Validation of a Parent Report on Externalizing Symptoms Scale: A Downward Extension of the Behavioral Health Screen

Alannah Shelby Rivers, PhD[a]

Payne Winston-Lindeboom, M.A.[a]

Linda Ruan-Iu, PhD, NCSP[a]

Tita Atte, MPH, CPH[a]

Allen Tien, MD, MHS[b]

Guy Diamond, PhD[a]

[a]Center for Family Intervention Science, Drexel University, 3141 Chestnut St, Philadelphia, PA 19104, USA

[b]Medical Decision Logic, 1501 St Paul St, Baltimore, MD 21202, USA

Author Note

The corresponding author (Alannah Shelby Rivers; https://orcid.org/0000-0001-8826-4606) is now at Texas Woman's University and may be contacted at arivers1@twu.edu. Funding for this paper was also provided by SAMHSA's 1U79SM061750 Garrett Lee Smith Suicide Prevention Program. The Behavioral Health Screening tool is licensed to Medical Decision Logic, Inc., a health science informatics and computer science engineering company. Dr. Tien is the Founder, President, and Chief Science Officer of this company. Dr. Diamond might one day receive a small royalty payment for his part

in developing the tool. Dr. Rivers and the other coauthors do not report financial interests or potential conflicts of interest.

Abstract

Externalizing problems are common in children ages 6-14, can have lifelong consequences, and may pose particular risk when combined with other risk factors and symptoms (like depression and anxiety). Schools are uniquely positioned to assess and address these types of behavioral health concerns, but many school-based assessments do not focus on mental health distress (partially because they often lack the infrastructure for identification, screening, and referral). To address this gap, the bhworks student mental health software system has integrated teacher training, psychometrically strong assessments, feedback, and referral tools. However, this self-report tool for adolescents needed to be adapted for younger children. Thus, a parent report version was added as well as new scales for better assessing this age group. The current study examines the psychometric properties of the new parent-report attention-deficit/hyperactivity and oppositional defiant/conduct scales within a sample of 440 children referred for school-based assessments. Overall, the new scales demonstrated good structural validity, measurement invariance across most demographic groups, discrimination in item response theory analyses, and evidence of convergent validity and good classification accuracy in relation to a validation battery. These externalizing scales are distinct and precise and show promise for improving the effectiveness of school-based programs for identifying at-risk children.

Keywords: ADHD, conduct disorder, externalizing, psychometrics

Impact and Implications Statement

To address challenges of behavioral health school-based screening, this project tested new measures assessing parent-reported attention-deficit/hyperactivity and oppositional-defiant/conduct symptoms for elementary-age children. This study finds that the

new scales show initial promise and work well for a sample of children referred to behavioral health services. These scales are part of an assessment system which may help close gaps in receipt of services for students with these concerns.

Validation of a Parent Report on Externalizing Symptoms Scale: A Downward Extension of the Behavioral Health Screen

Externalizing behaviors, such as symptoms of attention-deficit hyperactivity disorder (ADHD) and oppositional defiant disorder (ODD) or conduct disorder (CD), are quite common in elementary and middle-school children. Between 7 and 9% of elementary- and middle-school children are diagnosed with ADHD (Danielson et al., 2018; Xu et al., 2018), and between 4 and 7% of children meet the criteria for ODD or CD (Boat & Wu, 2015; Ghandour et al., 2019). There is evidence that having symptoms for one of these disorders as a youth increases risk for the other (Atherton et al., 2018). School-aged children experiencing such externalizing problems are at greater risk of consequences like academic underachievement, peer victimization, and, in the long-term, depression, risky behavior, and suicide (Duprey et al., 2020; Klein et al., 2012; Masten et al., 2005; van Lier et al., 2012). Externalizing problems also have economic implications, like higher costs in behavioral and medical care services delivery (Christenson et al., 2016; Matza et al., 2005). Moreover, many children with sub-clinical problems display related symptoms at an early age; if untreated, these symptoms may develop into a psychiatric disorder in adolescence or adulthood (McMahon, 1994; Nock et al., 2007). Unfortunately, up to 25-50% of youth with externalizing problems are not receiving treatment (Danielson et al., 2018; Ghandour et al., 2019), partly due to challenges with identification, referral, and treatment systems. Therefore, there is a need for effective early identification and referral to appropriate care.

Behavioral Health in Schools

Schools are a unique context for identifying behavioral health needs. Although other contexts (e.g., primary care) are certainly important, schools have more contact with children and fewer barriers to access, particularly for marginalized children (Bruns et al., 2016). In schools, parents and teachers have unique opportunities to identify potential concerns, and externalizing behaviors are particularly salient within a school context; these behavior problems disrupt the learning experience not only for the target child but for other children as well (Lane & Walker, 2015). Schools may not want the burden of treating these youth, but there is an academic and ethical imperative to develop procedures for screening, assessment, and triage. To address this need, there has been a greater focus on social-emotional development and behavioral health within a multi-tiered system of support (MTSS; NASP, 2014). In such systems, screening is often used to identify and evaluate students who could benefit from support at higher tiers.

Unfortunately, mental health screening in schools remains underdeveloped. Many school-based screening programs are primarily concerned with skills-based, academic outcomes (Bruhn et al., 2014), which may have limited utility for detecting behavioral health concerns. In many schools, few students receive more clinically focused screening (Connors et al., 2015; Dowdy et al., 2010). There are often gaps in receipt of services for children with more psychiatric needs (often due to screening difficulties or lack of continuity between systems; Bruhn et al., 2014; Walker et al., 2005). Although widely-available, high-quality screening increases access to services and referrals (Husky et al., 2011), school decision-makers can be hesitant to evaluate behavioral health problems, fearing they will be overwhelmed (Severson et al., 2007). Teachers may also be less comfortable identifying clinical symptoms and may need additional training for behavioral health concerns (Severson et al., 2007). Finally, when schools want to refer students for services, their behavioral health “neighborhoods” often remain small (J. G. Green et al., 2013; Stiffman et al., 2000). These broader, systems-level challenges suggest that in addition to psychometrically-sound screening tools, schools need well-developed screening *programs* that can help schools manage training, identification, and referrals within a multi-tiered framework.

