Routledge
Taylor & Francis Group

Journal of Clinical Child & Adolescent Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hcap20

Developing Cutoff Scores for the Inventory of Callous-Unemotional Traits (ICU) in Justice-Involved and Community Samples

Emily C. Kemp, Paul J. Frick, Tatiana M. Matlasz, Julia E. Clark, Emily L. Robertson, James V. Ray, Laura C. Thornton, Tina D. Wall Myers, Laurence Steinberg & Elizabeth Cauffman

To cite this article: Emily C. Kemp, Paul J. Frick, Tatiana M. Matlasz, Julia E. Clark, Emily L. Robertson, James V. Ray, Laura C. Thornton, Tina D. Wall Myers, Laurence Steinberg & Elizabeth Cauffman (2021): Developing Cutoff Scores for the Inventory of Callous-Unemotional Traits (ICU) in Justice-Involved and Community Samples, Journal of Clinical Child & Adolescent Psychology, DOI: 10.1080/15374416.2021.1955371

To link to this article: https://doi.org/10.1080/15374416.2021.1955371

JOURNAL OF CLINICAL CHILD & ADOLESCENT PSYCHOLOGY
https://doi.org/10.1080/15374416.2021.1955371

Developing Cutoff Scores for the Inventory of Callous-Unemotional Traits (ICU) in Justice-Involved and Community Samples

Emily C. Kemp[a], Paul J. Frick[a,b], Tatiana M. Matlasz[a], Julia E. Clark[c], Emily L. Robertson[a], James V. Ray[d], Laura C. Thornton[e], Tina D. Wall Myers[f], Laurence Steinberg[g], and Elizabeth Cauffman[h]

[a]Department of Psychology, Louisiana State University; [b]Institute for Learning Sciences and Teacher Education, Australian Catholic University; [c]Morrissey-Compton Educational Center; [d]Department of Criminal Justice, University of Central Florida; [e]Bureau of Family Health, Louisiana Office of Public Health; [f]Independent Scholar; [g]Department of Psychology, Temple University; [h]Department of Psychological Science, University of California

ABSTRACT

Objective: The recent addition of the callous-unemotional (CU) traits specifier, "with Limited Prosocial Emotions (LPE)," to major classification systems has prompted the need for assessment tools that aid in the identification of elevations on these traits for diagnostic purposes. The goal of the current study was to use and evaluate multiple methods for establishing cutoff scores for the multi-informant questionnaire, the Inventory of Callous-Unemotional Traits (ICU).

Method: The present study compared the clinical utility of various proposed cutoff methods and scores (i.e., empirically derived cutoffs using receiver operating characteristic (ROC), normative cutoffs, and rational scoring approximations of LPE criteria) in both a longitudinal sample of justice-involved male adolescents (N = 1,216; M_{age} = 15.29, SD = 1.29) and a cross-sectional sample of school children (N = 289; M_{age} = 11.47 years; SD = 2.26).

Results: Methods resulted in a range of cutoff scores with substantial diagnostic overlap and validity. Specifically, they designated justice-involved adolescents at risk for later delinquency, aggression, and rearrests, and they designated school children more likely to be rated by parents and teacher as having conduct problems and rated by peers as being rejected and mean.

Conclusions: The results lead to ranges of ICU scores that have support for their validity and can help to guide clinical decisions about children and adolescents who may be elevated on CU traits.

Establishing Cutoff Scores for the Inventory of Callous-Unemotional Traits (ICU)

Latest versions of major diagnostic systems, including the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (DSM-5; American Psychiatric Association, 2013) and the *International Classification of Disease, 11th Revision* (ICD-11; World Health Organization, 2018), have added the specifier, "with Limited Prosocial Emotions (LPE)," to refer to a subgroup of youth with severe behavior problems who display significant levels of callous-unemotional (CU) traits. CU traits are defined by a callous-lack of empathy, deficient guilt, lack of concern about performance in important activities, and constricted or superficial affect (Frick et al., 2014a). CU traits are theorized to represent an affective dimension of the broader construct of psychopathy in adults (Hare & Neumann, 2008) and the affective components of conscience in children (Frick et al., 2014b). Further, their inclusion in diagnostic classification was based on research evidencing associations between CU traits and particularly severe (i.e., violent) and stable forms of antisocial behavior (Frick et al., 2014a). For example, youth with elevated CU traits are shown to engage in greater weapon use (i.e., gun carrying and gun use; Robertson et al., 2020; Saukkonen et al., 2016) and more harmful forms of aggression (Crapanzano et al., 2011; Fanti & Kimonis, 2012; Golmaryami et al., 2016; Lawing et al., 2010). In addition, youth with elevated CU traits, even in the absence of serious conduct problems, are found to experience greater psychosocial and interpersonal impairment (Ciucci et al., 2014; Graziano et al., 2016; Waller et al., 2016). For example, youth with elevated CU traits are shown to experience greater peer rejection and are more likely to be described as "mean" by peers (Graziano et al., 2016; Matlasz et al., 2020). Further, CU traits seem to designate an etiologically distinct group of youth with severe behavior problems who

CONTACT Emily C. Kemp ekemp4@lsu.edu Department of Psychology, Louisiana State University, 236 Audubon, Baton Rouge, LA 70803
Supplemental material for this article can be accessed online at https://doi.org/10.1080/15374416.2021.1955371

display reduced emotional reactivity to others' distress and lower sensitivity to punishment (Blair et al., 2014; Frick et al., 2014a).

With the inclusion of the LPE specifier in the latest versions of major classification systems, there is an enhanced need to develop and refine assessments for CU traits. The Inventory of Callous-Unemotional Traits (ICU; Kimonis et al., 2008) was developed to provide a comprehensive assessment of CU traits as they are operationalized in the DSM-5 LPE specifier for Conduct Disorder (CD). It is a relatively short (24-item) multi-informant (i.e., self-, parent-, and teacher-report) questionnaire. The ICU has six items assessing each of the four symptoms of the LPE specifier, with three positively worded (i.e., indicating higher CU traits) and three negatively worded (i.e., indicating lower CU traits) items assessing each symptom. Items are rated on a 4-point Likert-type scale, from 0 (*Not at all true*) to 3 (*Definitely true*), that does not allow for a median or central tendency rating.

The ICU has been translated into over 28 languages and used in over 250 peer-reviewed studies, which generally provide strong support for its ability to measure CU traits continuously. That is, across various factor analyses, there is support for an overarching CU dimension that is captured well by a summed unit-weighting of items (Ray & Frick, 2018).[1] In addition, a meta-analysis of 115 independent samples (N = 27,947) by Cardinale and Marsh (2020) provide support for the validity of the ICU self-report total score, including good internal consistency (α_{mean} = .83) and significant associations in expected directions with relevant constructs (e.g., positive correlations with externalizing problems, delinquency, and proactive aggression; negative correlations with measures of empathy). Though there have been fewer studies validating ICU informant-report versions, several studies have reported similar estimates of internal consistency for ICU parent-report (α_{mean} = .82; Ueno et al., 2019; Yoshida et al., 2019) and teacher-report (α_{mean} = .88; see Ueno et al., 2019) total scores. In addition, support has been found for the construct validity of both informant-report total scores, with these versions showing significant associations with aggression and rule-breaking behavior (Ueno et al., 2019).

Although there is strong support for the ICU as a continuous measure of CU traits, a major limitation in its use for diagnostic purposes is the absence of well-validated cutoff scores. One way that the ICU has been used to identify youth with elevated CU traits has been to use deviations from the mean or median score in a sample. For example, both Lawing et al. (2010) and Viding et al. (2012) used a median split to distinguish groups of 150 detained adolescents (ages 12 to 20) and 46 community youth (ages 10 to 16), respectively, as either high (above the median) or low (below the median) on CU traits. These studies reported that youth with elevated CU traits, based on this method, showed higher levels of risk associated with sex offending (Lawing et al., 2010) and reduced amygdala activation in response to fearful faces (Viding et al., 2012). However, using such normative cutoffs in small, non-representative samples could make the findings difficult to replicate. Thus, the use of normative cutoffs would be aided by scores from large, representative samples. Recently, one source of normative data for the ICU self-report total score has become available from Kemp et al. (2019) based on a sample of 4,683 boys and girls ages 11–17 (separate data provided for girls ages 11–14 (n = 1,475), boys 11–14 (n = 1,444), girls 15–17 (n = 833), and boys 15–17 (n = 931) from four different countries. Unfortunately, the clinical utility of these normative cutoffs has yet to be tested and such a large normative base has yet to be established for either of the ICU informant versions.

