The Clinical Assessment of Prosocial Emotions (CAPE): Initial Tests of Reliability and Validity in a Clinic-Referred Sample of Children and Adolescents

Courtney M. Goetz, Taylor A. Miller, and Paul J. Frick
Department of Psychology, Louisiana State University

Recent changes to diagnostic criteria for serious conduct problems in children and adolescents have included the presence of elevated callous-unemotional traits to define etiologically and clinically important subgroups of youth with a conduct problem diagnosis. The Clinical Assessment of Prosocial Emotions (CAPE) is an intensive assessment of the symptoms of this limited prosocial emotions specifier that uses a structured professional judgment method of scoring, which may make it useful in clinical settings when diagnoses may require more information than that provided by behavior rating scales. The present study adds to the limited tests of the CAPE's reliability and validity, using a sample of clinic-referred children ages 6–17 years of age, who were all administered the CAPE by trained clinicians. The mean age of the sample was 10.13 years ($SD = 2.64$); 54% of the sample identified as male and 46% identified as female; and 67% of participants identified as White, 29% identified as Black, and 52% identified as another race/ethnicity (i.e., Asian, Hispanic/Latinx, or other). The findings indicated that CAPE scores demonstrated strong interrater reliability. The scores also were associated with measures of conduct problems and aggression, even when controlling for behavior ratings of callous-unemotional traits. Further, when children with conduct problem diagnoses were divided into groups based on the presence of the limited prosocial emotions specifier from the CAPE, the subgroup with the specifier showed more severe conduct problems and aggression. The results support cautious clinical use of the CAPE, its further development and testing, and research into ways to make its use feasible in many clinical settings.

Public Significance Statement
This study tests the reliability and validity of the Clinical Assessment of Prosocial Emotions, a comprehensive diagnostic assessment of the limited prosocial emotions specifier for conduct disorder. The authors found that the Clinical Assessment of Prosocial Emotions is a reliable and valid tool that adds valuable information not captured in other measures of limited prosocial emotions.

Keywords: limited prosocial emotions, callous-unemotional traits, diagnosis, psychometric properties, Clinical Assessment of Prosocial Emotions

Callous-unemotional (CU) traits, defined by limited remorse, callous lack of empathy, limited concern about performance, and restricted affect, are theorized to represent the affective dimension of psychopathy and a failure in the development of the affective components of conscience (Frick, 2022). Further, CU traits are associated with an earlier onset of more chronic and serious antisocial behavior (see Ritchie et al., 2022 for meta-analysis).

Importantly, while youth with elevated CU traits do respond to some evidence-based treatment of conduct problems, they often still end treatment with more severe behavior problems (Perlstein et al., 2023). In addition, CU traits have been critical to many causal theories for serious conduct problems, as antisocial youth with elevated CU traits display unique genetic, emotional, cognitive, and social characteristics that implicate distinct etiological processes in

the development of their antisocial behavior, as compared to other youth with conduct problems (De Brito et al., 2021). Based on the clinical and etiological significance of CU traits, they have been recently included in latest editions of major diagnostic classification systems as a specifier for diagnoses involving serious conduct problems, such as conduct disorder (CD) in the *Diagnostic and Statistical Manual of Mental Disorders, fifth edition* (*DSM-5*; American Psychiatric Association, 2013) and oppositional defiant (ODD) and conduct-dissocial disorder in the *International Classification of Diseases, 11th edition* (*ICD-11*; World Health Organization, 2019). In both systems, the specifier for those with elevated CU traits is called "with Limited Prosocial Emotions."

Given this relatively recent inclusion in major classification systems used to diagnose mental health problems worldwide, there is a critical need for methods of assessing this construct that can be used in a wide variety of research and applied settings. Much of the research that has been conducted on CU traits to date have relied on behavior rating scales for the assessment of the construct (Andershed et al., 2002; Frick & Hare, 2002; Kimonis et al., 2008; López-Romero et al., 2012; Lynam, 1997). These rating scales are easy and time-efficient to administer, making them very useful for many research purposes. However, while rating scales can be useful for screening people at risk for a mental health problem, more comprehensive assessments are typically needed to make clinical diagnoses. That is, behavior rating scales have persons rate themselves or have informants (e.g., parents and teachers) rate the person being evaluated on the key indicators of CU traits. In many clinical assessments, it would be important for clinicians to follow-up to determine what led to the ratings in order to be sure that the questions were understood by the informant and to evaluate whether or not the informant was carefully considering the questions and attempting to rate them accurately. Further, while behavior rating scales all have some method of anchoring the ratings of the items (e.g., frequency of the trait being assessed, how typical it is), the LPE specifier requires that the CU traits are not transient (i.e., displayed persistently over at least 12 months) and are displayed in multiple relationships and settings (American Psychiatric Association, 2013). When relying on behavior ratings scales, it is often difficult to obtain such detailed information on whether or not the indicators reflect the individual's typical pattern of interpersonal or emotional functioning and not just occasional occurrences in some situations. Finally, it is widely accepted that the assessment of children's social, emotional, and behavioral functioning requires getting information from multiple informants (De Los Reyes & Epkins, 2023). However, it is also well-established that information from different informants is often discordant (e.g., a child rating themselves as being empathetic, while the parent rates them as being callous and uncaring) and it is difficult for a clinician to determine how to integrate such discrepant information when making a diagnosis (Makol et al., 2020).

Thus, there is a critical need for reliable and valid methods for assessing the CU traits that form the LPE specifier that can aid a clinician in obtaining the detailed information needed to make a diagnosis. One clinician rating that can be used for this purpose is the Psychopathy Checklist–Youth Version (PCL-YV; Forth et al., 2003). The PCL-YV is a clinician-rating system that assesses constructs similar to those included in the LPE specifier. Trained clinicians obtain detailed information to rate items as being 0 (*definitely not present*), 1 (*somewhat present*), or 2 (*definitely present*). Further, there has been substantial support for the reliability

and validity of the scores from this measure (Forth et al., 2016). However, CU traits are assessed as part of the broader dimension of psychopathy and, as a result, only four of the 21 items on the PCL-YV are relevant for the LPE specifier. Further, the items on the PCL-YV that assess CU traits were not designed to specifically assess the LPE specifier symptoms. For example, the LPE symptom of "Failure to put forth effort in important activities" is best assessed by the PCL-YV item of "Unable to take responsibility for actions," which is not an exact correspondence. Finally, the PCL-YV was developed for and has been primarily tested for adolescents ages 12–18 years of age in forensic settings (Forth et al., 2016), making its usefulness for children prior to adolescence and its validity in mental health settings less clear.