One systems-change solution has been the Behavioral Health Works program (bhworks). bhworks is a multi-component, web-based, commercially available system that supports all aspects of the screening process in schools and other settings. The platform for schools includes prerecorded teacher training modules to increase staff comfort, readiness, and willingness to address behavioral health, and to help school personnel develop skills to better engage parents. The platform then delivers a comprehensive mental health screening tool, which generates a fully automated report with acute risk factors, symptoms, risk behaviors, and strengths. The system also includes safety planning, an electronic resource guide to assist with referrals, and a real-time data dashboard that aggregates and analyzes data for program monitoring and policy decisions. This platform has been implemented widely in Pennsylvania, Michigan, California, and Kansas.

The Behavioral Health Screen

A cornerstone of any screening system is the screening tool itself. bhworks uses the Behavioral Health Screen (BHS; G. Diamond et al., 2010), a brief (10-minute) electronic screener that has previously demonstrated good psychometric characteristics in adolescents (Bevans et al., 2012; G. Diamond et al., 2010). The adolescent version assesses 14 domains: demographics, medical, school, family, safety and violence, substance use, sexual risk, nutrition and eating, anxiety, depression, suicide and self-harm, psychosis, trauma, and bullying, using 61 main (mostly Likert-type) questions and 46 skip-out questions (which are only required if certain symptoms are endorsed). The program scores the survey and instantly generates a report. An electronic referral system allows up-to-date contact with potential community partners to expedite referrals (Twymon et al., 2020). Finally, the electronic administration allows for immediate feedback and real-time aggregate data analysis (Bickman et al., 2011).

The BHS was originally developed as a self-report tool for adolescents and young adults. However, given the large number of children referred for evaluations, the bhworks team felt the need to develop a downward extension: a parent report on child functioning, appropriate for younger children. Although teachers make most referrals in school settings, parents have opportunities to observe their child's behavior over longer periods of time and across multiple contexts (K. E. Diamond & Squires, 1993). However, parents often do not know what to look for, making well-designed scales particularly important. This expansion also includes new scales examining attention-deficit/hyperactivity (ADHD) and oppositional defiant/conduct (ODD/CD) symptoms, problems both common to younger children and easily observable. For these types of externalizing symptoms, high-quality parent reports are particularly relevant for reducing subjectivity and bias. Psychometric validation is necessary to ensure these scales produce precise and meaningful assessments for school-based behavioral health; notably, an initial validation has already been completed of the adapted depression and anxiety and new sleep scales (Ruan-Iu et al., 2022), with future research focusing on the other adapted and new scales. For externalizing symptoms, it is especially important to ensure not only generally good scale functioning but also measurement invariance across gender, given established gender differences across the lifespan (wherein male individuals typically display more externalizing symptoms; Boyd et al., 2015; Lau et al., 2021).

Therefore, this paper begins the validation of the two new externalizing scales that were added to the parent-report version of the BHS. Specifically, we explore *structural validity* (items are good indicators of distinct underlying constructs), *measurement invariance* (items hold their meaning across demographic groups and are not biased), *item discrimination* (items have adequate precision to detect meaningful differences in responses), *convergent and discriminant validity* (scales are strongly associated with their own validation scales, and associated positively with negative school behavior and negatively with grades), and *classification accuracy* (scales have adequate sensitivity and specificity in identifying children within clinical ranges). The current study evaluates these psychometric characteristics within a sample of children referred for an evaluation from school-based behavioral health assessment teams across Pennsylvania.

Method

Participants

Overall, 440 parents reported on their child. Children were ages 6-14 (*M* = 8.79, *SD* = 2.03) and 35.5% were girls (64.5% were boys, including one transgender boy). Children were approximately 48.6% White/Caucasian, 29.5% Black/African American, 0.5% American Indian/Alaska Native, 1.1% Asian, 0.2% Native Hawaiian/Other Pacific Islander, and 15.0% more than one race (5.0% uncertain). Approximately 20.2% were Hispanic.

Procedure

Data were collected between April 2018 and June 2019 by 12 Pennsylvania Student Assistant Program (SAP) agencies located across the state, mostly in rural and small urban communities. SAP is the main behavioral health assessment process for students in Pennsylvania. The referral process generally begins when a teacher identifies a youth with academic or behavioral health problems and refers the youth to a school committee who evaluates the student’s needs. If the committee determines that a behavioral health assessment is needed (about 70% of referrals), the school invites their designated behavioral health specialist from a local community mental health center to do an evaluation at the school (most of these specialists use bhworks and the BHS to conduct this evaluation). In this regard, this study sample was not a normative, general population of students in schools, but instead, students identified as at risk for behavioral health problems.

Before the behavioral health assessment and after the initial evaluation, parents were contacted for consent and invited to participate in the assessment. Consent included permission for the SAP agency to share deidentified data from the assessment with the research team. If the child was in the target age range, the parent completed a battery of assessments (the BHS screening and other assessment tools used for validation) as standard care. Only those fully completing both assessments were included in the current study. The online system did not allow for items to be missed, although parents could refuse to respond to items. No parents refused items used in the current investigation. This study was reviewed and approved by the Drexel University IRB.

Measures

Behavioral Health Screen

The Behavioral Health Screen (BHS) is a brief, comprehensive behavioral health screening tool (G. Diamond et al., 2010) hosted and distributed by Medical Decision Logic, Inc. ("mdlogix"), a health informatics technology company (www.bh-works.com). Because the BHS was originally designed for adolescent self-report, we reworded the items for parent-report on 6–14-year-old children. The sexual risk and psychosis domains were removed, and four new domains were added. The current study evaluates two of these new domains: attention-deficit/hyperactivity symptoms, and oppositional defiant/conduct symptoms. Of the remaining entirely new domains, one was assessed in a separate study (sleep; Ruan-Iu et al.,

2022) and one underwent initial pilot testing for a future validation project (autism spectrum disorder). Items for the new domains described here were initially created based on DSM-5 criteria. Then, a team of psychologists selected and revised the items. For the ADHD symptom items, items were selected to represent the two main categories of behavior that characterize the disorder: Inattention (1 item) and Hyperactivity/Impulsivity (2 items), plus an item assessing overall impairment. For the ODD/CD items, items were selected to correspond to the two disorders: Oppositional Defiant Disorder (2 items, assessing two major areas of Angry/Irritable Mood and Argumentative/Defiant Behavior) and Conduct Disorder (2 items assessing Aggression to People or Animals and Destruction of Property). Items were chosen based on symptoms most likely to be developmentally appropriate for this age range.