Another method for developing cutoff scores is to select individual ICU items that most closely match each of the four DSM-5 LPE symptom criteria and to select a minimum rating on each item that indicates symptom presence. An individual is considered "elevated" by this approach if at least two of the four items reach this symptom threshold, as specified by the LPE criteria. A recent review of nine studies using this approach indicated that this method of grouping children and adolescents resulted in a wide variation in prevalence rates, ranging from as low as 6% when using ICU self-report in clinic-referred youth to as high as 77% when using parent-report in detained youth, and this method of grouping youths led to inconsistent findings in support of its validity (Colins et al., 2020). The inconsistent findings using this method are likely due to several limitations. First, this method uses a very limited item pool (e.g., four items) and a very restricted range of scores (i.e., 0 to 4) on which to determine elevations (see, for example, Kimonis et al., 2015). More stable estimates are likely to be obtained if

[1]Factor analyses have often identified subdomains within this overarching factor. However, the validity of these subdomains have been questioned, since they seem to be, at least in part, a result of item wording (Ray et al., 2016). That is, a callousness factor often emerges consisting largely of items worded in the positive, callous direction, and an uncaring factor often emerges consisting largely of items worded in the negative, prosocial direction. However, in the few factor analyses that consider item wording as a method factor, the results support the structure that led to the development of ICU, with 4 item clusters consisting of 6 items each (3 positively worded and 3 negatively worded) contributing to an overarching CU factor (Kliem et al., 2020; Koutsogiorgi et al., 2020).

multiple items assessing each symptom are used to estimate symptom presence, consistent with how the ICU was developed to have six items assessing each of the four symptoms. Second, past uses of such rational approaches did not consider findings from IRT analyses; that is, higher scores are less likely to be obtained on items worded in the positive (i.e., callous) direction than the negative (i.e., caring) direction, after the latter items are reverse scored (Ray et al., 2016). Thus, lower ratings on positively worded items are potentially more meaningful than higher ratings on negatively worded items, and such differences in endorsement rates should be considered when estimating symptom presence.

A final approach for forming cutoff scores includes the use of receiver operating characteristic (ROC) to empirically derive cutoff scores that optimize sensitivity and specificity for predicting clinically important outcomes (Youngstrom, 2013). Specifically, ROC analysis produces a graphical curve that plots the rate of true positives (i.e., sensitivity rate captured by those individuals who are positive for a dichotomous outcome) against the rate of false positives (i.e., the inverse of the specificity rate captured by those who are incorrectly categorized as positive for an outcome) resulting from all possible ICU total score cutoffs. The area under the ROC curve (i.e., AUC) that results from this graph provides an estimate of the test's overall classification accuracy (Fawcett, 2006). Further, with the use of ROC, Youden's index (YI) can be used to determine at what score both sensitivity and specificity are maximized and classification accuracy may be considered optimal. To illustrate the use of this approach, Docherty et al. (2017) used ROC analyses to develop optimal points at which ICU scores predicted indices of aggression, violence, and juvenile justice detention in a combined sample (N = 634) of community and incarcerated youth. They reported that ICU total scores of 28 for youth self-report, 30 for parent-report, and 33 for teacher-report were optimal for the prediction of these clinically relevant outcomes (Docherty et al., 2017). However, this approach, which was based on predicted probabilities, can be highly influenced by the base rates of the outcomes of interest (Youngstrom, 2013). Thus, such analyses need to be replicated in different samples and by using a diversity of outcomes that vary in their base rates to determine empirically derived cutoffs that can generalize across outcomes and samples.

Thus, several methods for determining useful cutoff scores for the ICU have been published, each with strengths and weaknesses in their approach. However, to date, no study has attempted to compare the results of these different methods in a single sample or using all three versions of the ICU. With this, the current study tested cutoffs for ICU self-, parent-, and teacher-report versions formed using a normative cutoff method, an expanded rational scoring method for determining the presence of the LPE specifier, and ROC analyses to determine empirically derived cutoffs. These different methods were compared to see if they identified similar persons as elevated on the ICU and to test how well they identified individuals with clinically important characteristics in two different samples. In a sample of adolescent boys who were arrested for the first time for an offense of moderate severity, we tested how well the ICU self-report cutoffs identified those elevated on self-reported aggression, self-reported delinquency, self-reported gun use, and those who were rearrested at least twice over the five years following their first arrest. Then, in a community sample of school children, we tested how well the cutoffs obtained for all three ICU versions identified youth who were rated as having significant conduct problems by parents and teacher and as being rejected and perceived as "mean" by peers.

Methods

Participants and Procedures

The first sample included 1,216 male adolescents who had been arrested for the first time. These participants were recruited from three regions in the United States as part of a multi-site prospective study and were considered eligible if they were male, English-speaking, between the ages of 13 and 17 years (M = 15.29, SD = 1.29), and recently arrested for an eligible offense of low-to-moderate severity (see Ray et al., 2016 for more details). Institutional review boards at all institutions approved the study procedures. Participants completed baseline interviews within six weeks of the disposition date for their first arrest and were then reassessed every six months during the first three years and once per year during years four and five, providing eight total waves of follow-up data. Participants were compensated $50 at baseline, and payments increased up to a total of $140 by year four. At baseline, self-reported race and ethnicity data showed that the sample was primarily Hispanic (46%), followed by African American (37%), White (15%), and self-identified other (3%). At baseline, the majority (82%) of participants' parents had an education level less than a college diploma.

The second sample included 289 boys and girls, ages 8 to 15 years (M = 11.47, SD = 2.26), who were recruited as part of a cross-sectional study from two school systems in the southeastern United States. Study approvals were received from the university institutional review board, school superintendent, and principals. Schools

were compensated $10 per participating child. The sample included students in the 3rd (n = 93), 6th (n = 69), and 8th (n = 127) grades and was comprised 60% of girls. According to parent-report, the sample was primarily Black, Afro-Caribbean, or African American (40%), followed by White (35%), Biracial (12%), Hispanic (5%), and other ethnic minorities (2% East Asian or Asian American, 1% Middle Eastern or Arab American, 1% Native American or Alaskan Native, and 1% self-identified other), while 4% of the sample chose not to report on race or ethnicity. The majority (86%) of participants' parents had an education level less than a college diploma.

Measures (Both Samples)

Callous-Unemotional (CU) Traits

CU traits were assessed with the ICU total scale in both samples. In the first sample, the ICU self-report version (M = 26.28, SD = 8.08) was collected during baseline interviews within six weeks of the adolescent's first arrest and showed acceptable internal consistency (α = .76). In the second sample, ICU self-, parent-, and teacher-report versions were administered. In this sample, ICU self-report (M = 18.93, SD = 8.04) showed acceptable internal consistency (α = .76), parent-report (M = 17.26, SD = 9.92) showed good internal consistency (α = .86), and teacher-report (M = 21.31, SD = 13.54) showed excellent internal consistency (α = .93).

Measures (Justice-Involved Only)

Self-Reported Aggression

In the sample of justice-involved adolescents, levels of self-reported aggression were measured at all follow-up points with the Peer Conflict Scale (PCS; Marsee et al., 2011, 2014). The 20 items assessing physical aggression, both reactive and proactive (10 items each), were used. Items were rated on a 4-point Likert-type scale from 0 (*Not at all true*) to 3 (*Definitely true*) and summed to form continuous total scores. Across eight follow-up points, the physical aggression total score showed good to excellent internal consistency (α = .85-.90), the physical proactive subscale showed acceptable to good internal consistency (α = .72-.83), and the physical reactive subscale showed good internal consistency (α = .83-.86). For ROC analyses, a cutoff z-score of 1.5 was applied to form dichotomized variables of high versus low total, proactive, and reactive aggression.