To address these limitations in the existing methods for assessing the LPE specifier, the Clinical Assessment of Prosocial Emotions (CAPE; Frick, 2013) was developed. The CAPE is a structured clinical assessment designed specifically to assess the four symptoms of the LPE specifier for children and adolescents ages 3–21 years. The CAPE is designed to be administered by experienced clinicians who have been trained in both the psychological assessment of children and adolescents and in the specific administration procedures of the CAPE. The CAPE provides comprehensive descriptions of characteristics associated with each of the four LPE symptoms and the clinician rates the person being assessed on each symptom using a 0 (*not descriptive or mildly descriptive*), 1 (*moderately descriptive*), and 2 (*highly descriptive*) scale. The CAPE uses the structured professional judgment approach to make these ratings (Lawing et al., 2017). That is, semistructured interviews are provided and are to be administered to the person being assessed and at least one additional informant who knows the person being assessed well (e.g., parent, teacher, case manager). These interviews provide questions that help the clinician gain information needed to make the ratings for each symptom but they allow for substantial leeway for the clinician to follow-up with questions designed to gain more information on why the informant answered the question in the way they did. This format also allows the clinician to adjust the follow-up questions to the educational and developmental level of the person being assessed. The interviews also provide standard prompts to assess the duration and pervasiveness of each symptom. The clinician then considers information from these interviews, as well any additional clinical information that is available, to make the ratings on the four LPE symptoms.

To date, there have been three published tests of the reliability and validity of the CAPE. First, Molinuevo et al. (2020) tested the reliability and validity of CAPE in a sample of 72 males ages 14–22 years incarcerated in two Spanish detention centers. They reported moderate interrater reliability for the diagnosis of the LPE specifier (Cohen's $\kappa = .66$), as well as convergent validity, with those meeting the LPE specifier cutoff on the CAPE showing higher scores on rating scales measure of CU traits and scoring higher on the CU dimension (i.e., the affective facet) of the PCL-YV. Further, those scoring above the LPE cutoff on the CAPE showed higher self-report and teacher ratings of antisocial behavior than those who did not meet this cutoff but not higher scores on ratings of internalizing problems. While promising, this study was limited in that, while the sample were all detained males, no conduct problem diagnosis was made. As a result, the actual *DSM-5* and *ICD-11* criteria for the specifier that requires a conduct problem diagnosis were not

actually tested. Further, there were no tests as to whether the time-intensive CAPE scores predicted outcomes over and above more time-efficient rating scales. Such tests of incremental validity are crucial given the amount of training and administration time required by the CAPE. Finally, the sample consisted only of adolescents and young adults, leaving the CAPE's reliability and validity for children untested.

Many of these limitations were addressed in a study by Hawes et al. (2020), who examined the psychometric properties of the CAPE in a sample of 82 clinic-referred children and adolescents (ages 3–15 years). They reported that symptom counts from the CAPE were positively associated with measures of conduct problems using parent report on structured interviews. Further, CAPE scores predicted parent-rated empathy and teacher-rated proactive aggression, even when controlling for parental ratings on a measure of CU traits. Finally, children with a diagnosis of either ODD or CD based on the structured interview, who also had two or more symptoms rated as present (e.g., who met the LPE criteria), scored significantly lower on parent-rated empathy than the group with ODD/CD but who did not meet the LPE criteria. However, the two ODD/CD groups did not differ significantly on measures of conduct problem severity or aggression. It is important to note that Hawes et al. (2020) did not test the interrater reliability of the CAPE diagnosis and the diagnosis was based solely on parental report, rather utilizing multiple informants as specified by the standard instructions for scoring the CAPE.

Finally, a study by Neo et al. (2023) further tested the psychometric properties of the CAPE using a sample of 232 clinic-referred children aged 2–8 years. The authors reported that the CAPE LPE symptom scores had good interrater reliability and internal consistency and, as in the Hawes et al. (2020) study, they predicted conduct problem severity and aggression over and above a parent rating of CU traits. Also, consistent with past research, children diagnosed with LPE using the CAPE had more severe externalizing problems and lower empathy than children without LPE, both overall in the sample and when the sample was limited to those with an ODD/CD diagnosis. Two unique findings from this test of the validity of the CAPE were that the CAPE provided incremental validity over parent ratings of CU traits in predicting a laboratory measure of emotion recognition accuracy and children diagnosed with the LPE specifier using the CAPE showed reduced responsiveness to an evidence-based treatment. Of note, like the Hawes et al. study, this study also relied solely on the parent interview to score the CAPE.

Thus, there are a few tests of the reliability and validity of the CAPE that have provided promising results for its reliability and validity but more tests in different samples are needed. Further, many of these past tests have been limited by failing to use the multiinformant assessment procedure that considers both the parent and youth report. To address this limitation and to provide further tests of the CAPE's reliability and validity, we administered the CAPE as part of a standard diagnostic procedure for children and adolescents (ages 6–17) who were referred to an outpatient mental health clinic for a psychological evaluation due to behavioral, emotional, social, or learning problems. In the present study, the CAPE was scored based on interviews with the child and at least one parent. A majority of the interviews were observed and the CAPE was scored independently by the observer to test the interrater reliability of the CAPE scores. Then, the association of symptoms from the CAPE in predicting conduct problem severity and aggression, both alone and controlling for a behavior rating of CU traits, was tested. Finally, children diagnosed with either ODD or CD were divided into those who either did or did not meet the LPE criteria based on the CAPE and these groups were compared to each other and to children without a conduct problem diagnosis on the measures of conduct problem severity and aggression.

Method

Transparency and Openness

This study is a retrospective chart review of a sample of consecutive referrals to a mental health clinic, and its analysis was not preregistered. Data, analysis code, and research materials are not available. Data were analyzed using SPSS, Version 28.0.0.0 (IBM Corp, 2021).

Participants

The current sample was consecutive referrals to an outpatient mental health training clinic at a university in the Southeastern United States. The clinic provides comprehensive psychological evaluations for children ages 6–17 with emotional, behavioral, social, or learning problems. Participants with an IQ below 70 or with a diagnosis of autism spectrum disorder were not included in analyses. This resulted in a sample of 97 children with a mean age of 10.13 ($SD = 2.64$). Fifty-four percent of the sample identified as male and 46% identified as female. Sixty-seven percent of participants identified as White, 28% identified as Black, and 52% identified as another race/ethnicity (i.e., Asian, Hispanic/Latinx, or other). The Full-Scale Intelligence Quotient of the sample ranged from 70 to 129 ($M = 94.13$, $SD = 12.09$). Finally, the least socioeconomically disadvantaged quartile of the sample lived in a 2020 census tract with approximately 4% of the population living below the poverty line, and the most disadvantaged quartile lived in a census tract with approximately 19% of the population living under the poverty line (national average = 12%; Creamer et al., 2022).

Procedure

Institutional review board approval for the use of de-identified clinic files to be used in research was obtained. Prior to testing, parents gave informed consent for the clinical evaluation and for the use of the information in research and the children gave assent for the data to be used in research. Both parents and children were informed that allowing the information to be used in research would not alter any clinical services they received. The measures used in the present study were administered as part of a comprehensive battery during an all-day assessment. The entire battery was administered in a standardized order by graduate students trained in the assessment procedures. Because measures were administered as part of a clinical assessment, the order of measures was determined so that measures that were considered more important for diagnoses were given first. For children below age 9, or for any child that showed any reading difficulties, behavior rating scales were read out loud to the participant. Following the testing, the child was given a clinical diagnosis and a report was written that summarized the testing information. This report, which included recommendations for treatment, was reviewed with the child and parent on a separate day.