Attention-deficit/Hyperactivity. Four items assessed ADHD symptoms: three items assessing indicators of difficulty concentrating, impulsivity, and hyperactivity, and one item assessing overall impairment in daily life. All items were scored on a three-point scale ("Never" [0], "Sometimes" [2] "Often" [4]). Internal reliability was good (alpha = .88).

Oppositional Defiant/Conduct. Five items assessed ODD/CD symptoms: four items assessing indicators of defiance, anger, interpersonal aggression, and property damage, and one item assessing overall impairment in daily life. All items were initially scored on a three-point scale ("Never" [0], "Sometimes" [2] "Often" [4]); however, based on preliminary analyses indicating that the intermediate option was rarely endorsed and, in item response theory analyses, failed to provide any additional information beyond the endorsement of the "often" option, all items were dichotomized (combining "Sometimes" and "Often" into a single response [4]). Internal reliability was acceptable (alpha = .70).

School Behavior and Grades. Two additional BHS items in the school domain were selected for convergent validity, given expected associations with externalizing behaviors. One item asked how often parents had been contacted by the school about their children's negative behavior ("Never" [0], "Sometimes" [2] "Often" [4]), and one item asked about the child's average grades (in letter grades).

Child and Adolescent Symptom Inventory Progress Monitor

The Child and Adolescent Symptom Inventory Progress Monitor (CASI-PM; Lavigne et al., 2009) is a 28-item battery assessing psychiatric symptoms for youth aged 3 to 18. The CASI-PM has parent and teacher versions; selected scales from the parent version (CASI-PM-P) were used as the validation battery. The CASI-PM-P was developed using items from the Symptom Inventories scale which has been tested and validated in a 26-year long program (Gadow & Sprafkin, 1994; Lavigne et al., 2009). When developing the CASI-PM-P, the authors used items that had the strongest correlations (moderate to good) with the Symptom Inventories items. Additionally, the CASI-PM-P showed high intercorrelations with other externalizing scales, high internal consistency among all the externalizing scales, high test-retest reliability (>.7), and concurrent validity. All items are scored as Never, Sometimes, or Often.

ADHD Subscales. Eight items assessed ADHD symptoms; four items assessing inattention: inattention to details, difficulty paying attention, difficulty following instructions, and difficulty organizing tasks; and four items assessing hyperactivity/impulsivity: difficulty remaining seated, difficulty being quiet, on the go, and difficulty awaiting turn. Internal reliability was good (alpha = .92).

Oppositional-Defiant and Conduct-Aggression Subscales. Eight items assessed oppositional-defiant and conduct-aggression symptoms. Three items assessed CD indicators: bullying or threatening others, starting fights, and destroying others' property. Five items assessed ODD indicators: defiance, anger and resentment, deliberate annoyance of others, and arguing with adults. Internal reliability was good (alpha = .89).

Analytic Plan

Internal Structure

First, structural validity was tested using confirmatory factor analysis (CFA) in the R package "lavaan" (Rosseel, 2012) using the diagonally weighted least squares estimator for ordinal data. A well-fitting model was defined *a priori* (CFI ≥ .95, SRMR ≤ .09, RMSEA ≤ .08; (Browne & Cudeck, 1992; Hu & Bentler, 1999), and standardized loadings were expected to be "good" or better (.55 or above; Comrey & Lee, 1992).

Then, measurement invariance was examined across groups with cell sizes of at least 100: gender (boys and girls), race (Black and White only, due to small cell sizes in other racial categories), age (ages 6-9 and ages 10-14), and special education usage (yes and no). Invariance was tested using the four-step procedure for ordinal data (Bowen & Masa, 2015): 1) testing separate baseline models, 2) testing a single factor model for all groups (configural invariance), 3) testing a model constraining the item loadings to be equal across groups (metric invariance), and 4) testing a model constraining item loadings and thresholds to be equal across groups (scalar invariance). Models were tested separately for the two scales to detect any initial concerns with the functioning of each independent subscale. In line with recommendations for ordinal data (Brown, 2006; Chen, 2008), we did not assess residual invariance in the current models. In addition to the absolute model fit criteria, relative fit between models was examined. For ordinal data, there is a suggested cutoff of ΔCFI more than -.004 along with ΔRMSEA ≥ .050 (Rutkowski & Svetina, 2017).

Then, item response theory analyses were tested in the R package “ltm” (Rizopoulos, 2006), using the graded response model (Samejima, 1969) for polytomous data, and the two-parameter logistic latent trait model for binary data (Birnbaum, 1968). All items were expected to have at least “moderate” discrimination (0.65 or higher; Baker, 2001).

Convergent and Discriminant Validity

Convergent validity was tested using correlations with validation scales and other criterion measures (using Spearman correlation coefficients for the latter, as they were ordinal), as well as standardized estimates controlling for the other scale(s). Then, to examine discriminant validity, relative strength of association with validation scales was tested using a structural equation model constraining the betas to be equal. All variables were standardized prior to this analysis, so this effectively constrains the standardized betas to be equal. When this model is compared to the baseline model using a chi-square test with one degree of freedom, a significant chi-square would indicate poor fit for the model where the betas are the same (difference between the betas).

Diagnostic Accuracy

To assess diagnostic accuracy and identify clinical cutoffs, receiver operating characteristic (ROC) analyses were tested using the R packages “pROC” (Robin et al., 2011) and “ROCit” (Khan & Brandenburger, 2019). Published clinical cutoffs for the CASI-PM subscales (6- to 12-year old version; Lavigne et al., 2009) were used to dichotomize those variables, and the Youden index was used to identify cutoffs. Because gender was distinguished within the CASI, separate cutoffs were identified for girls and boys. Sensitivity and specificity (inversely proportional true positive and negative rates), along with positive and negative predictive values (probability of true positive and negative classification), were explored. Combined sensitivity and specificity were expected to be at least 1.5, indicating adequate utility (Power et al., 2013).

Results

Internal Structure

Single-factor CFA models for ADHD ($\chi^2(2) = 1.49$, $p = .48$; CFI = 1, RMSEA = .000, SRMR = .012) and ODD/CD ($\chi^2(5) = 13.72$, $p = .02$; CFI = .991, RMSEA = .063, SRMR = .073) symptoms fit well. Standardized loadings are found in Table 1. Loadings ranged from .67 to .96, all “good” or better (Comrey & Lee, 1992).