Self-Reported Delinquency

Levels of self-reported delinquency were assessed at all follow-up points with the 24-item revised version of the Self-Report of Offending Scale (SRO; Huizinga et al., 1991). For each item, participants were asked (yes or no) if, in the last six months, they engaged in each crime. Total offending and violent offending (i.e., crimes against persons) variety scores were calculated, with higher scores representing a greater variety of crimes committed. Across the eight follow-up points, the total variety score showed acceptable to good internal consistency (α = .75-.83), while the violent offending variety score showed poor to questionable internal consistency (α = .51-.64); however, due to low base rates of endorsements on this subscale, mean inter-item correlations are also reported (r = .16-.29). For ROC analyses, a cutoff z-score of 1.5 was applied to form dichotomized variables of high versus low total and violent self-reported offending.

Self-Reported Gun Use

Self-reported gun carrying was assessed at all follow-up points using gun-related items of the SRO. Specifically, each item asked participants (yes or no) if they carried a gun at any time since the last interview and, if yes, how many times. A total gun carrying variable was created by summing the number of times participants endorsed carrying a gun across all follow-up points. This item was chosen given that gun carrying prior to adulthood is illegal, and carrying a gun predicts later offending and violence (Emmert et al., 2018). For ROC analyses, and due to a relatively low base rate (Table 1), a dichotomized variable of any gun carrying over follow-up was used. Self-reported gun use during a crime was also assessed using items from the SRO at each follow-up. Specifically, participants were asked (yes or no) if, since the last interview, they had committed a violent crime (e.g., carjacked someone; see Robertson et al., 2020 for more details). If participants endorsed engagement in any of these offenses, they were then asked (yes or no) if they used a gun during the commission of that crime. Due to a low base rate (Table 1), a dichotomous variable for any endorsement of gun use during a crime was used in all analyses.

Official Arrests

Official rearrest data were obtained for both juvenile and adult arrests within the jurisdictions in which the participants were initially arrested. Only new charges during follow-up were included (i.e., probation and technical violations were excluded). Rearrest data for all crimes, as well as violent crime (i.e., crimes against persons), were assessed. Due to a low base rate of rearrest for violent crime (Table 1), a dichotomous variable (i.e., at least one rearrest for violent crime at any point during follow-up) was used in all analyses. For ROC analyses, and due to

Table 1. Main study variables: distributions and zero-order correlations.

Justice-Involved Sample					
Variables	Mean (SD) or % (n)	Range	Skewness	Kurtosis	Corr. with ICU-SR
ICU-SR	26.28 (8.08)	0– 55	.07	.08	–
Total aggression	4.57 (4.94)	0– 40	2.30	7.66	.35**
Proactive aggression	.98 (1.70)	0– 17	3.78	19.51	.31**
Reactive aggression	3.59 (3.54)	0– 23	1.68	3.68	.34**
Total SR offending	8.02 (11.44)	0– 85	2.85	10.58	.31**
Violent SR offending	3.06 (4.30)	0– 33	2.69	9.37	.28**
Any gun carrying	19% (n = 221)	–	–	–	–
Any criminal gun use	9% (n = 102)	–	–	–	–
Any with 2+ rearrests	28% (n = 330)	–	–	–	–
Any violent rearrests	18% (n = 217)	–	–	–	–

School Sample							
Variables	Mean (SD)	Range	Skewness	Kurtosis	Corr. with ICU-SR	Corr. with ICU-PR	Corr. with ICU-TR
ICU-SR	18.93 (8.04)	3– 50	.71	.71	–	–	–
ICU-PR	17.26 (9.92)	1– 47	.76	.12	.24**	–	–
ICU-TR	21.31 (13.54)	0– 61	.51	−.50	.29**	.24**	–
Conduct problems	25.65 (8.71)	18– 67	1.77	3.39	.29**	.49**	.52**
Peer rejection	.01 (1.00)	−.7– 5.1	2.44	7.19	.30**	.15*	.20**
Meanness	.002 (.72)	−1 – 5	2.95	12.60	.33**	.17**	.21**

ICU = Inventory of Callous-Unemotional Traits; SR = self-report; PR = parent-report; TR = teacher-report; SD = standard deviation. Correlations ("Corr.") reported with Pearson's *r*. Gun use and rearrest reported as frequency scores (e.g., number of times endorsed across follow-up).
**$p < .01$, *$p < .05$.

a relatively high base rate of any rearrests over follow-up (45%), a dichotomous variable was created such that any participant with two or more rearrests (Table 1) were coded as 1 and those with one or fewer were coded as 0.

Measures (School Sample Only)

Informant-Reported Conduct Problems

In the sample of school children, informant-reported conduct problems were measured with the Disruptive Behavior Disorders Scale (DBD; Pelham et al., 1992). Only items assessing Oppositional Defiant Disorder (ODD) and CD symptoms were included in analyses. Items were rated on a 4-point Likert-type scale from 1 (*Not at all*) to 4 (*Very much*). The DBD was completed by both parent and teacher, and both showed excellent internal consistency (parent-report α = .95; teacher-report α = .96) and were correlated with one another (r = .29, $p < .001$). Based on the recommendation from Piacentini et al. (1992), the highest rating was taken to yield a resolved score for each item. These resolved scores were then summed and averaged to create a composite score of conduct problems that also showed excellent internal consistency (α = .93). For ROC analyses, criteria set forth by Pelham et al. (1992) to approximate diagnostic criteria were used to create a dichotomous variable of high versus low conduct problems. Based on these criteria, 50 youth (17%) in the sample met criteria for either elevated ODD or CD symptoms, or both.

Peer Rejection

Peer rejection was measured with a standard peer-nomination question that asked students "who do you like the least?" out of all classmates in the same grade (McMullen et al., 2014).[2] Participants were allowed to nominate same- and other-gender peers within their grade at their school. Nominations for this sociometric item were converted to proportion scores that were then standardized within grade. For ROC analyses, a cutoff z-score of 1.5 was applied to form a dichotomous variable of peer-rejection.

Peer-Nominated "Meanness"

Levels of peer-reported "meanness" were assessed with peer nominations that asked participants to name: "who is mean?"; "who doesn't care who they hurt?"; "who always has to get his or her own way?"; "who doesn't care about having friends?"; and "who is hard to get to know well?" (see Matlasz et al., 2020). Participating youth were given unlimited nominations and were instructed to write in the names of classmates in their same grade who they felt best fit each description. These scores were converted to proportion scores and standardized within grade. For ROC analyses, a cutoff z-score of

[2]We recognize that peer rejection is often measured as a difference score between nominations of being "liked least" by classmates subtracted from nominations of being "liked most" by classmates (McMullen et al., 2014). However, due to concerns about the psychometric properties of difference scores (De Los Reyes, 2017), only the "liked least" nomination was used to validate the ICU. When analyses were run using than more traditional method of assessing peer rejection, the results were relatively unchanged.

1.5 was used to form a dichotomous variable of high versus low peer-nominated meanness.

Analytic Plan

All analyses were conducted using IBM SPSS Statistics. There was minimal missing data across both samples. ROC analyses were performed using dichotomized versions of each outcome variable (described above) in each sample. Nonparametric distributions were assumed. The classification accuracies (i.e., AUCs) resulting from each test were reported. Further, resulting measures of sensitivity and specificity were used to calculate YI (sensitivity + (specificity – 1)) for all possible ICU cutoffs. For each sample and respective dichotomized outcome, the ICU total score with the largest YI was selected as the optimal cutoff. Based on these analyses, the average of these scores, across all outcomes, was selected as the empirically derived cutoff.