Measures

Clinical Assessment of Prosocial Emotions

The CAPE (Frick, 2013) is a clinician-rating system that assesses the four diagnostic indicators of LPE (e.g., lack of remorse or guilt, callous lack of empathy, unconcerned about performance, shallow or deficient affect). To score the CAPE, the clinician rates how well each indicator fits the client on a 3-point scale (0 = *not descriptive or mildly descriptive*, 1 = *moderately descriptive*, 2 = *highly descriptive*). Prototypes for each indicator are provided in order to guide the clinician. Next, the clinician records how many indicators of CU traits were rated "highly descriptive," providing a continuous rating ranging from 0 to 4. When this value is 2 or greater (i.e., two or more symptoms are rated as "highly descriptive"), severity is considered sufficient to reach the diagnostic threshold for applying the *DSM-5* or *ICD-11* LPE specifier (American Psychiatric Association, 2013; World Health Organization, 2019).

The clinicians administering the CAPE were all students in a doctoral program for clinical or school psychology, they all had passed classes on child mental health problems and psychological testing, they were trained in CAPE administration as outlined in the manual (Frick, 2013), and they were supervised in their administration by a licensed psychologist, who was also trained in the administration of the CAPE. The scoring of the CAPE is based on the clinician's weighing of all available clinical information. It is aided by semistructured interviews with both a primary custodial parent and the child. These interviews include 2–3 stem questions (nine total) related to each LPE symptoms (e.g., "Does _____ seem to feel bad or guilty if he/she does something wrong or if he/she hurts someone?") that are answered as either "yes" or "no." The clinician can follow-up with questions to determine the reason for the answer (e.g., "please give me some examples," "what makes you say that"). If the stem questions are answered in a way that might suggest symptom presence, they are followed by a series of questions to determine how persistent and pervasive the trait might be (e.g., "Is this how he/she is most of the time and with most people?") and other questions that would help in scoring the CAPE (e.g., "Does he/she only feel bad or guilty if he/she is caught doing something wrong and is going to be trouble?"). Following the administration of both interviews, the clinician scores the CAPE using their professional judgment after weighing all information. For 72 (74%) of the cases, a second trained clinician observed both interviews and independently rated the CAPE symptoms in order to test the interrater reliability of the scores. Collection of reliability data began in the third year of the clinical service and was collected on consecutive cases afterward.

Computerized Diagnostic Interview Schedule for Children

The computerized version of the National Institute of Mental Health Diagnostic Schedule (Computerized Diagnostic Interview Schedule for Children [C-DISC]; Shaffer et al., 2000) is a highly structured diagnostic interview that assesses mental health disorders in children and adolescent, using questions designed to assess criteria from the *DSM-5*. The structured interview was administered to parents and to the child or adolescent, if they were over 9 years old. This was based on concerns about the reliability of structured interviews in the assessment of children below this age (Frick et al., 2020). For the present study, we used the parent and child report of ODD and CD symptoms. The CD and combined ODD/CD symptoms were significantly correlated between parent and child report at $r = .42$ ($p < .001$) and $r = .47$ ($p \leq .001$), respectively. The ODD symptoms were not significantly correlated between child and parent report ($r = .23$, $p = .088$). The C-DISC was also administered by doctoral students trained in administration procedures and supervised by a licensed psychologist. The clinician administering the C-DISC interview was also the clinician who was scoring the CAPE. Research has demonstrated that the ODD and CD symptoms assessed in the C-DISC interviews are highly related to clinical diagnoses of these disorders (Rolon-Arroyo et al., 2014). As recommended by Piacentini et al. (1992), a symptom was considered present if reported by either the parent or child.

Inventory of Callous-Unemotional Traits

The Inventory of Callous-Unemotional Traits (ICU; Frick, 2004) is a widely used measure of CU traits (Cardinale & Marsh, 2020; Deng et al., 2019). The ICU consists of 24 items assessing the four diagnostic indicators of LPE. For each symptom, three positively worded items (worded in the callous direction) and three negatively worded items (worded in the prosocial dimension) are rated on a Likert scale ranging from 0 = *not at all true* to 3 = *definitely true*. Items worded in the prosocial direction were inversely coded prior to summing items to create a composite score. Although past work has found that this measure of CU traits can be broken into subscales, the total score was used in the present study due to consistent findings of a general factor that accounts for a large portion of the variance in the subscales and that is associated negatively with empathy and positively with aggression in a variety of child, adolescent, and adult samples (Cardinale & Marsh, 2020; Ray & Frick, 2020). The ICU was completed by both parent (Cronbach's $\alpha = .91$) and children (Cronbach's $\alpha = .78$) and reports from these two informants were correlated at $r = .40$ ($p < .001$). Due to findings that different ICU informant versions are better at predicting relevant outcomes in different age groups (see Matlasz et al., 2022), parent and child ratings were combined by taking the highest score on each item (see Matlasz et al., 2022 for evidence supporting the validity of this method of scoring).

Peer Conflict Scale

The Peer Conflict Scale (PCS; Marsee et al., 2011) is a 40-item measure that includes items describing 20 items reflecting both relational and physical proactive aggression (e.g., "I start fights to get what I want") and 20 items assessing both relational and physical reactive physical aggression (e.g., "When someone hurts me, I end up getting into a fight") aggression. Items are rated on a 4-point scale ranging from 0 (*not at all true*) to 3 (*definitely true*) and summed to form a continuous measures of aggression severity. These aggression subscales of the PCS have been correlated with laboratory measures of aggression and other indicators of aggression and violence (Muñoz et al., 2008), and the two subscales have been found to be associated with differences in several hypothesized correlates. Specifically, the Reactive Aggression scale has been uniquely associated with reaction to provocation and poor emotion regulation, while the Proactive Aggression scale has been uniquely associated with CU traits and biased outcome expectations for aggressive behavior in samples of adolescents and young adults

(Marsee et al., 2011; Marsee & Frick, 2007; Muñoz et al., 2008; Vagos et al., 2021). The PCS was completed by both parent (proactive Cronbach's α = .93; reactive Cronbach's α = .95) and child (proactive Cronbach's α = .91; reactive Cronbach's α = .91). Across these informants, the ratings on the Proactive subscale were correlated at $r = .30$ ($p = .024$) but the correlations between parent and child ratings on the Reactive subscale did not reach statistical significance ($r = .23$, $p = .092$). As with other measures, composite scores for the Proactive and Reactive Aggression subscales were created by taking the highest rating on each item across informants.

