

# Positive and Negative Item Wording and Its Influence on the Assessment of Callous-Unemotional Traits

James V. Ray  
University of Texas, San Antonio

Paul J. Frick  
University of New Orleans and Australian Catholic University

Laura C. Thornton  
University of New Orleans

Laurence Steinberg  
Temple University and King Abdulaziz University

Elizabeth Cauffman  
University of California, Irvine

This study examined the item functioning of the Inventory of Callous-Unemotional Traits (ICU) in an ethnically diverse sample 1,190 of first-time justice-involved adolescents (mean age = 15.28 years,  $SD = 1.29$ ). On elimination of 2 items, the total ICU score provided a reliable (internally consistent and stable) and valid (correlated with and predictive of measures of empathy, school conduct problems, delinquency, and aggression) continuous measure of callous and unemotional (CU) traits. A shortened, 10-item version of the total scale, developed from item response theory (IRT) analyses, appeared to show psychometric properties similar to those of the full ICU and, thus, could be used as an abbreviated measure of CU traits. Finally, item analyses and tests of validity suggested that the factor structure of the ICU reported in a large number of past studies could reflect method variance related to the ICU, including equal numbers of positively and negatively worded items. Specifically, positively worded items (i.e., items for which higher ratings are indicative of higher levels of CU traits) were more likely to be rated in the lower response categories, showed higher difficulty levels in IRT analyses (i.e., discriminated best at higher levels of CU traits), and were more highly correlated with measures of antisocial and aggressive behavior. On the basis of these findings, we recommend using the total ICU as a continuous measure of CU traits and do not recommend continued use of the subscale structure that has been reported in multiple past studies.

**Keywords:** Inventory of Callous-Unemotional Traits, item response theory, ICU-10, method variance, with limited prosocial emotions

Callous and unemotional (CU) traits (e.g., lack of empathy and remorse, shallow affect, lack of concern over performance in important activities) demarcate a unique subgroup of antisocial youths whose behavior tends to be more severe, chronic, and aggressive (Frick, Ray, Thornton, & Kahn, 2014; Pardini & Fite,

2010) than that of other young people. Youths who exhibit higher levels of CU traits also engage in more calculating and cold-blooded forms of aggression (i.e., proactive aggression; Marsee & Frick, 2007); show little concern for the feelings of others (Pardini & Byrd, 2012); and show little concern for the negative consequences of their actions, even when punishment is imminent (D. J. Hawes & Dadds, 2005). In addition, antisocial youths with CU traits show a number of genetic, neurocognitive, emotional, personality, and social differences compared with other antisocial youths, suggesting that the factors leading to the development of antisocial behavior may differ across the two groups (for a review, see Frick et al., 2014). On the basis of this extensive body of research, the most recent revision of the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed. [DSM-5]; American Psychiatric Association, 2013) integrated CU traits into the diagnostic criteria for conduct disorder. The criteria now include the specifier “with limited prosocial emotions” to designate those youths with serious conduct problems who also show elevated levels of CU traits.

In light of this recent change, there is now a greater emphasis on developing and testing optimal methods for assessing CU traits. In research to date, CU traits have been assessed using several dif-

---

This article was published Online First June 29, 2015.

James V. Ray, Department of Criminal Justice, University of Texas at San Antonio; Paul J. Frick, Department of Psychology, University of New Orleans, and Learning Sciences Institute of Australia, Australian Catholic University; Laura C. Thornton, Department of Psychology, University of New Orleans; Laurence Steinberg, Department of Psychology, Temple University, and King Abdulaziz University; Elizabeth Cauffman, Department of Psychology and Social Behavior, University of California, Irvine.

The Crossroads Study is supported by grants from the U.S. Department of Justice’s Office of Juvenile Justice and Delinquency Prevention (2010-JF-FX-0612) and the John D. and Catherine T. MacArthur Foundation (09-94942-000 HCD and 10-95802-000 HCD). We are grateful to the many individuals responsible for the data collection and preparation.

Correspondence concerning this article should be addressed to James V. Ray, Department of Criminal Justice, University of Texas at San Antonio, San Antonio, TX 78207. E-mail: [james.ray@utsa.edu](mailto:james.ray@utsa.edu)

ferent measures and assessment formats (Frick, 2009). However, most of these measures assess these traits as one part of the broader construct of psychopathy. As a result, the number of items specifically assessing CU traits has often been limited to as few as four (Forth, Kosson, & Hare, 2003) or six (Frick, Bodin, & Barry, 2000). Further, the response options for rating the severity or frequency of the CU items on these scales has also been limited, often with only three options (Forth et al., 2003; Frick et al., 2000). The few items and the limited range in response options have led to significant psychometric limitations in many measures, such as poor internal consistency (Poythress, Dembo, Wareham, & Greenbaum, 2006).