Model fit information is found in Table 2. All baseline models for ADHD symptoms fit well, and the model demonstrated up to scalar invariance across all groups. Most baseline models for ODD/CD fit well, but the initial model for White children produced problematic fit. Examination of modification indices revealed a large residual correlation between the “defiance” and “anger” items. The “anger” item was dropped from all further analyses and the final scale score for the following reasons: because these indicators shared a high degree of overlap, because item response theory analyses indicated lower item information for “anger” than for “defiance”, and because internal reliability showed the smallest decrease if “anger” was dropped compared to other items (from .70 to .66). After this, all baseline models fit well, and the scale demonstrated up to scalar invariance across all groups except across race.

Item discrimination and thresholds (for polytomous items) and difficulty (for binary items) are found in Table 1. Item discrimination values (representing the overall ability of the items to detect levels of the latent trait) ranged from 1.79 to 4.47, "very high" or above (1.70; Baker, 2001). Thresholds and difficulty parameters represent the latent trait level needed to have a 50% probability of moving up to the next response option (for polytomous items) or endorsing an item (for binary items). For example, for the item regarding difficulty concentrating, a parent rating a child with a latent score of around –1.14 (more than one standard deviation below the mean) would be equally likely to endorse "Never" or "Sometimes," and a parent rating a child with a latent score around 0.04 (around the average) would be equally likely to respond "Sometimes" or "Often."

Convergent and Discriminant Validity

First, items were averaged to create scale scores, to parallel how these scales are expected to be used in practice and allow for comparability across samples (Comrey & Lee, 1992). Overall means were 2.38 (*SD* = 1.31) for ADHD symptoms and 1.74 (*SD* = 1.25) for ODD/CD. Means for both scales were significantly higher for boys than for girls (both small effects): for ADHD, *d* = 0.48, *t*(438) = 4.23, *p* < .001; for ODD/CD, *d* = 0.27, *t*(438) = 2.73, *p* = .007. The ADHD mean was also higher among children using special education resources, *d* = 0.28, *t*(438) = 2.59, *p* = .01. There were no significant differences across age or race groups.

Correlations are found in Table 3. As expected, both scales were significantly associated with both validation scales in the expected directions (medium to large, positive), grades (negative), and parents being contacted by the school because of negative behavior (positive). Partial correlations showed that, as expected, both scales retained strong relationships with their respective validation scales after controlling for the other scale. Notably, although both scales were uniquely related to negative behavior, only ADHD remained associated with grades. Finally, chi-square tests indicated that the unique association between each scale and its validation counterpart was significantly stronger than the association of the validation scale and the other BHS scale, supporting the distinctiveness of the two scales. Therefore, the convergent and discriminant validity of the scales were supported.

Follow-up analyses examined whether findings were the same using the CASI component scales (inattentiveness and hyperactivity, and oppositional-defiant and conduct-aggressiveness), as well as whether associations remained significant after controlling for demographic variables (gender, age, and race). There were no changes to the primary findings.

Diagnostic Accuracy

Cutoffs, classification accuracy, and total area under the curve (AUC) for each BHS scale predicting CASI clinical cutoffs are found in Table 4. ROC curves are found in Figure 1. All cutoffs except one (for boys, based on oppositional-defiant) met the criteria of a combined value of 1.5 (Power et al., 2013). AUCs ranged from .78 to .87 (fair to good).

For ADHD symptoms, the Youden index identified different clinical cutoffs for boys and girls. For girls, a cutoff around 2 was the most optimal. However, for boys, cutoffs of 2.5 (based on inattentive) and 3.5 (based on hyperactive) were the most optimal. For ODD/CD symptoms, a cutoff of 2 demonstrated the best performance regardless of gender or criterion scale. Overall, scale cutoffs appeared to have higher sensitivity (better true positive rate) for boys and higher specificity (better true negative rate) for girls, with the exception of the ADHD-hyperactivity cutoff.

Discussion

The current study evaluated the psychometric characteristics of two new scales added to the BHS to assess parent-reported externalizing symptoms in children ages 6 to 14, and found good evidence for structural validity, measurement invariance (except across race for ODD/CD symptoms), item discrimination, convergent and discriminant validity, and classification accuracy. The new scales are an important addition to the bhworks electronic screening program.

These new scales represent a substantial contribution to the literature. Externalizing behaviors are particularly important among younger age groups, and detection and intervention still need improvement (Danielson et al., 2018; Ghandour et al., 2019). Comprehensive screeners including both externalizing and internalizing symptoms are especially important because this combination poses an additional risk for outcomes like suicide (Duprey et al., 2020). Therefore, having psychometrically strong scales assessing these behaviors significantly improves an already widely used and psychometrically strong tool, and makes it more appropriate and useful within a school-based, referred setting. In line with the adolescent BHS, the new subscales are quite brief (each subscale is only four items). Because length is often a barrier to use (Connors et al., 2015, 2021), particularly for repeated assessments, this may improve feasibility for use within a referred population if future research indicates the scales are appropriate for progress monitoring. Moreover, the strong psychometric characteristics contrasts with many parent-report behavioral health scales, which may not have published information on structural validity, measurement invariance, or item response theory performance, and have been criticized for potential demographic biases (Cauce, 1995; Piper et al., 2014; Raadal et al., 1994). Given demographic differences such as disparities in externalizing problems across gender (Boyd et al., 2015; Lau et al., 2021), evidence of invariance across most groups and scales can increase confidence that observed differences in scores are not mere measurement artifacts. Moreover, by assessing risk factors that play a role in the development and course of externalizing (e.g., family environment; Stormont, 2002), alongside possible comorbid symptoms like depression, anxiety, and sleep disturbance (Duprey et al., 2020), the BHS provides assessors with additional context that may guide triage decisions for youth in need of more services (Scott & Lewis, 2015).

These findings have implications for school-based behavioral health screening practice. The use of psychometrically strong measures is crucial to improve accuracy in screening and confidence in the findings but remains low in school- and community-based mental health agencies (Bruhn et al., 2014; Connors et al., 2019). School-based behavioral health programs are crucial for improving outcomes among youth, but assessment-based limitations can often prevent them from maximum effectiveness (Connors et al., 2015). These limitations can include factors like poor ease of use, lack of immediate feedback, low interpretability, and poor continuity across programs or sites (Connors et al., 2015; Jensen-Doss & Hawley, 2010). Platforms like bhworks represent advancements in the school-based assessment field intended to address these problems within the broader program of assessment, feedback, intervention, and referral. The use of validated scales on a feasible and scalable electronic platform is a major strength of the bhworks platform.