Behavior Assessment Schedule for Children–Third Edition

The Behavior Assessment Schedule for Children–Third Edition (BASC-3; Reynolds & Kamphaus, 2015) is a comprehensive measure of adaptive and problem behaviors that provides norm-referenced T scores based on a large norming sample collected to match the U.S. Census on major demographic characteristics. The parents' ratings on the Conduct Problems and Aggression subscales of the BASC-3 Parent Rating Scale were used. The Conduct Problems subscale measures broad range of conduct problems including lying, rule breaking, and stealing. This subscale has shown good to excellent test–retest reliability over a period of 7–70 days (corrected $r = .78$–.91) in the norming sample and was significantly correlated with other measures of conduct problems, such as the Child Behavior Checklist Rule-Breaking Behavior subscale ($r = .61$–.77; Reynolds & Kamphaus, 2015). The Aggression subscale measures overt conduct problems including such items as threatening others, bullying, and hitting others. This subscale has also showed good to excellent test–retest reliability in the normative sample (corrected $r = .83$–.90) and was significantly associated with other measures of overt conduct problems, such as the Child Behavior Checklist Aggressive Behavior subscale ($r = .66$–.72; Reynolds & Kamphaus, 2015).

Data Analysis

There were very few instances of missing data on various study measures. These ranged from three missing responses on the PCS to five missing responses on the C-DISC. Missing data occurred due to (a) the measure not being given to the family or (b) data loss due to computer error. In these cases, participants were included in all analyses except for those including missing responses.

Several indicators of the interrater reliability of CAPE scores were obtained. Intraclass correlations (ICC) using a single-rating, consistency, two-way mixed-effects model were calculated to determine the associations between the primary clinician and the reliability clinician for each LPE symptom (rated 0, 1, 2) and for the sum of the number of symptoms rated as being 2. The ICC was used as an estimate of reliability, given that it considers both the rank order of ratings, as well as the consistency in the absolute value of the ratings (Shrout & Fleiss, 1979). Cohen's κ was also calculated to test the level of agreement on whether each LPE symptom indicator met the threshold of "highly descriptive" or not and whether the child met the diagnostic threshold (i.e., two or more LPE symptoms are present) or not across raters. Cohen's κ was used as an estimate of agreement, since it controls for the level of agreement that would be expected by chance (Fleiss & Cohen, 1973).

To test the validity of the CAPE symptom scores, zero-order correlations were obtained between the CAPE symptom count and the ICU (i.e., convergent validity) and between the CAPE symptom count and measures of conduct problems and aggression (i.e., construct validity). Next, the incremental validity of the CAPE relative to the ICU was assessed using multiple regression analyses. For each measure of conduct problems and aggression, a hierarchical linear regression analysis was conducted with the ICU entered at the first step and then the CAPE symptom count added in the second step to determine the amount of variance in the validator variables accounted for by the CAPE symptoms over and above the ICU. Finally, to test the validity of the LPE specifier based on the CAPE, participants were separated into three groups based on their diagnostic status. The first group contained those who did not meet for ODD/CD based on the DISC or the LPE based on the CAPE (i.e., control group), the second group were those who met criteria for ODD/CD on the DISC but did not meet the threshold for LPE based on the CAPE (i.e., ODD/CD only), and the final group were those who met criteria for both ODD/CD and LPE (i.e., ODD/CD + LPE). One-way analyses of variance (ANOVAs) were then performed to compare diagnostic groups based on the validator variables and, for any significant ANOVA, pairwise comparisons (Tukey's) were used to compare groups to determine which groups differed significantly from each other on the measure of conduct problems or aggression.

Results

The results of the tests of the interrater reliability of the CAPE scores are provided in Table 1. The ICC for individual symptom scores were generally in the good (callous/lack of empathy, ICC = .73, 95% CI [.60, .82], $p < .001$) to excellent (unconcerned about performance; ICC = .82, 95% CI [73, 89], $p < .001$) range. The reliability for the total symptom count was excellent (ICC = .85, 95% CI [.76, .90], $p < .001$). Cohen's κ for the presence of individual symptoms ranged from moderate (callous/lack of empathy, κ = .51, SE = .08, $p < .001$) to good (unconcerned about performance, κ = .72, SE = .08, $p < .001$). The κ for

Table 1
Interrater Reliability of CAPE Scores (n = 72)

CAPE score	Cohen's κ	SE	ICC	95% CI
Lack of remorse/guilt	.58***	.08	.77***	[.66, .85]
Callous/lack of empathy	.51***	.09	.73***	[.60, .82]
Unconcerned about performance	.72***	.09	.82***	[.73, .89]
Shallow/deficient affect	.71***	.08	.77***	[.66, .85]
Total symptom count			.85***	[.76, .90]
Meets diagnostic threshold for LPE	.84***	.11		

Note. Cohen's κ is the level of agreement on whether the symptom meets the threshold of "2" or not by the primary rater and reliability rater or the agreement on whether the diagnostic threshold of two or more symptoms present is met by the primary rater and reliability rater; CAPE = Clinical Assessment of Prosocial Emotions; SE = standard error; ICC = intraclass correlations between the primary rater and reliability rater; using a single-rating, consistency, two-way mixed-effects model CI = confidence interval; LPE = limited prosocial emotions.
*** $p < .001$.

Table 2
Correlations Among Main Study Variables

Study variable	M	SD	1	2	3	4	5	6	7
1. CAPE symptom count	0.47	1.01	—						
2. DISC ODD symptoms[a]	1.86	2.17	.33** (n = 92)	—					
3. DISC CD symptoms[a]	1.13	1.73	.49*** (92)	.77*** (92)	—				
4. DISC ODD/CD symptoms[a]	2.99	3.67	.43*** (92)	.95*** (92)	.93*** (92)	—			
5. ICU[b]	35.98	12.01	.47*** (97)	.59*** (92)	.51*** (92)	.59*** (92)	—		
6. PCS proactive aggression[b]	5.27	8.14	.60*** (94)	.41*** (89)	.56*** (89)	.50*** (89)	.52*** (94)	—	
7. PCS reactive aggression[b]	9.72	10.97	.50*** (94)	.36*** (89)	.46*** (89)	.43*** (89)	.51*** (94)	.89*** (94)	—

Note. CAPE = Clinical Assessment of Prosocial Emotions; DISC = Diagnostic Interview Schedule for Children; ODD = oppositional defiant disorder; CD = conduct disorder; ICU = Inventory of Callous-Unemotional Traits; PCS = Peer Conflict Scale.
[a] Denotes composite score based on whether the symptom was present based on either parent or youth report. [b] Denotes composite based on highest score from either parent or youth report.
** $p < .01$. *** $p < .001$.

agreement on meeting the LPE diagnostic threshold was excellent ($\kappa = .84$, $SE = .11$, $p < .001$).

Zero-order correlations of study variables with several demographic variables (i.e., age, race/ethnicity, gender, census tract poverty level) were tested and only poverty level showed a significant correlation with CAPE symptom count ($r = .23$, $p = .031$). However, poverty level was not significantly correlated with any of the outcome variables and, as a result, was not included as a control variable in other analyses. The correlations among the main study variables are provided in Table 2. CAPE symptom count was significantly correlated with the ICU total score ($r = .47$, $p < .001$), as well as with all of the measures of conduct problems and aggression.