The Inventory of Callous-Unemotional Traits (ICU; Kimonis et al., 2008) was developed to overcome these limitations. The ICU was developed from the Antisocial Process Screening Device (APSD; Frick & Hare, 2001). Specifically, the ICU was developed from the four items from the callous-unemotional subscale of the APSD, which (a) loaded most consistently on a callous-unemotional factor across raters and samples (Frick et al., 2000) and (b) formed the basis for the *DSM-5* specifier (Kimonis et al., 2014). To form the items on the ICU, six items (three positively worded and three negatively worded) were developed to assess a similar content to each of the four core traits. These 24 items were then placed on a four-point scale that could be rated from 0 (*not at all true*) to 3 (*definitely true*). Thus, the ICU content was designed to provide a continuous measure of CU traits, similar to how they have been operationalized in a great amount of past work and to how they are operationalized for the *DSM-5* specifier. However, the content was also designed to overcome limitations of past measures by (a) including a more comprehensive assessment of CU traits (i.e., 24 items), (b) using an expanded anchor system (i.e., four points), and (c) including an equal number of items worded in the positive (e.g., “I do not care who I hurt to get what I want”) and the negative (e.g., “I am concerned about the feelings of others”) direction.

Several studies have tested the construct validity of the ICU using factor analyses and have generally reported that the best fitting factor structure is one specifying a general CU factor and three subfactors: callousness (a lack of empathy and remorse), uncaring (an uncaring attitude about performance on tasks and others’ feelings), and unemotional (deficient emotional affect). This structure was supported in a sample of 1,443 German community adolescents ages 13–18 years (Essau, Sasagawa, & Frick, 2006), a sample of 248 juvenile offenders ages 12–20 from the United States (Kimonis et al., 2008), and a community sample of 347 Greek Cypriot adolescents between the ages of 12 and 18 (Fanti, Frick, & Georgiou, 2009). A similar factor structure has been found in younger samples, including a sample of 540 Italian middle-school children (Ciucci, Baroncelli, Franchi, Golmaryami, & Frick, 2014) and, using parent report, samples of 540 high-risk 9-year-olds in the United States (Waller et al., 2014) and 622 Spanish preschool (3-year-old) children (Ezpeleta, de la Osa, Grano, Penelo, & Domènech, 2013). Further, Roose, Bijttebier, Decoene, Claes, and Frick (2010) reported that this factor structure was similar for both self- and other (i.e., parent and teacher) reports in a community sample of 455 Dutch adolescents (ages 14–20), and both Essau et al. (2006) and Ciucci et al. (2014) reported that the structure was invariant across boys and girls.

In summary, it appears that this three-dimensional structure of the ICU is robust across age, language, rater, and gender (for exceptions, see Feilhauer, Cima, & Arntz, 2012; Houghton, Hunter, & Crow, 2013), and these analyses support the presence of an overarching dimension of CU traits. Further, the total score of the ICU has shown acceptable internal consistency (Cronbach’s alphas ranging between .77 and .89) and similar correlations with antisocial behavior and other emotional and cognitive characteristics to those that have been reported in studies using other measures of CU traits (Frick et al., 2014). Thus, the total score from the ICU provides a continuous measure of the overall construct of CU traits that overcomes many of the limitations in past measures.

However, the available research also highlights several significant limitations of the ICU scale. First, although the three subfactors, with an overarching general factor, consistently emerge as the best fitting factor structure across diverse samples, the fit indices tend to be modest, typically only reaching acceptable fit after elimination of certain items with poor item-total correlations and after post hoc modifications are made to the model (S. W. Hawes et al., 2014; Kimonis et al., 2008; Waller et al., 2014). Second, there was no a priori specification for this three-factor structure for CU traits based on a clear theoretical model, nor have the subfactors shown consistent associations with external criteria (Frick & Ray, 2014; Waller et al., 2014). Third, it is possible that the factors to some extent represent shared-method variance in that the callousness dimension tends to be largely positively scored items (e.g., “I do not care who I hurt to get what I want”), whereas the uncaring dimension tends to be largely negatively scored items (e.g., “I try not to hurt others’ feelings”); S. W. Hawes et al., 2014).