There are several limitations to the current study. First, there is the exclusive use of parent-report. Although parent and teacher reports are often highly concordant for externalizing behaviors (Bied et al., 2017), additional teacher report scales could improve flexibility and continuity of these scales. Second, the ODD/CD scale had somewhat low internal consistency; this is not surprising for a brief, broad scale (Taber, 2018), and is not highly concerning given the good model fit. More concerning is the lack of scalar invariance across Black and White students, which could lead to biased results. Given the small group sizes, this finding needs to be tested in a second sample but may indicate a need to revise the scale to reduce any bias. Similarly, the scale structure with the item dropped should be validated in a second sample. Third, in invariance testing, some groups could not be examined due to small cell sizes (e.g., ethnicity, racial categories outside of Black and White) or insufficient information (e.g., socioeconomic status). The gender imbalance of the current sample, likely due to higher referral rates for boys (e.g., M. T. Green et al., 1996), may also affect results. Finally, and most importantly, the current investigation examined only the new externalizing scales in a non-normative and geographically restricted sample. Further research will be needed to ensure the entire parent-reported BHS functions well in a variety of samples and settings. Given the referred nature of the current sample, these results may not generalize to other uses (e.g., universal screening), although universal screening is currently being piloted. Similarly, it is possible that parents who declined to participate systematically differed from those who participated in the current study, and this was not possible to investigate.

In conclusion, externalizing symptoms are consequential for a broad spectrum of outcomes ranging from academic failure to divorce and even suicide (Duprey et al., 2020; Klein et al., 2012; Masten et al., 2005; van Lier et al., 2012). Schools are in a particularly advantageous position to evaluate and treat high-risk students, but there is a need for multi-tiered and comprehensive programs that utilize training, intervention and referral systems, and high-quality assessment. The new parent-report externalizing scales on the BHS are a helpful and psychometrically robust contribution within the context of the bhworks platform.

Acknowledgments: A special thank you to all of the participating agencies and individuals for their contributions to this work**.**

References

Atherton, O. E., Ferrer, E., & Robins, R. W. (2018). The development of externalizing symptoms from late childhood through adolescence: A longitudinal study of Mexican-origin youth. *Developmental Psychology, 54*(6), 1135–1147. https://doi.org/10.1037/dev0000489

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. https://eric.ed.gov/?id=ED458219

Bevans, K. B., Diamond, G., & Levy, S. (2012). Screening for adolescents' internalizing symptoms in primary care: Item response theory analysis of the behavior health screen depression, anxiety, and suicidal risk scales. *Journal of Developmental and Behavioral Pediatrics: JDBP*, *33*(4), 283–290. https://doi.org/10.1097/DBP.0b013e31824eaa9a

Bickman, L., Kelley, S. D., Breda, C., de Andrade, A. R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services (Washington, D.C.)*, *62*(12), 1423–1429. https://doi.org/10.1176/appi.ps.002052011

Bied, A., Biederman, J., & Faraone, S. (2017). Parent-based diagnosis of ADHD is as accurate as a teacher-based diagnosis of ADHD. *Postgraduate Medicine*, *129*(3), 375–381. https://doi.org/10.1080/00325481.2017.1288064

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Boat, T. F., & Wu, J. T. (2015). Prevalence of oppositional defiant disorder and conduct disorder. In *Mental Disorders and Disabilities Among Low-Income Children*. National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK332874/

Bowen, N. K., & Masa, R. D. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research*, *6*(2), 229–249. https://doi.org/10.1086/681607

Boyd, A., Van de Velde, S., Vilagut, G., de Graaf, R., O'Neill, S., Florescu, S., Alonso, J., Kovess-Masfety, V., & EU-WMH Investigators. (2015) Gender differences in mental disorders and suicidality in Europe: Results from a large cross-sectional population-based study. *Journal of Affective Disorders, 173,* 245-254. https://doi.org/10.1016/j.jad.2014.11.002.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Publications.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Bruhn, A. L., Woods-Groves, S., & Huddle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K–12 schools. *Education & Treatment of Children*, *37*(4), 611–634. https://doi.org/10.1353/etc.2014.0039

Bruns, E. J., Duong, M. T., Lyon, A. R., Pullmann, M. D., Cook, C. R., Cheney, D., & McCauley, E. (2016). Fostering SMART partnerships to develop an effective continuum of behavioral health services and supports in schools. *The American Journal of Orthopsychiatry*, *86*(2), 156–170. https://doi.org/10.1037/ort0000083

Cauce, A. M. (1995). “Behavior problems in 5- to 11-year-old children from low-income families”: Of norms and cutoffs: Response. *Journal of the American Academy of Child & Adolescent Psychiatry*, *34*(5), 537–538. https://doi.org/10.1097/00004583-199505000-00004

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Psychology Press.

Connors, E. H., Arora, P., Curtis, L., & Stephan, S. H. (2015). Evidence-based assessment in school mental health. *Cognitive and Behavioral Practice*, *22*(1), 60–73. https://doi.org/10.1016/j.cbpra.2014.03.008

Connors, E. H., Douglas, S., Jensen-Doss, A., Landes, S. J., Lewis, C. C., McLeod, B. D., Stanick, C., & Lyon, A. R. (2021). What gets measured gets done: How mental health agencies can leverage measurement-based care for better patient care, clinician supports, and organizational goals. *Administration and Policy in Mental Health and Mental Health Services Research*, *48*(2), 250–265. https://doi.org/10.1007/s10488-020-01063-w

Connors, E. H., Schiffman, J., Stein, K., LeDoux, S., Landsverk, J., & Hoover, S. (2019). Factors associated with community-partnered school behavioral health clinicians' adoption and implementation of evidence-based practices. *Administration and Policy in Mental Health and Mental Health Services Research*, *46*(1), 91–104. https://doi.org/10.1007/s10488-018-0897-3

Danielson, M. L., Bitsko, R. H., Ghandour, R. M., Holbrook, J. R., Kogan, M. D., & Blumberg, S. J. (2018). Prevalence of parent-reported ADHD diagnosis and associated treatment among U.S. children and adolescents, 2016. *Journal of Clinical Child & Adolescent Psychology*, *47*(2), 199–212. https://doi.org/10.1080/15374416.2017.1417860