The results of the multiple regression analyses testing the incremental association of the CAPE symptoms with the measures of conduct problems and aggression, after controlling for the ICU total scores, are summarized in Table 3. CAPE symptoms remained associated with most outcome measures, after controlling for ICU. The two exceptions were that the LPE symptoms were not independently associated with the number of ODD symptoms or the number of ODD/CD symptoms.[1] Importantly, in the analyses in which the CAPE symptoms were associated with outcomes after controlling for the ICU, the CAPE accounted for between 7% (for CD symptoms and parent ratings of conduct problems) and 16% (for proactive aggression) of additional variance in the outcomes.

The results of the one-way ANOVAs comparing diagnostic groups (control [$n = 70–75$], ODD/CD [$n = 15–16$], ODD/CD + LPE [$n = 6$]) are presented in Table 4. The one-way ANOVA's were significant for each dependent variable, with effect sizes ranging from $\eta^2 = .16$ for the measure of reactive aggression to $\eta^2 = .72$ for the measure of ODD/CD symptoms. Most importantly, those children with an ODD or CD diagnosis and who met criteria for the LPE specifier based on the CAPE show significantly higher levels of conduct problems or aggression than those who only met criteria for ODD/CD on five of the seven dependent measures that assessed conduct problem severity or aggression. Participants with ODD/CD and LPE did not differ from the ODD/CD only group on their number of ODD symptoms or on the BASC-3 Conduct Problems scale, although both groups were significantly different than controls. Unexpectedly, ICU scores in the ODD/CD with LPE and ODD/CD only groups did not differ.

Discussion

The current findings provide some encouraging preliminary support for the reliability and validity of the CAPE as a clinical assessment for the limited prosocial emotions specifier, which is now part of the diagnostic criteria for conduct disorder in the *DSM-5* and for oppositional defiant and conduct-dissocial disorder in the *ICD-11*. Of critical importance, the CAPE was designed to be a clinician-rated measure that guides the clinician through a comprehensive assessment of each of the indicators of the LPE specifier. While providing some guidance and structure, it also allows the clinician to obtain detailed information needed to rate the symptoms in an individualized manner and it requires the clinician to subjectively weigh information from different sources in making the ratings. Thus, testing whether the training and degree of guidance provided by the CAPE is sufficient for this to be done reliably is a critical test. The current findings, along with those provided by Molinuevo et al. (2020) in sample of detained adolescents and Neo et al. (2023) in sample of clinic-referred young children, support such interrater reliability.

The clearest indication of this reliability is that the Cohen's κ for the interrater agreement for a child meeting the threshold for the LPE criteria across informants was quite high (i.e., Cohen's $\kappa = .84$). This is the critical decision for which the CAPE was designed. There was also some support that individual symptoms could be assessed reliably as well, although the reliability for the symptoms was substantially lower than that found for the full diagnosis. One consistent finding across our study and previous studies assessing interrater reliability on individual symptoms is that the symptom of "lack of remorse/guilt" tends to be the symptom that shows lower interrater reliability than other CAPE symptoms (Molinuevo et al., 2020; Neo et al., 2023). It seems that informants have difficulty determining if the child is truly remorseful after misdeeds or if any displays of guilt are done solely

---

[1] CAPE symptoms predicted CD symptoms and proactive and reactive aggression. Further, CD symptoms were also related to both measures of aggression. Thus, an interesting question that was not included in our main hypotheses was whether CAPE symptoms predicted aggression when controlling for CD symptoms. When CD and CAPE symptoms were included in a multiple regression as predictors, CAPE symptoms remained associated with both proactive (standardized $\beta = .47$; $p < .001$) and reactive aggression (standardized $\beta = .40$; $p < .001$) after controlling for CD symptoms.

Table 3

Incremental Validity of the CAPE Symptoms Over Parent and Child Ratings on the ICU

Predictor	Unstd. B	SE	Std. B	R^2	ΔR^2
DISC ODD symptom[a] ($N = 92$)					
Step 1				.35***	
Constant	−1.80	.56			
ICU	.10	.02	.59***		
Step 2				.35***	.001
Constant	−1.71	.60			
ICU	.10	.02	.57***		
CAPE	.10	.22	.04		
DISC CD symptom[a] ($N = 92$)					
Step 1				.27***	
Constant	−1.40	.47			
ICU	.07	.01	.51***		
Step 2				.34***	.07**
Constant	−.88	.48			
ICU	.05	.01	.36***		
CAPE	.54	.17	.31**		
DISC CD/ODD symptom[a] ($N = 92$)					
Step 1				.35***	
Constant	−3.20	.94			
ICU	.17	.03	.59***		
Step 2				.36***	.02
Constant	−2.59	.99			
ICU	.15	.03	.51***		
CAPE	.63	.36	.17		
PCS proactive aggression[b] ($N = 94$)					
Step 1				.27***	
Constant	−6.75	2.18			
ICU	.34	.06	.52***		
Step 2				.43***	.16***
Constant	−3.54	2.03			
ICU	.20	.06	.31***		
CAPE	3.65	.71	.46***		
PCS reactive aggression[b] ($N = 94$)					
Step 1				.26***	
Constant	−6.10	2.96			
ICU	.45	.08	.51***		
Step 2				.35***	.09***
Constant	−2.89	2.94			
ICU	.31	.09	.35***		
CAPE	3.64	1.03	.34***		
BASC-3 Parent Rating Scales—Conduct Problems subscale ($N = 93$)					
Step 1				.34***	
Constant	32.58	3.88			
ICU	.70	.10	.58***		
Step 2				.40***	.07**
Constant	36.44	3.90			
ICU	.53	.11	.44***		
CAPE	4.29	1.35	.29**		
BASC-3 Parent Rating Scales—Aggression subscale ($N = 93$)					
Step 1				.29***	
Constant	35.00	3.54			
ICU	.58	.09	.54***		

(table continues)

Table 3 (*continued*)

Predictor	Unstd. B	SE	Std. B	R^2	ΔR^2
Step 2				.41***	.11***
Constant	39.43	3.44			
ICU	.39	.10	.36***		
CAPE	4.94	1.20	.38***		

Note. CAPE = Clinical Assessment of Prosocial Emotions; ICU = Inventory of Callous-Unemotional Traits; Unstd. = unstandardized; SE = standard error; Std. = standardized; DISC = Diagnostic Interview Schedule for Children; ODD = oppositional defiant disorder; CD = conduct disorder; PCS = Peer Conflict Scale; BASC = Behavior Assessment System for Children.
[a] Denotes composite score based on whether the symptom was present based on either parent or youth report. [b] Denotes composite based on highest score from either parent or youth report.
** $p < .01$. *** $p < .001$.

to avoid punishment. In our study, the "callous lack of empathy" symptom also had noticeably lower reliability but this was not consistent with past research. Further research into the reliability and validity of individual symptoms can be used to enhance the training of raters on the CAPE.

Also consistent with past work, we found support that the CAPE scores were substantially correlated with other measures CU traits. Molinuevo and colleagues reported that the CAPE scores were substantially correlated with behavior ratings of CU traits (i.e., the ICU) and another clinician rating of CU traits (i.e., PCL-YV). In the present study, we showed that scores from the CAPE were also substantially correlated with ratings of CU traits on the ICU. Also, consistent with past research (Hawes et al., 2020; Molinuevo et al., 2020; Neo et al., 2023), we found that scores from the CAPE were correlated with measures of conduct problems and aggression. Of note, we also demonstrated the CAPE scores were associated with aggression, even when controlling for CD symptoms. These results are important tests of the CAPE's validity, given that one reason that the LPE specifier was added to diagnostic classification was research showing that CU traits were associated with a more severe and aggressive pattern of antisocial behavior (Frick et al., 2014).