In order to assess the clinical utility of normative cutoffs, separate ICU self-, parent-, and teacher-report norms were applied to each sample. Participants in the first sample were classified as having clinically elevated CU traits based on the following sex- and age-based multinational norms at the 90th percentile: ICU self-report total score ≥34 for boys ages 11–14 and ≥37 for boys ages 15–17 (Kemp et al., 2019). In the second sample, the same age- and sex-based multinational norms provided by Kemp et al. (2019) were applied, that is ICU self-report total score ≥34 for boys ages 14 and younger; ≥37 for boys age 15; ≥29 for girls ages 14 and younger; and ≥32 for girls age 15. In addition, sex-based local norms at the 90th percentile were used: ICU parent-report total score ≥34 for boys and ≥30 for girls and ICU teacher-report total score ≥46 for boys and ≥35 for girls.

As noted previously, the rational scoring method used in much of the past research was limited in the number of items used to approximate the LPE specifier (Colins et al., 2020). To overcome this limitation, six items were used to assess each symptom and different methods for determining symptom presence were used for positively and negatively worded items. Specifically, for positively worded items, ratings of "very true" or "definitely true" were scored as indicating symptom presence. For negatively worded items, only "extreme" responses of "not at all true" were scored as indicating symptom presence. A symptom was considered present if two or more of the six items representing that symptom reached the threshold, and the LPE criteria was considered present if two or more of the symptoms were considered present.

Once cutoff scores were established, intercorrelations among cutoff methods (i.e., associations between binary variables) were tested with the phi coefficient. In addition, degrees of classification overlap (i.e., percentage agreement) across scoring methods were reported in each sample. To compare the clinical utility resulting from each cutoff method in each sample, group-means difference tests in the forms of chi-square and independent-samples *T*-tests were performed across outcomes. Using meta-analytic methods, we converted the inferential statistics into a common effect size (i.e., Pearson's *r*) and examined the average effect sizes across outcomes for each cutoff method in each sample.

Results

The distributions of main study variables are described in Table 1. In addition, all zero-order correlations between ICU total scores (self-, parent-, and teacher-report) and continuous outcomes in each sample are reported in Table 1.

Receiver Operating Characteristic (ROC) Analysis

Justice-Involved Sample

First, ICU self-report scores at baseline performed significantly better than chance at distinguishing participants with high versus low levels of total, proactive, and reactive aggression over follow-up (Supplementary Figure 1). Further, YIs of .41, .38, and .36 indicated optimal cutoffs of 30, 34, and 28 on ICU self-report for distinguishing participants with high versus low total, proactive, and reactive aggression, respectively. Second, ICU self-report performed significantly better than chance at distinguishing participants with high versus low total and violent self-reported offending over follow-up (Supplementary Figure 1). YIs of .37 and .19 indicated optimal cutoffs of 28 and 24 for distinguishing high versus low total and violent offending, respectively. Third, ICU self-report performed significantly better than chance at distinguishing participants who reported any gun carrying over the five-year follow-up and those who reported any gun use during crime over follow-up (Supplementary Figure 1). Further, YIs of .20 and .29 indicated optimal cutoffs of 30 and 28 for distinguishing participants who reported any gun carrying and any gun use during crime five years after first arrest. Lastly, ICU self-report performed significantly better than chance at distinguishing participants with two or more rearrests and those with any violent rearrests over follow-up (Supplementary Figure 1). Further, YIs of .13 and .08 indicated optimal cutoffs of 27 and 33 for distinguishing participants with more rearrests and those with any violent rearrests, respectively. Thus, across these

different outcomes, an average cutoff of 29 was optimal for ICU self-report.

School Sample

ICU self-report scores performed significantly better than chance at distinguishing participants with high versus low informant-reported conduct problems and high versus low peer-nominated "meanness" (Supplementary Figure 2). Further, YIs of .25 and .31 indicated optimal cutoffs of 22 and 26, respectively, for distinguishing participants with significant conduct problems and high levels of peer-nominated meanness. ICU self-report scores did not perform significantly better than chance at distinguishing high versus low peer rejection (Supplementary Figure 2), and thus, YI was not calculated or factored into the computation of an average optimal cutoff. Thus, across these outcomes, an average cutoff of 24 was considered optimal for ICU self-report.

ICU parent-report scores performed significantly better than chance at distinguishing participants with high versus low conduct problems, peer rejection, and peer-nominated "meanness" (Supplementary Figure 3). Further, YIs of .41, .33, and .28 indicated optimal cutoffs of 23, 18, and 23 for distinguishing participants with significant conduct problems, high levels of peer rejection, and high levels of peer-nominated meanness, respectively. Thus, across these outcomes, an average cutoff of 21 was considered optimal for ICU parent-report.

ICU teacher-report scores performed significantly better than chance at distinguishing participants with high versus low conduct problems and peer-nominated "meanness" (Supplementary Figure 4). Further, YIs of .57 and .47 indicated optimal cutoffs of 32 and 37 for distinguishing participants with significant conduct problems and high levels of peer-nominated meanness, respectively. ICU teacher-report scores did not perform significantly better than chance at distinguishing high versus low peer rejection (Supplementary Figure 4), and thus, YI was not calculated or factored into the computation of an average optimal cutoff. Thus, across these outcomes, an average cutoff of 35 was considered optimal for ICU teacher-report.

Classification by Cutoff Methods

Justice-Involved Sample

The empirically derived cutoff of 29 classified 39% (n = 479) of participants as having elevated CU traits. The second cutoff method, sex- and age-based normative cutoffs at the 90th percentile, classified 13% (n = 152) of participants as having elevated CU traits. The final cutoff method, LPE criteria approximation using rational scoring, classified 27% (n = 331) of participants as having elevated CU traits. Across classification methods, empirical and normative cutoffs were most highly correlated (φ = .47), followed by correlations between both empirical and normative cutoffs and the rational LPE approximation (φs = .37). In terms of classification agreement (i.e., the proportion of participants classified by both methods as having either high or low CU traits), the agreement was fairly similar between normative cutoffs and the LPE approximation (78%) and the empirical cutoff and LPE approximation (71%).

School Sample

The empirically derived cutoffs of 24, 21, and 35 for ICU self-, parent-, and teacher-report, respectively, classified 25% (n = 64), 31% (n = 90), and 19% (n = 52), respectively, as having elevated CU traits. The second cutoff method used multinational sex- and age-based self-report norms (90^{th} percentile) and classified 9% (n = 23) of participants as having elevated CU traits, in addition to local sex-based parent- and teacher-report norms (90^{th} percentile) that classified 11% (n = 32) and 13% (n = 35), respectively, as having elevated CU traits. The third cutoff method, LPE approximation with rational scoring, classified 17% (n = 43) as having elevated CU traits based on ICU self-report, 10% (n = 27) based on parent-report, and 16% (n = 42) based on teacher-report.

For ICU self-report, empirical and normative cutoffs were most highly correlated (φ = .55), followed by correlations between normative cutoffs and LPE approximation (φ = .45) and the empirical cutoff and LPE approximation (φ = .44). In terms of classification agreement, normative cutoffs and LPE approximation resulted in the greatest agreement (87%), followed by similarly high agreement (81%) between the empirical cutoff and LPE approximation. For ICU parent-report, normative cutoffs and the LPE approximation were most highly correlated (φ = .61), followed by correlations between empirical and normative cutoffs (φ = .52) and the empirical cutoff and LPE approximation (φ = .42). In terms of classification agreement, normative cutoffs and LPE approximation resulted in the greatest agreement (93%), followed by agreement (76%) between the empirical cutoff and LPE approximation. For ICU teacher-report, correlations between the empirical cutoff and normative cutoffs were most highly correlated (φ = .79), followed by correlations between empirical cutoff and LPE approximation (φ = .75) and normative cutoffs and LPE approximation (φ = .60). In terms of classification agreement, the empirical cutoff and LPE

approximation resulted in the greatest agreement (93%), followed by similarly high agreement (90%) between normative cutoffs and LPE approximation

Validation of Cutoff Scores

Justice-Involved Sample

Results from chi-square and independent-samples *T*-tests are reported in Table 2, showing the differences between those low and high on CU traits based on the different cutoff methods on the various outcome measures in the sample of justice-involved adolescents. The average effect sizes produced by cutoff methods, along with exact total score cutoffs resulting from empirical and normative cutoff methods, are reported in Table 3. The three methods led to similar average effect sizes, with the empirical cutoff resulting in an average effect size of r = .19, the normative cutoffs resulting in an average effect size of r = .17, and the LPE approximation resulting in an average effect size of r = .15.