Diamond, G., Levy, S., Bevans, K. B., Fein, J. A., Wintersteen, M. B., Tien, A., & Creed, T. (2010). Development, validation, and utility of internet-based, behavioral health screen for adolescents. *Pediatrics*, *126*(1), e163-170. https://doi.org/10.1542/peds.2009-3272

Diamond, K. E., & Squires, J. (1993). The role of parental report in the screening and assessment of young children. *Journal of Early Intervention*, *17*(2), 107–115. https://doi.org/10.1177/105381519301700203

Dowdy, E., Ritchey, K., & Kamphaus, R. W. (2010). School-based screening: A population-based approach to inform and monitor children's mental health needs. *School Mental Health*, *2*(4), 166–176. https://doi.org/10.1007/s12310-010-9036-3

Duprey, E. B., Oshri, A., & Liu, S. (2020). Developmental pathways from child maltreatment to adolescent suicide-related behaviors: The internalizing and externalizing comorbidity hypothesis. *Development and Psychopathology*, *32*(3), 945–959. https://doi.org/10.1017/S0954579419000919

Gadow, K. D., & Sprafkin, J. (1994). *Child Symptom Inventories Manual*. Stony Brook, NY: Checkmate Plus.

Ghandour, R. M., Sherman, L. J., Vladutiu, C. J., Ali, M. M., Lynch, S. E., Bitsko, R. H., & Blumberg, S. J. (2019). Prevalence and treatment of depression, anxiety, and conduct problems in US children. *The Journal of Pediatrics*, *206*, 256-267.e3. https://doi.org/10.1016/j.jpeds.2018.09.021

Green, J. G., McLaughlin, K. A., Alegría, M., Costello, E. J., Gruber, M. J., Hoagwood, K., Leaf, P. J., Olin, S., Sampson, N. A., & Kessler, R. C. (2013). School mental health resources and adolescent mental health service use. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(5), 501–510. https://doi.org/10.1016/j.jaac.2013.03.002

Green, M. T., Clopton, J. R., & Pope, A. W. (1996). Understanding Gender Differences in Referral of Children to Mental Health Services. *Journal of Emotional and Behavioral Disorders*, *4*(3), 182–190. https://doi.org/10.1177/106342669600400305

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Husky, M. M., Kaplan, A., McGuire, L., Flynn, L., Chrostowski, C., & Olfson, M. (2011). Identifying adolescents at risk through voluntary school-based mental health screening. *Journal of Adolescence*, *34*(3), 505–511. https://doi.org/10.1016/j.adolescence.2010.05.018

Jensen-Doss, A., & Hawley, K. M. (2010). Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of Clinical Child and Adolescent Psychology : The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *39*(6), 885–896. https://doi.org/10.1080/15374416.2010.517169

Khan, M. R. A., & Brandenburger, T. (2019). *ROCit: Performance Assessment of Binary Classifier with Visualization*. https://cran.r-project.org/web/packages/ROCit/index.html

Lane, K. L., & Walker, H. M. (2015). The connection between assessment and intervention: How can screening lead to better interventions? In *Enduring Issues In Special Education*. Routledge.

Lau, T. W. I., Lim, C. G., Acharryya, S., Lim-Ashworth, N., Tan, Y. R., & Fung, S. S. D. (2021). Gender differences in externalizing and internalizing problems in Singaporean children and adolescents with attention-deficit/hyperactivity disorder. *Child and Adolescent Psychiatry and Mental Health*, *15*(1), 1–11. https://doi.org/10.1186/s13034-021-00356-8

Lavigne, J. V., Cromley, T., Sprafkin, J., & Gadow, K. D. (2009). The Child and Adolescent Symptom Inventory-Progress Monitor: A brief Diagnostic and Statistical Manual of Mental Disorder, 4th edition-referenced Parent-Report Scale for children and adolescents. *Journal of Child and Adolescent Psychopharmacology*, *19*(3), 241–252. https://doi.org/10.1089/cap.2008.052

McMahon, R. J. (1994). Diagnosis, assessment, and treatment of externalizing problems in children: The role of longitudinal data. *Journal of Consulting and Clinical Psychology*, *62*(5), 901–917. https://doi.org/10.1037/0022-006X.62.5.901

National Association of School Psychologists (2016). Ensuring high quality, comprehensive pupil services. NASP. https://www.nasponline.org/Documents/Research%20and%20Policy/ESSA_MTSS_Members.pdf.

Nock, M. K., Kazdin, A. E., Hiripi, E., & Kessler, R. C. (2007). Lifetime prevalence, correlates, and persistence of oppositional defiant disorder: Results from the National Comorbidity Survey Replication. *Journal of Child Psychology and Psychiatry*, *48*(7), 703–713. https://doi.org/10.1111/j.1469-7610.2007.01733.x

Piper, B. J., Gray, H. M., Raber, J., & Birkett, M. A. (2014). Reliability and validity of Brief Problem Monitor, an abbreviated form of the Child Behavior Checklist. *Psychiatry and Clinical Neurosciences*, *68*(10), 759–767. https://doi.org/10.1111/pcn.12188

Power, M., Fell, G., & Wright, M. (2013). Principles for high-quality, high-value testing. *BMJ Evidence-Based Medicine*, *18*(1), 5–10. https://doi.org/10.1136/eb-2012-100645

Raadal, M., Milgrom, P., Cauce, A. M., & Mancl, L. (1994). Behavior problems in 5- to 11-year-old children from low-income families. *Journal of the American Academy of Child & Adolescent Psychiatry*, *33*(7), 1017–1025. https://doi.org/10.1097/00004583-199409000-00013

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(1), 1–36. https://doi.org/10.18637/jss.v048.i02

Ruan-Iu, L., Winston-Lindeboom, P., Rivers, A. S., Atte, T., Tien, A., & Diamond, G. (2022). Validation of the parent report Behavioral Health Screen among elementary school children: A look at internalizing disorders and sleep. Unpublished manuscript.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*, *30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243540

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Byrd Press.

Scott, K., & Lewis, C. C. (2015). Using measurement-based care to enhance any treatment. *Cognitive and Behavioral Practice*, *22*(1), 49–59. https://doi.org/10.1016/j.cbpra.2014.01.010

Severson, H. H., Walker, H. M., Hope-Doolittle, J., Kratochwill, T. R., & Gresham, F. M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues, approaches, emerging innovations, and professional practices. *Journal of School Psychology*, *45*(2), 193–223. https://doi.org/10.1016/j.jsp.2006.11.003

Stiffman, A. R., Hadley-Ives, E., Dore, P., Polgar, M., Horvath, V. E., Striley, C., & Elze, D. (2000). Youths' access to mental health services: The role of providers' training, resource connectivity, and assessment of need. *Mental Health Services Research*, 14.