An important test of the CAPE's validity is whether it predicts clinically important outcomes over the prediction afforded by rating scale measures of CU traits. Such tests of incremental validity are critical because of the training and supervision required for clinicians to administer the CAPE and the amount of time required to administer and score the CAPE, which only assesses a single clinical construct. Our findings support the incremental utility of the CAPE symptom scores because they remained significantly associated with five of the seven measures of conduct problems and aggression after controlling for the scores on the ICU. In fact, the CAPE accounted for 16% of additional variance in the measure of proactive aggression. These findings provide strong support that the CAPE is providing clinically important information that may not be captured by behavior ratings scales, which is supported by the results of Hawes et al. (2020) who reported that CAPE

Table 4

One-Way ANOVA Comparing Diagnostic Groups on Validators

Study variable	F (df)	η^2	M (SD) (N)		
			Control	ODD/CD only	ODD/CD and LPE
DISC ODD symptoms[a]	84.66*** (2, 89)	.66	.89 (1.21)$_a$ (70)	4.69 (1.45)$_b$ (16)	5.67 (1.75)$_b$ (6)
DISC CD symptoms[a]	81.49*** (2, 89)	.65	0.43 (0.67)$_a$ (70)	2.63 (1.89)$_b$ (16)	5.33 (1.51)$_c$ (6)
DISC CD/ODD symptoms[a]	113.73*** (2, 89)	.72	1.31 (1.69)$_a$ (70)	7.31 (2.65)$_b$ (16)	11.00 (2.90)$_c$ (6)
ICU[b]	17.19**** (2, 94)	.27	32.19 (10.99)$_a$ (75)	44.50 (9.63)$_b$ (16)	52.83 (8.80)$_b$ (6)
PCS proactive aggression[b]	18.66*** (2, 91)	.29	3.55 (5.52)$_a$ (73)	7.27 (9.51)$_a$ (15)	21.17 (13.51)$_b$ (6)
PCS reactive aggression[b]	8.91*** (2, 91)	.16	7.95 (8.89)$_a$ (73)	12.00 (13.07)$_a$ (15)	25.67 (16.01)$_b$ (6)
BASC-3 Parent Rating Scale—Conduct Problems subscale	21.76*** (2, 90)	.33	52.94 (11.72)$_a$ (72)	69.73 (15.47)$_b$ (15)	80.17 (15.08)$_b$ (6)
BASC-3 Parent Rating Scales—Aggression subscale	9.08*** (2, 90)	.17	53.01 (12.39)$_a$ (72)	60.67 (13.31)$_a$ (15)	73.33 (7.34)$_b$ (6)

Note. Subscript letters denote significant differences between groups using Tukey's Honestly Significant Difference. Means with different letters are significantly different at $p < .05$. Group 0 = no clinically significant ODD or CD based on DISC or LPE based on CAPE; Group 1 = clinically significant ODD or CD but no significant LPE; Group 2 = clinically significant ODD or CD and LPE. ANOVA = analysis of variance; df = degrees of freedom; ODD = oppositional defiant disorder; CD = conduct disorder; LPE = limited prosocial emotions; DISC = Diagnostic Interview Schedule for Children; ICU = Inventory of Callous-Unemotional Traits; PCS = Peer Conflict Scale; BASC = Behavior Assessment System for Children.
[a] Denotes composite score based on whether the symptom was present based on either parent or youth report. [b] Denotes composite based on highest score from either parent or youth report.
*** $p < .001$.

symptom scores predicted parent-rated empathy and teacher-rated proactive aggression, even when controlling for parental ratings of CU traits on the ICU. An important finding in our tests of incremental validity is that the CAPE provided incremental validity over the ICU in predicting not only validity indicators outcomes assessed by structure interviews but in predicting validators based on parent and self-report ratings as well. Thus, its incremental validity does not appear to be solely due to shared method variance.

A final critical test of the CAPE was our test of the resulting LPE diagnosis as a specifier for clinical diagnoses of either ODD or CD. That is, the finding that CAPE scores are associated with measures of conduct problems and aggression could mean that the CAPE is measuring a risk factor for antisocial behavior, rather than designating a distinct subgroup of antisocial individuals who show more severe behavior problems (see Frick, 2022 for a discussion of this distinction). In the present study, within those clinic-referred youth who met criteria for either ODD or CD, 27% also showed two or more symptoms of the LPE specifier on the CAPE. This prevalence of the specifier is similar to what has been reported in other samples, using other methods for making the diagnosis (Kahn et al., 2012). When comparing the groups of youth with a conduct problem diagnosis based on the presence of the specifier, the group with the LPE specifier exhibited more CD symptoms and more aggression that the group with ODD/CD only (see Table 4). This evidence for the LPE specifier is much stronger than those reported by Hawes et al. (2020), who reported that children with ODD/CD and LPE did not differ from those with ODD/CD only on their level of conduct problems. While these discrepant results clearly indicate the need for more tests of LPE specifier as measured by the CAPE, one possible reason for these

differences in findings was that the CAPE was based solely on parental report in the Hawes and colleagues study and this may have limited its utility in forming diagnostic groups.

An interesting and unexpected finding is that the diagnostic groups formed by CAPE did not differ on the parent and child ratings on the ICU in our sample. That is, both the ODD/CD only and ODD/CD and LPE groups showed higher ratings on the ICU than the control group but the two ODD/CD groups did not differ significantly from each other. This finding needs to be interpreted cautiously because it was not predicted a priori and it was not found by Neo et al. (2023), who did report differences between these groups on the parent-rated ICU in a sample of young children. However, it is possible that ratings on the ICU are not as good as interviews at distinguishing between children's behavior, that may lead to harming others or that breaks rules, and their emotions in response to such behavior. That is, parents may assume a lack of guilt or empathy in their ratings when a child is showing a chronic pattern of antisocial behavior that creates a "halo effect" in their ratings of the emotions and behaviors.