School Sample

Results from independent-samples *T*-tests are reported in Table 4, showing the differences between those low and high on CU traits based on the different cutoff methods on the various outcome measures in the school sample. The average effect sizes produced by cutoff methods, along with exact total score cutoffs resulting from empirical and normative cutoff methods, are reported in Table 5. For ICU self-report, the empirical cutoff method produced the largest average effect size (r = .18), followed by normative cutoffs (r = .15) and the LPE approximation (r = .10). For ICU parent-report, both empirical and normative cutoffs resulted in equivalently large effect sizes (r = .18), followed by the LPE approximation (r = .10). Finally, for ICU teacher-report, normative cutoffs provided the largest average effect size (r = .27), followed by the empirical cutoff (r = .26) and LPE approximation (r = .24).

Table 3. Comparisons of effect size across ICU cutoff methods in justice-involved sample.

ICU Self-Report Cutoff Method	ES (r)	SE	95% CI
Empirical Cutoff: ICU total = 29	.19	.03	.14 – .23
Normative Cutoffs: Male age ≤14 = 34; Male age ≥ 15 = 37	.17	.04	.10 – .24
Rational LPE Approximation	.15	.02	.10 – .19

ICU = Inventory of Callous-Unemotional Traits; LPE = "Limited Prosocial Emotions"; ES = average effect size (*r*) across 9 total outcomes; *SE* = standard error; CI = confidence interval.

Table 2. Group-means difference tests across ICU cutoffs in justice-involved sample.

	Low-CU	High-CU	Means-Difference Tests		
	Mean (SD) or % (n)	Mean (SD) or % (n)	df	t or χ2	ES (r)
Empirical Cutoff					
Total physical aggression	3.44 (3.69)	6.35 (6.01)	1145	9.20***	.26
Proactive aggression	.63 (1.17)	1.52 (2.19)	1145	7.85***	.23
Reactive aggression	2.80 (2.83)	4.83 (4.15)	1145	9.07***	.26
Total SR offending	5.72 (8.48)	11.62 (14.24)	1143	7.91***	.23
Violent SR offending	2.30 (3.32)	4.24 (5.28)	1143	6.92***	.20
Gun carrying	.30 (.92)	.64 (1.35)	1143	4.73***	.14
Any gun use during crime	5% (n = 36)	15% (n = 66)	1	31.23***	.17
Total rearrests	1.04 (1.80)	1.56 (2.28)	1182	4.18***	.12
Any violent rearrests	17% (n = 123)	20% (n = 94)	1	1.98	.04
Normative Cutoffs					
Total physical aggression	4.01 (4.28)	8.68 (7.05)	1145	7.59***	.22
Proactive aggression	.80 (1.39)	2.32 (2.82)	1145	6.24***	.18
Reactive aggression	3.42 (3.33)	7.06 (5.50)	1145	4.82***	.14
Total SR offending	7.00 (9.91)	15.45 (17.58)	1143	5.53***	.16
Violent SR offending	2.70 (3.83)	5.68 (6.24)	1143	5.47***	.16
Gun carrying	.35 (.99)	1.01 (1.69)	1143	4.49***	.13
Any gun use during crime	6% (n = 67)	23% (n = 35)	1	52.35***	.21
Total rearrests	1.16 (1.93)	1.80 (2.47)	1182	2.98**	.09
Any violent rearrests	17% (n = 181)	24% (n = 36)	1	4.26*	.06
Rational LPE Approximation					
Total physical aggression	3.84 (4.13)	6.56 (6.24)	1145	7.11***	.21
Proactive aggression	.76 (1.36)	1.57 (2.30)	1145	5.76***	.17
Reactive aggression	3.08 (3.05)	5.00 (4.34)	1145	7.15***	.21
Total SR offending	6.63 (9.33)	11.81 (15.24)	1143	5.58***	.16
Violent SR offending	2.51 (3.61)	4.54 (5.51)	1143	6.00***	.17
Gun carrying	.33 (.98)	.71 (1.38)	1143	4.43***	.13
Any gun use during crime	6% (n = 49)	16% (n = 53)	1	36.09***	.18
Total rearrests	1.16 (1.95)	1.46 (2.18)	1182	2.28*	.07
Any violent rearrests	17% (n = 153)	19% (n = 64)	1	.82	.03

CU = callous-unemotional traits; LPE = "Limited Prosocial Emotions"; SR = self-reported; *SD* = standard deviation; df = degrees of freedom; ES = effect size in the form of Pearson's *r*. χ2 values and *t*-scores from chi-square and independent-samples *T*-tests. ***p < .001, **p < .01, *p < .05

Table 4. Group-means difference tests across ICU cutoffs in school sample.

	Low-CU	High-CU	Means-Difference Tests		
Empirical Cutoff	Mean (SD)	Mean (SD)	df	t	ES (r)
ICU Self-Report					
Conduct problems	24.45 (7.93)	27.10 (8.61)	259	2.27*	.14
Peer rejection	−.15 (.74)	.42 (1.42)	259	3.07**	.19
Peer-nominated meanness	−.12 (.51)	.33 (1.03)	259	3.41**	.21
Normative Cutoffs					
Conduct problems	24.60 (7.77)	30.30 (10.34)	259	2.58*	.16
Peer rejection	−.07 (.82)	.63 (1.90)	259	1.77	.11
Peer-nominated meanness	−.08 (.58)	.67 (1.29)	259	2.76*	.17
Rational LPE Approx.					
Conduct problems	24.81 (8.20)	26.58 (7.93)	259	1.30	.08
Peer rejection	−.09 (.82)	.41 (1.49)	259	2.12*	.13
Peer-nominated meanness	−.06 (.59)	.22 (1.07)	259	1.66	.10
Empirical Cutoff					
Conduct problems	23.27 (6.42)	30.91 (10.63)	287	6.32***	.35
Peer rejection	−.11 (.79)	.28 (1.31)	287	2.59*	.15
Peer-nominated meanness	−.09 (.57)	.20 (.95)	287	2.63*	.15
Normative Cutoffs					
Conduct problems	24.36 (7.15)	36.98 (11.64)	280	6.00***	.34
Peer rejection	−.03 (.92)	.37 (1.46)	280	1.51	.09
Peer-nominated meanness	.03 (.63)	.36 (1.20)	280	1.82	.11
Rational LPE Approx.					
Conduct problems	24.64 (7.24)	36.65 (13.34)	280	4.61***	.27
Peer rejection	.02 (1.00)	.06 (1.08)	280	.20	.01
Peer-nominated meanness	−.01 (.73)	.01 (.72)	280	.05	.003
ICU Parent-Report					
Conduct problems	23.39 (5.88)	34.44 (11.27)	269	6.85***	.39
Peer rejection	−.10 (.82)	.52 (1.37)	269	3.17**	.19
Peer-nominated meanness	−.09 (.54)	.43 (1.11)	269	3.30**	.20
Normative Cutoffs					
Conduct problems	23.72 (6.21)	37.63 (11.10)	269	7.25***	.40
Peer rejection	−.06 (.88)	.57 (1.35)	269	2.71*	.16
Peer-nominated meanness	−.06 (.65)	.45 (.94)	269	3.11**	.19
Rational LPE Approx.					
Conduct problems	23.99 (6.49)	33.81 (12.23)	269	5.07***	.30
Peer rejection	−.11 (.81)	.71 (1.45)	269	3.56***	.21
Peer-nominated meanness	−.09 (.55)	.50 (1.19)	269	3.15**	.19

CU = callous-unemotional traits; SD = standard deviation; df = degrees of freedom; ES = effect size in the form of correlation coefficient (Pearson's *r*); LPE Approx. = "Limited Prosocial Emotions" approximation. χ2 values and *t*-scores from chi-square and independent-samples *T*-tests.
***$p < .001$, **$p < .01$, *$p < .05$.