Stormont, M. (2002). Externalizing behavior problems in young children: Contributing factors and early intervention. *Psychology in the Schools*, *39*(2), 127–138. https://doi.org/10.1002/pits.10025

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Twymon, K., Dorsey, M., Hudson, N., Kane, J., Valleru, J., & Fristad, M. A. (2020). Improving timely access to school-based mental health services: Getting through the referral gate. *Evidence-Based Practice in Child and Adolescent Mental Health*, *5*(2), 118–127. https://doi.org/10.1080/23794925.2020.1784055

Walker, B., Cheney, D., Stage, S., Blum, C., & Horner, R. H. (2005). Schoolwide screening and positive behavior supports: Identifying and supporting students at risk for school failure. *Journal of Positive Behavior Interventions*, *7*(4), 194–204. https://doi.org/10.1177/10983007050070040101

Xu, G., Strathearn, L., Liu, B., Yang, B., & Bao, W. (2018). Twenty-year trends in diagnosed attention-deficit/hyperactivity disorder among US children and adolescents, 1997-2016. *JAMA Network Open*, *1*(4), e181471. https://doi.org/10.1001/jamanetworkopen.2018.1471

Table 1

Confirmatory factor analysis loadings and item response theory parameters

Item	Standardized loadings		Item	Difficulty/threshold	
	ADHD	ODD/CD	discrimination	1	2
Difficulty concentrating	.84		3.32	-1.14	0.04
Impulsivity	.79		2.25	-1.20	0.20
Hyperactivity	.90		3.51	-0.77	0.23
Impairment	.96		4.47	-0.70	0.38
Defiance		.85	2.05	-0.97	
Anger		.85	Dropped from analyses		
Interpersonal aggression		.82	2.59	0.41	
Damaging property		.67	1.79	1.42	
Impairment		.74	2.18	0.18	

Note. ADHD = attention-deficit/hyperactivity disorder symptoms. ODD/CD = oppositional defiant/conduct disorder symptoms. Thresholds and difficulty parameters represent the latent trait level needed to have a 50% probability of moving up to the next response option (for polytomous items) or endorsing an item (for binary items).

Table 2

Model fit information

Group	Model	χ^2	CFI	RMSEA	SRMR
Gender, ADHD	Girls only	2.38	.999	.035	.027
	Boys only	0.32	1	.000	.007
	Configural	2.70	1	.000	.014
	Metric	8.94	.999	.036	.028
	Scalar	5.92	1	.000	.014
Gender, ODD/CD	Girls only	0.27	1	.000	.011
	Boys only	4.08	.993	.061	.056
	Configural	4.11	.999	.011	.040
	Scalar[a]	5.43	1	.000	.043
Age, ADHD	Under 10 only	0.97	1	.000	.012
	10+ only	0.31	1	.000	.008
	Configural	1.27	1	.000	.011
	Metric	3.99	1	.000	.018
	Scalar	5.65	1	.000	.011
Age, ODD/CD	Under 10 only	1.14	1	.000	.034
	10+ only	2.74	.997	.048	.060
	Configural	3.88	1	.000	.044
	Scalar[a]	7.51	.997	.034	.050
Race, ADHD	Black only	0.39	1	.000	.014
	White only	0.63	1	.000	.009
	Configural	1.02	1	.000	.011
	Metric	4.62	1	.000	.020
	Scalar	2.91	1	.000	.012
Race, ODD/CD	Black only	1.03	1	.000	.053
	White only	0.97	1	.000	.035
	Configural	2.00	1	.000	.042
	Scalar[a]	11.83	.986	.075	.058
Special education, ADHD	Yes only	2.18	.999	.027	.030
	No only	0.23	1	.000	.005
	Configural	2.40	1	.000	.012
	Metric	5.83	1	.000	.019
	Scalar	6.05	1	.000	.013
Special education, ODD/CD	Yes only	0.21	1	.000	.022
	No only	5.03	.991	.069	.060
	Configural	5.24	.998	.038	.054
	Scalar[a]	5.34	1	.000	.050

Note. Values for ODD/CD scale reflect models after dropping "anger" item. All χ^2 values were nonsignificant ($p > .05$). When $df > \chi^2$, CFI = 1, RMSEA = 0.

[a] Metric invariance could not be tested due to the dichotomous indicators.

Table 3

Associations with validation measures and other criterion variables

	Simple bivariate association		Controlling for other scale		
	Correlations		Standardized estimates		χ^2 (difference in associations)
Variable	ADHD	ODD/CD	ADHD	ODD/CD	
CASI ADHD	**.74*****	.42***	**.71*****	.07	90.38***
	[.62, .86]	[.39, .60]	[.63, .74]	[.00, .14]	
CASI ODD/CD	.48***	**.67*****	.20***	**.57*****	28.41***
	[.38, .58]	[.56, .78]	[.12, .28]	[.49, .63]	
School behavior	**.49*****	**.48*****	**.31*****	**.30*****	-
	[.41, .58]	[.40, .57]	[.25, .50]	[.23, .29]	
Grades	**-.26*****	**-.22****	**-.21****	**-.10**	-
	[-.37, -.14]	[-.34, -.10]	[-.38, -.07]	[-.24, .04]	

Note. School variables were treated as ordinal. Bolded values indicate hypothesized relationships. Confidence intervals are in brackets. Significant χ^2 values indicate a significant difference between standardized estimates, controlling for the other scale.