This last finding, that positively and negatively worded items load on different factors, is especially critical for determining whether CU traits are better considered a multi- or a unidimensional construct. The goal in developing the ICU was to have equal numbers of positive and negatively worded items to minimize the influence of response sets by forcing informants to consider the direction of ratings across items (Kimonis et al., 2008). However, it is possible that this methodology leads to very different distributions of item endorsement for positively and negatively worded items. That is, people may be less likely to endorse extreme levels of callousness (e.g., “I do not care who I hurt to get what I want”) than they are to endorse very low levels of prosocial emotions (e.g., “I try not to hurt others’ feelings”). In terms of item response theory (IRT), this could indicate that the positively worded items (i.e., items for which higher ratings indicate more CU traits) are more difficult and, thus, are more discriminating at higher levels of the latent trait and that negatively worded items (i.e., items for which higher ratings indicate lower levels of CU traits) are less difficult, with optimal discrimination obtained at lower levels along the latent trait (de Ayala, 2009). The *difficulty* of an item in IRT refers to the level of the latent trait necessary to either endorse the item or endorse a higher response category (or a lower response category if the item is negatively worded; Embretson & Reise, 2000). Applied to the ICU, items that are more difficult would require higher levels of CU traits to positively endorse them. Item functioning is important for the predictive validity of the full scale, because it is generally considered important to have a test with items showing a range of difficulty levels (Embretson & Reise, 2000). However, if the factors that have been identified in past samples simply reflect different levels of item difficulty, the

separate factors would then be theoretically unimportant for understanding the construct.

Although item difficulty may account for the scale structure of the ICU, there could also be substantive reasons for the findings from past factor analyses. Many early definitions of *CU traits*, defined as indicators of an “undersocialized” interpersonal style (Quay, 1993), used the absence of significant prosocial features to define the construct rather than the presence of indicators of a CU interpersonal style. This methodology is exemplified in the definition for *undersocialized conduct disorder* in the *DSM-III* (American Psychiatric Association, 1980), which, unlike the criteria for other disorders requiring the presence of a certain number of symptoms, required that no more than one of a list of prosocial features could be present (e.g., “extends him or herself for others even when no immediate advantage is likely,” “apparently feels guilt or remorse when such a reaction is appropriate [not just when caught or in difficulty]”). Thus, it could be that lower levels of prosocial features are a better indicator of the construct than are higher levels of CU features.

To investigate these potential reasons for the subscale structure of the ICU, we conducted a test of the differential item functioning on the self-report version of the scale in a large and ethnically diverse sample of justice-involved adolescents assessed at two points in time over a 6-month period. We conducted a number of tests of item functioning using analyses based on classical test theory (e.g., item distributions, internal consistency of items, test-retest reliability) as well as analyses using IRT. We tested whether positively and negatively worded items showed different distributions and/or item difficulty. Further, we tested whether using only positively worded or negatively worded items led to better reliability and construct validity, relative both to each other and to the full scale combining both types of item formats. Our tests of validity included tests of convergent validity (i.e., correlations with a measure of empathy) and criterion validity (i.e., positive correlations with measures of aggression and antisocial behavior and negative correlations with anxiety). In these tests of validity, correlations at baseline were examined, as were predictive associations between the ICU scales and the external criteria at the 6-month follow-up, controlling for initial levels of the criterion measures.

Finally, S. W. Hawes et al. (2014) posited that the limitations in the factor structure of the ICU could suggest that the total score could be refined by limiting it to items that show the best discrimination in the overall level of CU traits. Specifically, in a sample of 250 boys ages 6–12, they selected 12 items from the parent-report version of the ICU on the basis of IRT analyses and found that this refined scale showed high internal consistency and support for its validity in terms of correlations with measures of conduct problems and social competence. Importantly, the significantly shorter revised total scale and the full 24-item scale exhibited nearly identical correlations with external measures (see also Waller et al., 2014, who also used the parent-report version). We followed a similar procedure in the present study to determine whether a similar shortened version of the ICU could be developed using IRT analyses in an older sample, using the self-report version of the ICU, and whether it too would show levels of reliability and validity similar to those of the full version of the scale.

## Method

### Participants

The Crossroads Study is an ongoing multisite longitudinal study examining the outcomes of juveniles who have either been formally or informally processed for offenses of moderate severity in Jefferson Parish, Louisiana; Orange County, California; and Philadelphia, Pennsylvania. To be eligible for the Crossroads Study, juveniles had to be first-time male offenders, be English speakers between the ages of 13 and 17 years at the time of arrest, and have an eligible offense. Eligible charges were midrange offenses, such as theft of goods, simple battery, and vandalism. Across sites, 72.32% ( $n = 1,216$ ) of individuals eligible to participate enrolled in the study. The current sample comprised 1,190 participants who participated in the first two waves (6 months apart) of the study (retention rate = 97.9%). The breakdown of participants across the three sites was as follows: Jefferson Parish ( $n = 149$ ), Orange County ( $n = 524$ ), and Philadelphia ( $n = 517$ ). The mean age of participants at the first assessment was 15.28 ( $SD = 1.29$ ). The sample was predominately White Latino (45.9%) and Black (36.8%), followed by White non-Latino (14.7%) and other (2.5%).