Table 5. Comparisons of effect size across ICU cutoff methods in school sample.

	ES (*r*)	*SE*	95% CI
ICU Self-Report Cutoff Method			
Empirical Cutoff: ICU total = 24	.18	.04	.11 – .25
Normative Cutoffs			
Male age ≤14 = 34; Male age ≥ 15 = 37	.15	.04	.08 – .22
Female age ≤14 = 29; Female age ≥ 15 = 32			
Rational LPE Approximation	.10	.04	.03 – .17
ICU Parent-Report Cutoff Method			
Empirical Cutoff: ICU total = 21	.18	.08	.02 – .35
Normative Cutoffs:	.18	.08	.02 – .35
Male = 34; Female = 30			
Rational LPE Approximation	.10	.09	−.08 – .27
ICU Teacher-Report Cutoff Method			
Empirical Cutoff: ICU total = 35	.26	.08	.10 – .42
Empirical Cutoff			
Male = 46	.27	.07	.13 – .41
Female = 35			
Rational LPE Approximation	.24	.04	.17 – .31

ICU = Inventory of Callous-Unemotional Traits; LPE = "Limited Prosocial Emotions"; ES = average effect size (*r*) across 3 total outcomes. SE = standard error; CI = confidence interval.

Discussion

While there is substantial research that supports the validity of the ICU as a continuous measure of CU traits, its usefulness for many clinical decisions is limited by the absence of well-validated cutoff scores to determine elevations that predict important clinical criteria. This

limitation has become more problematic as CU traits have been integrated into major systems for diagnosing conduct disorders in children and adolescents, which require a determination of whether clinically elevated levels of the traits are present for an individual (American Psychiatric Association, 2013; World Health Organization, 2018). Thus, in the present study, we tested the clinical utility of three distinct methods for forming ICU cutoffs in two samples of children and adolescents.

First, consistent with past research on ICU total scores (Cardinale & Marsh, 2020), results of ROC analyses showed that total scores on all versions of the ICU (i.e., self-report in the justice-involved sample; self-, parent-, and teacher-report in the school sample) performed significantly better than chance at differentiating youth who showed a number of clinically important outcomes. That is, ICU self-report scores predicted self-reported aggression, self-reported offending, self-reported gun use, and official rearrests over a five-year period after first arrest in a sample of justice-involved adolescents (see Supplementary Figure 1). In addition, total scores for ICU self- and informant-report were associated with conduct problems, peer rejection (parent-report only), and peer-nominated "meanness" in a school sample of older children and young adolescents (see Supplementary Figures 2–4). Of note, there was evidence of the effects of shared method variance in both samples, consistent with past research (e.g., Docherty et al., 2017; Gao & Zhang, 2016; Roose et al., 2010). Specifically, classification accuracies (i.e., AUCs) were highest in analyses with ICU self-report scores predicting self-reported outcomes rather than official arrests in the justice-involved sample (see Supplementary Figure 1), and ICU parent- and teacher-report scores being associated with informant-reported conduct problems in the school sample, rather than peer nominations (see Supplementary Figures 3–4). Such findings emphasize the importance of testing independently measured outcomes when assessing construct validity. Specifically, in our study, ICU self-report still predicted official reports of rearrests in justice-involved adolescents, and all versions of the ICU were associated with peer ratings of meanness in school children.

The high levels of classification accuracy provided by the teacher-report version of the ICU (see Supplementary Figure 4), especially relative to the other versions, are also of importance. Unfortunately, the vast majority of studies using the ICU have tended to use the self-report version. Our findings, along with those of Ueno et al. (2019), suggest that increased use of the ICU teacher-report version is warranted, at least in pre-adolescent samples. For example, Ueno et al. (2019) showed that, in a sample of 955 youth ages 6 to 18 years, teacher-reported ICU scores were associated with youth self-reported levels of psychopathy and externalizing problems as well as parent-reported levels of externalizing problems and oppositionality/aggression. However, our results also suggest that the same cutoffs may not be appropriate across the different informant versions, with teachers generally reporting higher levels of CU traits than either youth or parents, which is consistent with other studies that have compared teacher ratings to either self-report (Allen et al., 2016) or to both self-report and parent-rating (Ueno et al., 2019).

Next, we compared cutoff scores derived from the three methods. The different cutoff methods led to rates of elevations ranging from 13% (normative) to 39% (empirical) for ICU self-report in the justice-involved sample and from 9% (normative self-report) to 31% (empirical parent-report) across all ICU versions in the school sample. While this is a large range of scores, the methods tended to show very high agreement as to who was considered elevated. Further, the average effect sizes testing the validity of these cutoffs derived from the different methods all showed evidence for substantial and similar levels of validity for predicting important outcomes in the justice-involved sample (i.e., aggression, delinquency, and rearrests over five years) and for being associated with important variables in the school sample (i.e., parent- and teacher-reported conduct problems, peer rejection, and peer-nominated meanness).

Thus, our results provide a range of cutoff scores for all versions of the ICU that possess some clinical utility (see Tables 2–5). The choice of cutoff score within this range can be guided by whether more or less stringent cutoffs are desired. Specifically, empirical cutoffs generally resulted in lower cutoffs that led to more youth (i.e., 19–39%) being classified as having elevated CU traits. Such less stringent cutoffs can be used when higher sensitivity (i.e., identifying the most youth who are likely to show clinically relevant or problematic outcomes) is more important than minimizing specificity (i.e., identifying some youth as being at risk who may not show problematic outcomes). Such low cutoff scores are most appropriate for situations where benefits of classification are high (e.g., inclusion in a preventive intervention with minimal side effects) and/or dangers associated with misclassification are low (e.g., placement in an elevated CU group for research). However, higher cutoff scores, which generally result from using normative methods,

are more appropriate when the dangers of misclassification are high (e.g., labeling someone as high-risk for violence in legal proceedings or forensic settings). Of note, we tested a new method for approximating LPE criteria that attempted to overcome limitations in past approximations (see Colins et al., 2020 for a review) by using multiple items for each symptom and considering the different endorsement rates for positively and negatively worded items. Using this method, we obtained evidence for validity that was comparable to other methods for determining cutoffs. However, this new method for approximating the LPE criteria needs to be tested in other samples to determine if it leads to more stable prevalence rates for the specifier.

