbard, S., Muthén, B., Lambert, S. F., et al. (2004b). 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, 16*, 919–941.

Sawyer, A. C. P., Chittlrborough, C. R., Lynch, J. W., Baghurst, P., Mittiny, M. N., Kaim, A. L. E., & Sawyer, M. G. (2014). Can screening 4-5 year olds accurately identify children who will have teacher-reported mental health problems when children are aged 6-7 years? *Australian and New Zealand Journal of Psychiatry, 48*, 554–563.

Stormont, M. (2000). Early child risk factors for externalizing and internalizing behaviors: A 5-year follow-forward assessment. *Journal of Early Intervention, 23*, 180–190. https://doi.org/10.1177/105381510002300307O1.

Stormshak, E. A., Bierman, K. L., & Conduct Problems Prevention Research Group. (1998). The implications of different developmental patterns of disruptive behavior problems for school adjustment. *Development and Psychopathology, 10*, 451–467.

Welsh, B. C., Loeber, R., Stevens, B. R., Stouthamer-Loeber, M., Cohen, M. A., & Farrington, D. P. (2008). Costs of juvenile crime in urban areas: A longitudinal perspective. *Youth Violence and Juvenile Justice, 6*, 3–27. https://doi.org/10.1177/1541204007308427.

Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Teacher observation of classroom adaptation–revised. *PsycTESTS*. https://doi.org/10.1037/t31163-000.

Wichstrøm, L., Skogen, K., & Oia, T. (1996). Increased rate of conduct problems in urban areas: What is the mechanism? *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 471–479.

Yates, T. M., & Marcelo, A. K. (2014). Through race-colored glasses: Preschoolers' pretend play and teachers' ratings of preschooler adjustment. *Early Childhood Research Quarterly, 29*, 1–11. https://doi.org/10.1016/j.ecresq.2013.09.003.

Zoccolillo, M., Tremblay, R., & Vitaro, F. (1996). DSM-III-R and DSM-III criteria for conduct disorder in preadolescent girls: Specific but insensitive. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*, 461–470. https://doi.org/10.1097/00004583-199604000-00012.