In summary, our tests suggest that CAPE scores can be obtained reliably by trained clinicians, they show incremental utility in predicting conduct problem diagnoses over more time-efficient rating scales, and they designate a subgroup of youth with conduct problems diagnoses who seem to show a more severe level of antisocial behavior than children with conduct problems who do not show the specifier. These findings were obtained in a sample of clinic-referred children, a type of sample in which a clinical assessment of the LPE specifier is often required. It was also obtained by scoring the CAPE using multiple informants. However, the results also need to be interpreted in light of several limitations. Of most importance, the number of youth who met criteria for

ODD/CD in the clinic-referred sample was fairly small ($n = 22$) and, when this group was further divided into those with and with an LPE specifier, it led to a very small number of youth in the group showing the specifier ($n = 6$). Thus, clearly more tests of its use in larger samples are needed. Further, this limitation in sample size required us to group those with either an ODD or CD diagnosis, which is consistent with how the LPE specifier is used in the *ICD-11* but prevented us from testing it within children showing a diagnosis of CD, which is how it is used in the *DSM-5*. It also prevented us from testing whether the reliability and validity of the CAPE was consistent across the wide age range used in our study (i.e., 6–17 years of age). The CAPE was designed to allow clinicians to tailor their follow-up questions to the developmental level of the child, and the results from Neo et al. (2023) suggested that this can be done in a reliable and valid manner for children ages 2–8. However, more research is needed on how best collect information and weigh information from different informants across different developmental levels. Also, the test of the CAPE's validity in the present study focused on its association with conduct problems and aggression and its ability to designate a subgroup of those with a conduct problems diagnosis who showed more severe antisocial behavior. While this is an important test of validity of the CAPE and the LPE diagnoses, elevated CU traits have also been shown to designate subgroups of youth with conduct problems who show very different emotional (e.g., differences in emotional reactivity to distress in others) and cognitive (e.g., differences in their responses to punishment) characteristics, as well as differences in their response to treatment (see Frick, 2022 for a review). While these outcomes were tested by Neo et al. (2023) in young children, future studies need to continue to test the CAPE's validity in designating an important subgroup of youth with serious conduct problems, using other indicators of potential etiological and clinical utility.

Constraints on Generality

Our results provide further evidence on the psychometric properties of the CAPE. Previous studies have provided support for the reliability and validity of the CAPE using Spanish detained male adolescents (Molinuevo et al., 2020) and children and adolescents (ages 3–15; Hawes et al., 2020) and young children (ages 2–8; Neo et al., 2023) referred to outpatient mental health clinics in Australia. Thus, the present study was the first to test the psychometric properties of the CAPE in a sample from the United States that had substantial (28%) representations of youth who identified as Black. Given the positive findings across these studies with differing samples, we are optimistic that the results will generalize to other clinical settings. However, the available research has not directly tested the generalizability of results across gender, race, ethnicity, and socioeconomic status. This will be an important focus of future research. Further, all of the samples used in published research contain a substantial number of youth with significant conduct problems (e.g., detained adolescents, clinic-referred youth). Thus, it is not clear how well the results would generalize to samples with different base rates of serious conduct problems,

Conclusions

In summary, the current results, along with past findings (Hawes et al., 2020; Molinuevo et al., 2020; Neo et al., 2023) support cautious

use of the CAPE in clinical assessments and the continued development and testing of its psychometric properties. Of most importance, the current findings suggest that CAPE scores predicted clinically important constructs, even after controlling for the level of CU traits obtained on behavior ratings. However, a critical question is whether this incremental benefit is worth the added administration burden of having specifically trained assessors to administer the CAPE and the administration time needed to assess only four symptoms. As a result, it will be important to test other potentially less time-intensive methods for assessing the LPE specifier (see e.g., Walker et al., 2021) and compare their validity relative to diagnoses made by the CAPE. Also, the administrative burden suggests that, in most clinical settings, the CAPE could not be administered to all referrals, as was done in the clinic used in the present study. However, it is possible to use rating scales, like the ICU, as a screening method. The CAPE would then only be administered when a possible LPE diagnosis is likely based on scores from the rating scales. This would substantially limit the number of youth who require the CAPE and, as a result, the number of clinicians needed to administer it. However, such a multiple-gating diagnostic procedure has not been tested but doing so should be an important focus of clinical research to determine how to use the CAPE in the most cost-effective manner.

References

American Psychiatric Association. (2013). Disruptive, impulse-control, and conduct disorders. *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596.dsm15

Andershed, H., Kerr, M., Stattin, H., & Levander, S. (2002). Psychopathic traits in non-referred youths: A new assessment tool. In E. Blaauw & L. Sheridan (Eds.), *Psychopaths: Current international perspectives* (pp. 131–158). Elsevier.

Cardinale, E. M., & Marsh, A. A. (2020). The reliability and validity of the Inventory of Callous Unemotional Traits: A meta-analytic review. *Assessment*, 27(1), 57–71. https://doi.org/10.1177/1073191117747392

Creamer, J. S. E. A.; Burns, K., & Chen, F. (2022). *Poverty in the United States: 2021*. United States Census Bureau. https://www.census.gov/library/publications/2022/demo/p60-277.html

De Brito, S. A., Forth, A. E., Baskin-Sommers, A. R., Brazil, I. A., Kimonis, E. R., Pardini, D., Frick, P. J., Blair, R. J. R., & Viding, E. (2021). Psychopathy. *Nature Reviews Disease Primers*, 7(1), Article 49. https://doi.org/10.1038/s41572-021-00282-1

De Los Reyes, A., & Epkins, C. C. (2023). A dozen years of demonstrating that informant discrepancies are more than measurement error: Toward guidelines for integrating data from multi-informant assessments of youth mental health. *Journal of Clinical Child and Adolescent Psychology*, 52(1), 1–18. https://doi.org/10.1080/15374416.2022.2158843

Deng, J., Wang, M.-C., Zhang, X., Shou, Y., Gao, Y., & Luo, J. (2019). The Inventory of Callous Unemotional Traits: A reliability generalization meta-analysis. *Psychological Assessment*, 31(6), 765–780. https://doi.org/10.1037/pas0000698

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3), 613–619. https://doi.org/10.1177/001316447303300309

Forth, A. E., Bergstrom, H., & Clark, H. (2016). Psychopathic traits in adolescence: Assessment and implications. In C. Gacono (Ed.), *Clinical and forensic applications of psychopathy* (2nd ed., pp. 115–136). Lawrence Erlbaum.

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Hare Psychopathy Checklist: Youth version*. Multi-Health Systems.

Frick, P. J. (2004). *The Inventory of Callous-Unemotional Traits* [Measurement instrument]. https://faculty.lsu.edu/pfricklab/icu.php

Frick, P. J. (2013). *Clinical Assessment of Prosocial Emotions: Version 1.1 (CAPE 1.1)* [Measurement instrument]. https://sites01.lsu.edu/faculty/pfricklab/wp-content/uploads/sites/100/2015/11/CAPE-Manual.pdf

Frick, P. J. (2022). Some critical considerations in applying the construct of psychopathy to research and classification of childhood disruptive behavior disorders. *Clinical Psychology Review*, *96*, Article 102188. https://doi.org/10.1016/j.cpr.2022.102188

Frick, P. J., Barry, C. T., & Kamphaus, R. W. (2020). *Clinical assessment of children's personality and behavior* (4th ed.). Springer. https://doi.org/10.1007/978-3-030-35695-8

Frick, P. J., & Hare, R. D. (2002). *Antisocial Process Screening Device* [Database record]. APA PsycTests. https://doi.org/10.1037/t00032-000

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Can callous-unemotional traits enhance the understanding, diagnosis, and treatment of serious conduct problems in children and adolescents? A comprehensive review. *Psychological Bulletin*, *140*(1), 1–57. https://doi.org/10.1037/a0033076