All of these interpretations should be considered in light of a number of limitations. First, it is important to note that our findings are only meant to guide decisions that require a dichotomous diagnostic decision. For most research purposes, maintaining continuous scores on a measure of CU traits is likely preferred due to the loss of power to detect associations in many cases when dichotomizing continuous scores (Royston et al., 2006). Continuous scores are also likely to be preferred in many clinical settings when diagnostic decisions are not required (e.g., monitoring treatment progress). Further, even when diagnostic decisions are required and a cutoff is used, it is important to recognize that the exact score is not a perfect indicator of a clinical construct due to measurement error and, as a result, will have some level of misclassification associated with it. Second, we used normative cutoffs for the ICU self-report that were based on a large, multinational normative sample of children (Kemp et al., 2019), but such normative samples were not available for ICU parent- or teacher-report versions. Thus, these normative cutoffs were based on the much smaller school-based sample. As noted above, this limitation is especially problematic for establishing cutoff scores, given that classification accuracy can be heavily influenced by the base rate of both the diagnosis (i.e., elevations on the ICU) and the validator (e.g., rate of rearrests). As a result, our findings need to be replicated in other, larger samples that likely vary in their base rate of CU traits. Third, while the justice-involved sample was much larger, it was limited to males arrested for the first time for offenses of moderate severity. Thus, the findings need to be replicated in samples of high-risk females and samples with higher levels of risk generally. Further, as noted previously, classification accuracy can be highly influenced by the base rate of the outcome of interest, and samples of more serious offenders are likely to have much higher rates of rearrest. Fourth, although this study included a number of clinically relevant outcomes (e.g., aggression, juvenile delinquency, informant-reported conduct problems, peer-nominated meanness), it also lacked a number of other validators that research has found to be related to CU traits. Future research should explore the utility of the ICU cutoffs derived in this study with other clinically relevant external criteria. Further, it is important to note that we felt it was important to have validators that did not overlap in method with the ratings on the ICU. This led to us including official arrests in the justice-involved sample and peer nominations in the school sample. However, we did not want to rely only on the well-validated sociometric method for assessing peer rejection since we were concerned that peer rejection may not be specific to persons elevated on CU traits but may be related to conduct problems more generally (Matlasz et al., 2020). Thus, we included the nominations for meanness, which we did feel would be more specifically related to CU traits (Matlasz et al., 2020), but these nominations have not been validated in other samples or with variables other than CU traits.

Finally, and perhaps most importantly, our results were designed to aid in the diagnosis of the LPE specifier. However, it is important to note that the specifier in the DSM-5 requires the presence of Conduct Disorder (American Psychiatric Association, 2013), and the ICD-11 requires the presence of either Conduct-Dissocial Disorder or Oppositional Defiant Disorder (World Health Organization, 2018). Assessment of the symptoms of these disorders are not included on the ICU and, as a result, the diagnosis cannot be made by the ICU alone. Another issue related to making a diagnosis of the LPE specifier is the requirement of the symptoms to be present across multiple settings. While the multi-informant ICU can aid in making this determination, especially by using both parent and teacher versions, the lack of association across informants complicates this interpretation. Specifically, in our school sample the correlations between raters ranged from $r = .24$ (between parent and self and parent and teacher) to $r = .29$ (between self and teacher, all $ps < .01$), which is similar to the level of cross-informant correlations found in the assessment of other emotional and behavioral problems in children and adolescents (De Los Reyes & Kazdin, 2005). However, these differences across informants suggest that many children who are elevated by one informant on the ICU, may not be elevated by another. For example, in our school sample, of the 90 children who were elevated using the most lenient cutoff score for ICU parent-report, 83% were also elevated on self-report, but only 63% were also elevated on teacher-report. We would argue, however, that once an elevation is present by any informant, which our analyses suggest provide some important clinical information, other sources of information can be used to determine if

some level of CU traits are present in other situations. Recommendations for guiding the use of multiple sources of information (e.g., clinical interviews, school records, behavioral observations) in making diagnostic decisions have been made for clinical assessments of youth in general (Frick et al., 2020) and for assessing CU traits specifically (Hawes et al., 2020; Seijas et al., 2018).

Within the context of these limitations, our results support the ICU's association with a number of clinically important outcomes. Further, we utilized and compared multiple methods for determining cutoff scores that validly designate persons at risk for these outcomes across the different rating formats. These cutoffs were similar to others that have been developed in other samples (e.g., Docherty et al., 2017; Kimonis et al., 2014). Thus, research is beginning to accumulate to guide decisions on when ICU scores should lead to concern that a child may be at risk for problematic outcomes and may warrant intervention. Importantly, our work provides a range of scores that can be used, depending on the context and the consequences of false positive and false negative decisions. Such ranges highlight the fact that any single score is associated with some level of error, and no decision that significantly impacts a child's future, clinically, legally, or otherwise, should ever be made on the basis of a single score from any assessment instrument (Frick et al., 2020). However, having empirically derived and tested cutoff scores can provide invaluable information to guide the use of the ICU in making such diagnostic decisions (Youngstrom, 2013).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The Crossroads Study is supported by grants from the Office of Juvenile Justice and Delinquency Prevention [2010-JF-FX -0612], the John D. and Catherine T. MacArthur Foundation [09-94942-000 HCD and 10-95802-000 HCD], the National Institute of Justice, the William T. Grant Foundation, the County of Orange, and the Fudge Family Foundation.

ORCID

Emily C. Kemp http://orcid.org/0000-0002-2008-7938

References

Allen, J. L., Morris, A., & Chhoa, C. Y. (2016). Callous–unemotional (CU) traits in adolescent boys and response to teacher reward and discipline strategies. *Emotional and Behavioural Difficulties*, *21*(3), 329–342. https://doi.org/10.1080/13632752.2016.1165968

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

Blair, R. J. R., Leibenluft, E., & Pine, D. S. (2014). Conduct disorder and callous–unemotional traits in youth. *New England Journal of Medicine*, *371*(23), 2207–2216. https://doi.org/10.1056/NEJMra1315612

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

Ciucci, E., Baroncelli, A., Franchi, M., Golmaryami, F. N., & Frick, P. J. (2014). The association between callous-unemotional traits and behavioral and academic adjustment in children: Further validation of the inventory of callous-unemotional traits. *Journal of Psychopathology and Behavioral Assessment*, *36*(2), 189–200. https://doi.org/10.1007/s10862-013-9384-z

Colins, O. F., Van Damme, L., Hendriks, A. M., & Georgiou, G. (2020). The DSM-5 with limited prosocial emotions specifier for conduct disorder: A systematic literature review. *Journal of Psychopathology and Behavioral Assessment*, *42*(2), 248–258. https://doi.org/10.1007/s10862-020-09799-3

Crapanzano, A. M., Frick, P. J., Childs, K., & Terranova, A. M. (2011). Gender differences in the assessment, stability, and correlates to bullying roles in middle school children. *Behavioral Sciences & the Law*, *29*(5), 677–694. https://doi.org/10.1002/bsl.1000

De Los Reyes, A. (2017). Inaugural editorial: Making the journal of clinical child & adolescent psychology your "Home Journal." Taylor & Francis. https://doi.org/10.1080/15374416.2016.1266649

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483–509. https://doi.org/10.1037/0033-2909.131.4.483

Docherty, M., Boxer, P., Huesmann, L. R., O'Brien, M., & Bushman, B. (2017). Assessing callous-unemotional traits in adolescents: Determining cutoff scores for the inventory of callous and unemotional traits. *Journal of Clinical Psychology*, *73*(3), 257–278. https://doi.org/10.1002/jclp.22313

Emmert, A. D., Hall, G. P., & Lizotte, A. J. (2018). Do weapons facilitate adolescent delinquency? An examination of weapon carrying and delinquency among adolescents. *Crime and Delinquency*, *64*(3), 342–362. https://doi.org/10.1177/0011128717714466

Fanti, K. A., & Kimonis, E. R. (2012). Bullying and victimization: The role of conduct problems and psychopathic traits. *Journal of Research on Adolescence*, *22*(4), 617–631. https://doi.org/10.1111/j.1532-7795.2012.00809.x

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Frick, P. J., Barry, C. T., & Kamphaus, R. W. (2020). *Clinical assessment of child and adolescent personality and behavior* (4th ed.). Springer.