Table 4

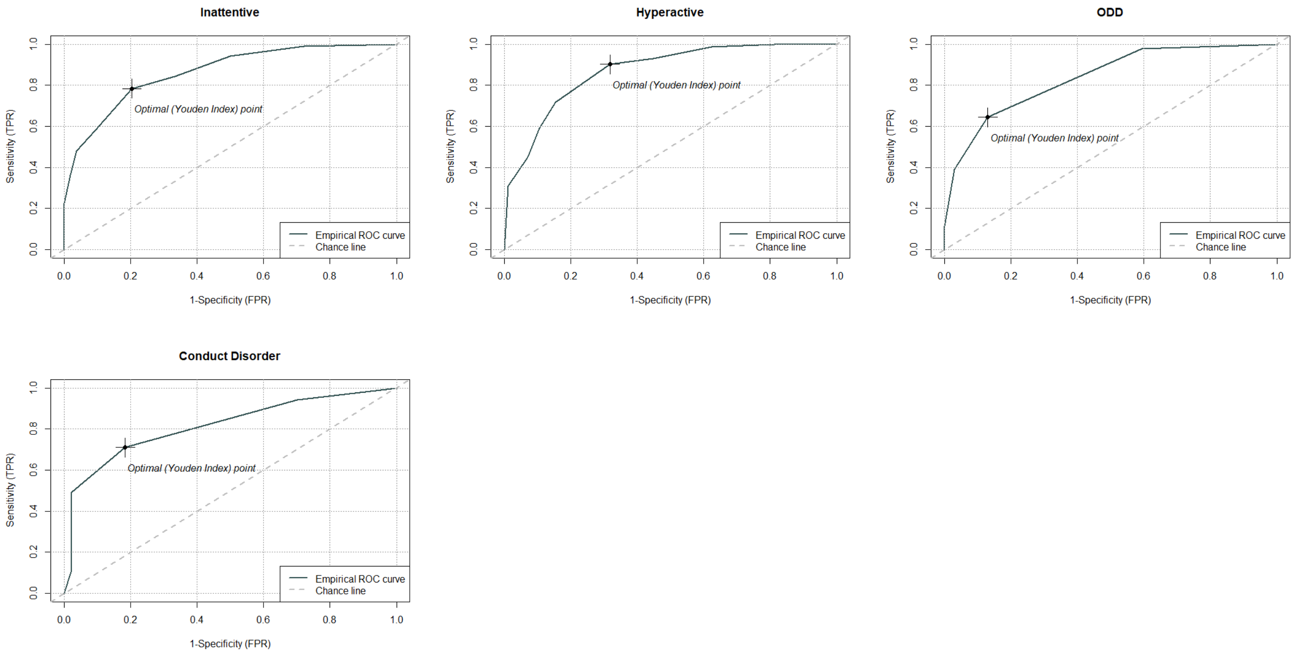
Receiver operator characteristic curve analyses

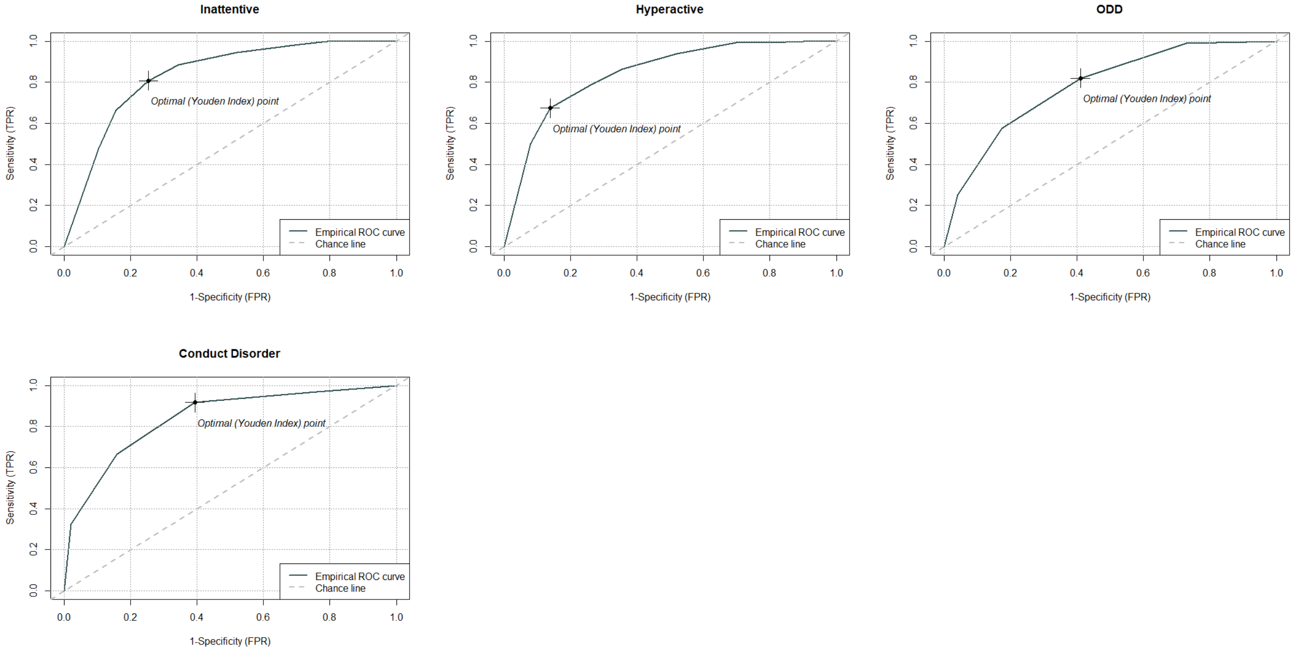
Criterion	ADHD				ODD/CD			
	Inattentive		Hyperactive		Oppositional-defiant		Conduct-aggression	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Cutoff	2.50	2.00	3.50	2.00	2.00	2.00	2.00	2.00
Sensitivity	.85	.78	.67	.90	.82	.64	.92	.71
Specificity	.72	.80	.86	.68	.59	.87	.61	.82
PPV	.79	.88	.81	.70	.56	.86	.54	.75
NPV	.80	.66	.75	.89	.84	.66	.93	.78
Total AUC	.85	.86	.84	.87	.78	.84	.82	.82

Note. ADHD = attention-deficit/hyperactivity disorder symptoms. ODD/CD = oppositional defiant/conduct disorder symptoms.

Figure 1

ROC curves for girls (top) and boys (bottom)




Inattentive
Optimal (Youden Index) point
Empirical ROC curve
Chance line
Sensitivity (TPR)
1-Specificity (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Hyperactive
Optimal (Youden Index) point
Empirical ROC curve
Chance line
Sensitivity (TPR)
1-Specificity (FPR)
ODD
Optimal (Youden Index) point
Empirical ROC curve
Chance line
Sensitivity (TPR)
1-Specificity (FPR)
Conduct Disorder
Optimal (Youden Index) point
Empirical ROC curve
Chance line
Sensitivity (TPR)
1-Specificity (FPR)