Hawes, D. J., Kimonis, E. R., Mendoza Diaz, A., Frick, P. J., & Dadds, M. R. (2020). The Clinical Assessment of Prosocial Emotions (CAPE 1.1): A multi-informant validation study. *Psychological Assessment*, *32*(4), 348–357. https://doi.org/10.1037/pas0000792

IBM Corp. (2021). *IBM SPSS statistics for Macintosh* (Version 28.0) [Computer software]. https://www.ibm.com/products/spss-statistics

Kahn, R. E., Frick, P. J., Youngstrom, E., Findling, R. L., & Youngstrom, J. K. (2012). The effects of including a callous-unemotional specifier for the diagnosis of conduct disorder. *Journal of Child Psychology and Psychiatry*, *53*(3), 271–282. https://doi.org/10.1111/j.1469-7610.2011.02463.x

Kimonis, E. R., Frick, P. J., Skeem, J. L., Marsee, M. A., Cruise, K., Munoz, L. C., Aucoin, K. J., & Morris, A. S. (2008). Assessing callous-unemotional traits in adolescent offenders: Validation of the Inventory of Callous-Unemotional Traits. *International Journal of Law and Psychiatry*, *31*(3), 241–252. https://doi.org/10.1016/j.ijlp.2008.04.002

Lawing, K., Childs, K. K., Frick, P. J., & Vincent, G. (2017). Use of structured professional judgment by probation officers to assess risk for recidivism in adolescent offenders. *Psychological Assessment*, *29*(6), 652–663. https://doi.org/10.1037/pas0000414

López-Romero, L., Romero, E., & Luengo, M. Á. (2012). Disentangling the role of psychopathic traits and externalizing behaviour in predicting conduct problems from childhood to adolescence. *Journal of Youth and Adolescence*, *41*(11), 1397–1408. https://doi.org/10.1007/s10964-012-9800-9

Lynam, D. R. (1997). Pursuing the psychopath: Capturing the fledgling psychopath in a nomological net. *Journal of Abnormal Psychology*, *106*(3), 425–438. https://doi.org/10.1037/0021-843X.106.3.425

Makol, B. A., Youngstrom, E. A., Racz, S. J., Qasmieh, N., Glenn, L. E., & De Los Reyes, A. (2020). Integrating multiple informants' reports: How conceptual and measurement models may address long-standing problems in clinical decision-making. *Clinical Psychological Science*, *8*(6), 953–970. https://doi.org/10.1177/2167702620924439

Marsee, M. A., Barry, C. T., Childs, K. K., Frick, P. J., Kimonis, E. R., Muñoz, L. C., Aucoin, K. J., Fassnacht, G. M., Kunimatsu, M. M., & Lau, K. S. L. (2011). Assessing the forms and functions of aggression using self-report: Factor structure and invariance of the Peer Conflict Scale in youths. *Psychological Assessment*, *23*(3), 792–804. https://doi.org/10.1037/a0023369

Marsee, M. A., & Frick, P. J. (2007). Exploring the cognitive and emotional correlates to proactive and reactive aggression in a sample of detained girls. *Journal of Abnormal Child Psychology*, *35*(6), 969–981. https://doi.org/10.1007/s10802-007-9147-y

Matlasz, T. M., Frick, P. J., & Clark, J. E. (2022). Understanding the social relationships of youth with callous-unemotional traits using peer nominations. *Journal of Clinical Child and Adolescent Psychology*, *51*(4), 530–542. https://doi.org/10.1080/15374416.2020.1823847

Molinuevo, B., Martínez-Membrives, E., Pera-Guardiola, V., Requena, A., Torrent, N., Bonillo, A., Batalla, I., Torrubia, R., & Frick, P. J. (2020). Psychometric properties of the Clinical Assessment of Prosocial Emotions: Version 1.1 (CAPE 1.1) in young males who were incarcerated. *Criminal Justice and Behavior*, *47*(5), 547–563. https://doi.org/10.1177/0093854819892931

Muñoz, L. C., Kerr, M., & Bešić, N. (2008). The peer relationships of youths with psychopathic personality traits: A matter of perspective. *Criminal Justice and Behavior*, *35*(2), 212–227. https://doi.org/10.1177/0093854807310159

Neo, B., Fleming, G. E., Kaouar, S., Chan, M. E., Huang, N. N., Hawes, D. J., Eapen, V., Briggs, N., & Kimonis, E. R. (2023). Clinical utility of diagnosing limited prosocial emotions in young children using the Clinical Assessment of Prosocial Emotions (CAPE). *Psychological Assessment*, *35*(12), 1085–1097. https://doi.org/10.1037/pas0001279

Perlstein, S., Fair, M., Hong, E., & Waller, R. (2023). Treatment of childhood disruptive behavior disorders and callous-unemotional traits: A systematic review and two multilevel meta-analyses. *Journal of Child Psychology and Psychiatry*, *64*(9), 1372–1387. https://doi.org/10.1111/jcpp.13774

Piacentini, J. C., Cohen, P., & Cohen, J. (1992). Combining discrepant diagnostic information from multiple sources: Are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology*, *20*(1), 51–63. https://doi.org/10.1007/BF00927116

Ray, J. V., & Frick, P. J. (2020). Assessing callous-unemotional traits using the total score from the Inventory of Callous-Unemotional Traits: A meta-analysis. *Journal of Clinical Child and Adolescent Psychology*, *49*(2), 190–199. https://doi.org/10.1080/15374416.2018.1504297

Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior assessment system for children* (3rd ed.). NCS Pearson.

Ritchie, M. B., Neufeld, R. W. J., Yoon, M., Li, A., & Mitchell, D. G. V. (2022). Predicting youth aggression with empathy and callous unemotional traits: A meta-analytic review. *Clinical Psychology Review*, *98*, Article 102186. https://doi.org/10.1016/j.cpr.2022.102186

Rolon-Arroyo, B., Arnold, D. H., & Harvey, E. A. (2014). The predictive utility of conduct disorder symptoms in preschool children: A 3-year follow-up study. *Child Psychiatry and Human Development*, *45*(3), 329–337. https://doi.org/10.1007/s10578-013-0404-8

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(1), 28–38. https://doi.org/10.1097/00004583-200001000-00014

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Vagos, P., Marinho, P. I., Pandeirada, J. N. S., Rodrigues, P. F. S., & Marsee, M. (2021). Measuring forms and functions of aggression in Portuguese young adults: Validation of the Peer Conflict Scale. *Journal of Psychoeducational Assessment*, *39*(7), 902–908. https://doi.org/10.1177/07342829211018106

Walker, T. M., Frick, P. J., Matlasz, T. M., Robertson, E. L., Mikolajewski, A. J., Mitchell, C., Lopez-Duran, N., Monk, C., & Hyde, L. W. (2021). Psychometric properties of a semistructured interview to assess limited prosocial emotions. *Assessment*, *28*(7), 1755–1764. https://doi.org/10.1177/1073191120947796

World Health Organization. (2019). *International statistical classification of diseases and related health problems* (11th ed.). https://icd.who.int/