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014a). 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

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014b). A developmental psychopathology approach to understanding callous-unemotional traits in children and adolescents with serious conduct problems. *Journal of Child Psychology and Psychiatry*, *55*(6), 532–548. https://doi.org/10.1111/jcpp.12152

Gao, Y., & Zhang, W. (2016). Confirmatory factor analyses of self-and parent-report inventory of callous-unemotional traits in 8-to 10-year-olds. *Journal of Psychopathology and Behavioral Assessment*, *38*(3), 331–340. https://dx.doi.org/10.1007%2Fs10862-015-9527-5

Golmaryami, F. N., Frick, P. J., Hemphill, S. A., Kahn, R. E., Crapanzano, A. M., & Terranova, A. M. (2016). The social, behavioral, and emotional correlates of bullying and victimization in a school-based sample. *Journal of Abnormal Child Psychology*, *44*(2), 381–391. https://doi.org/10.1007/s10802-015-9994-x

Graziano, P. A., Ros, R., Haas, S., Hart, K., Slavec, J., Waschbusch, D., & Garcia, A. (2016). Assessing callous-unemotional traits in preschool children with disruptive behavior problems using peer reports. *Journal of Clinical Child & Adolescent Psychology*, *45*(2), 201–214. https://doi.org/10.1080/15374416.2014.971460

Hare, R. D., & Neumann, C. S. (2008). Psychopathy as a clinical and empirical construct. *Annual Reviews in Clinical Psychology*, *4*(1), 217–246. https://doi.org/10.1146/annurev.clinpsy.3.022806.091452

Hawes, D. J., Kimonis, E. R., Diaz, A. M., 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

Huizinga, D., Esbensen, F. A., & Weiher, A. W. (1991). Are there multiple paths to delinquency? *The Journal of Criminal Law & Criminology*, *82*(1), 83–118. https://doi.org/10.2307/1143790

Kemp, E. C., Frick, P. J., & Robertson, E. L. (2019). *Inventory of callous-unemotional traits (ICU) - self-report youth version: T-scores & percentiles* [Normative dataset]. https://sites01.lsu.edu/faculty/pfricklab/icu/

Kimonis, E. R., Fanti, K. A., Frick, P. J., Moffitt, T. E., Essau, C., Bijttebier, P., & Marsee, M. A. (2015). Using self-reported callous-unemotional traits to cross-nationally assess the DSM-5 'with limited prosocial emotions' specifier. *Journal of Child Psychology and Psychiatry*, *56*(11), 1249–1261. https://doi.org/10.1111/jcpp.12357

Kimonis, E. R., Fanti, K. A., & Singh, J. P. (2014). Establishing cut-off scores for the parent-reported inventory of callous-unemotional traits. *Archives of Forensic Psychology*, *1*(1), 27–48. https://www.researchgate.net/profile/Kostas-Fanti/publication/303637241_Establishing_cut-off_scores_for_the_parent-reported_inventory_of_callous-unemotional_traits/links/60422f4992851c077f18c1f2/Establishing-cut-off-scores-for-the-parent-reported-inventory-of-callous-unemotional-traits.pdf

Kimonis, E. R., Frick, P. J., Skeem, J. L., Marsee, M. A., Cruise, K., Munoz, L. C., … 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

Kliem, S., Lohmann, A., Neumann, M., Glaubitz, C., Haselbach, S., Bergmann, M. C., & Baier, D. (2020). Factor structure of the inventory of callous-unemotional traits in a representative sample of German 9th grade students. *Journal of Abnormal Child Psychology*, *48*(1), 43–55. https://doi.org/10.1007/s10802-019-00590-x

Koutsogiorgi, C. C., Lordos, A., Fanti, K. A., & Michaelides, M. P. (2020). Factorial structure and nomological network of the inventory of callous-unemotional traits accounting for item keying variance. *Journal of Personality Assessment*, *103*(3), 1–12. https://doi.org/10.1080/00223891.2020.1769112

Lawing, K., Frick, P. J., & Cruise, K. R. (2010). Differences in offending patterns between adolescent sex offenders high or low in callous—unemotional traits. *Psychological Assessment*, *22*(2), 298. https://doi.org/10.1037/a0018707

Marsee, M. A., Barry, C. T., Childs, K. K., Frick, P. J., Kimonis, E. R., Muñoz, L. C., … Lau, K. S. (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., Barry, C. T., Kimonis, E. R., Centifanti, L. C. M., & Aucoin, K. J. (2014). Profiles of the forms and functions of self-reported aggression in three adolescent samples. *Development and Psychopathology*, *26*(3), 705–720. https://doi.org/10.1017/S0954579414000339

Matlasz, T. M., Frick, P. J., & Clark, J. E. (2020). Understanding the social relationships of youth with callous-unemotional traits using peer nominations. *Journal of Clinical Child and Adolescent Psychology*, 1–13. https://doi.org/10.1080/15374416.2020.1823847

McMullen, J. A., Veermans, K., & Laine, K. (2014). Tools for the classroom? An examination of existing sociometric methods for teacher use. *Scandinavian Journal of Educational Research*, *58*(5), 624–638. https://doi.org/10.1080/00313831.2013.838694

Pelham, W. E., Jr., Gnagy, E. M., Greenslade, K. E., & Milich, R. (1992). Teacher ratings of DSM-III-R symptoms for the disruptive behavior disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*(2), 210–218. https://doi.org/10.1097/00004583-199203000-00006

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. (2018). 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). https://doi.org/10.1080/15374416.2018.1504297

Ray, J. V., Frick, P. J., Thornton, L. C., Steinberg, L., & Cauffman, E. (2016). Positive and negative item wording and its influence on the assessment of callous-unemotional traits.

Psychological Assessment, *28*(4), 394. https://doi.org/10.1037/pas0000183

Robertson, E. L., Frick, P. J., Walker, T. M., Kemp, E. C., Ray, J. V., Thornton, L. C., . . . Cauffman, E. (2020). Callous-unemotional traits and risk of gun carrying and use during crime. *American Journal of Psychiatry*, *177*(9), 827–833. https://doi.org/10.1176/appi.ajp.2020.19080861

Roose, A., Bijttebier, P., Decoene, S., Claes, L., & Frick, P. J. (2010). Assessing the affective features of psychopathy in adolescence: A further validation of the inventory of callous and unemotional traits. *Assessment*, *17*(1), 44–57. https://doi.org/10.1177/1073191109344153

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*(1), 127–141. https://doi.org/10.1002/sim.2331

Saukkonen, S., Laajasalo, T., Jokela, M., Kivivuori, J., Salmi, V., & Aronen, E. T. (2016). Weapon carrying and psychopathic-like features in a population-based sample of Finnish adolescents. *European Child & Adolescent Psychiatry*, *25*(2), 183–191. https://doi.org/10.1007/s00787-015-0724-2

Seijas, R., Servera, M., García-Banda, G., Barry, C. T., & Burns, G. L. (2018). Evaluation of a four-item DSM–5 limited prosocial emotions specifier scale within and across settings with Spanish children. *Psychological Assessment*, *30*(4), 474. https://doi.org/10.1037/pas0000496

Ueno, K., Ackermann, K., Freitag, C. M., & Schwenck, C. (2019). Assessing callous–unemotional traits in 6-to 18-year-olds: Reliability, validity, factor structure, and norms of the German version of the inventory of callous–unemotional traits. *Assessment*, *28*(2), 567–584. https://doi.org/10.1177/1073191119847766

Viding, E., Sebastian, C. L., Dadds, M. R., Lockwood, P. L., Cecil, C. A., De Brito, S. A., & McCrory, E. J. (2012). Amygdala response to preattentive masked fear in children with conduct problems: The role of callous-unemotional traits. *American Journal of Psychiatry*, *169*(10), 1109–1116. https://doi.org/10.1176/appi.ajp.2012.12020191

Waller, R., Baskin-Sommers, A. R., & Hyde, L. W. (2016). Examining predictors of callous unemotional traits trajectories across adolescence among high-risk males. *Journal of Clinical Child & Adolescent Psychology*, *47*(3), 1–14. https://doi.org/10.1080/15374416.2015.1102070

World Health Organization. (2018). *The ICD-11 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*.

Yoshida, S., Adachi, M., Takahashi, M., Takanyanagi, N., Yasuda, S., Osada, H., & Nakamura, K. (2019). The factor structure and construct validity of the parent-reported inventory of callous-unemotional traits among school-aged children and adolescents. *PloS One*, *14*(8), 1–18. https://doi.org/10.1371/journal.pone.0221046

Youngstrom, E. A. (2013). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221. https://doi.org/10.1093/jpepsy/jst062