### Procedures

Institutional review board approval was obtained at each site before data collection began. Consent was obtained from a parent of each juvenile, and assent was obtained from the participant. The parent and youth were informed that the research project had obtained a certificate of confidentiality from the U.S. Department of Justice, which allowed the research information to be protected from being subpoenaed for use in legal proceedings. Interviews were administered from a laptop at a location convenient to the youth (e.g., participant’s home, retail location, one of the participating universities). Regardless of location, the interview was conducted in a private setting, with just the interviewer and participant present to reduce socially desirable responding. Participants were able to read the items themselves directly from the computer screen; however, to control for reading ability, items were also read aloud to participants by interviewers. In addition, participants were informed that they did not have to answer any questions that made them feel uncomfortable. Participants were paid \$50.00 for the first interview and \$65.00 at the 6-month follow-up interview.

### Measures

Descriptive statistics on all criterion variables used in the current study, including internal consistency estimates, are reported in Table 1.

**CU traits: ICU.** The ICU (Kimonis et al., 2008) is a 24-item instrument derived from the APSD (Frick & Hare, 2001), which is a rating scale commonly used to assess CU traits in children and adolescents. Participants rated items on a four-point scale ranging from 0 (*not at all true*) to 3 (*definitely true*). The ICU content is presented at a 6th-grade reading level. As noted earlier, use of the total score on the ICU has been supported in factor analyses conducted with both detained (Kimonis et al., 2008) and community (Essau et al., 2006; Fanti et al., 2009) samples of adolescents.

Table 1  
Descriptive Statistics for Criterion Measures

Measure	<i>M</i> ( <i>SD</i> )	Range	Skewness ( <i>SE</i> )	Kurtosis ( <i>SE</i> )	Cronbach's $\alpha$	<i>n</i>
Baseline						
Consideration of others	25.42 (5.38)	7–35	–0.65 (0.07)	0.48 (0.14)	.69	1,190
Proactive aggression	2.81 (4.76)	0–45	3.31 (0.07)	15.00 (0.14)	.87	1,190
Reactive aggression	6.98 (7.07)	0–47	1.74 (0.07)	4.03 (0.14)	.87	1,190
Self-reported offending	3.43 (3.08)	0–19	1.82 (0.07)	3.98 (0.14)	.82	1,190
School misconduct	5.36 (1.59)	0–7	–1.05 (0.08)	0.78 (0.16)	.63	894
Anxiety	12.6 (4.06)	7–28	1.06 (0.07)	1.01 (0.14)	.81	1,190
6-month follow-up						
Consideration of others	25.25 (5.37)	7–35	–0.55 (0.07)	0.34 (0.14)	.69	1,165
Proactive aggression	2.49 (5.00)	0–48	4.13 (0.07)	22.25 (0.14)	.88	1,165
Reactive aggression	5.96 (7.01)	0–57	2.22 (0.07)	2.24 (0.14)	.90	1,165
Self-reported offending	3.43 (3.08)	0–18	1.82 (0.07)	9.01 (0.14)	.81	1,165
School misconduct	15.08 (4.21)	10–30	1.10 (0.08)	0.98 (0.15)	.75	1,049
Anxiety	12.06 (4.14)	7–28	1.13 (0.07)	1.05 (0.14)	.84	1,165

Note.  $N = 1,190$ .

Further, the ICU has correlated positively with antisocial behavior and negatively with prosocial behavior (Essau et al., 2006; Fanti et al., 2009; Kimonis et al., 2008; Roose et al., 2010). Importantly, the same factor structure and evidence to support the validity of the scores have been found when items are read aloud to an adolescent (Kimonis et al., 2008) and when the adolescent reads the items him- or herself (Essau et al., 2006).

**Convergent validity: Weinberger Adjustment Inventory (WAI; Weinberger & Schwartz, 1990).** The current study used the seven-item consideration of others (e.g., “Doing things to help other people is more important to me than almost anything else”) subscale of the WAI as a measure of convergent validity for the ICU. This subscale has been negatively correlated with measures of delinquency ( $r = -.20$ ) and aggression ( $r = -.22$ ; Farrell & Sullivan, 2000) and positively correlated with other measures of empathy ( $r_s = .57-.64$ ; Espelage, Mebane, & Adams, 2004).

**Criterion validity: Revised Child Anxiety and Depression Scale (RCADS; Chorpita, Yim, Moffitt, Umemoto, & Francis, 2000).** The RCADS was developed to assess anxiety and major depressive disorders consistent with *DSM-IV* (American Psychiatric Association, 1994) diagnoses. Although the RCADS consists of six subscales, only the seven items comprised by the generalized anxiety disorder (GAD) subscale (e.g., “I worry that bad things will happen to myself”) were used in the current study. GAD scores were summed scores on those seven items, with higher scores indicating greater anxiety. The GAD subscale of the RCADS has shown convergent validity with other measures of anxiety (Chorpita, Moffitt, & Gray, 2005).

**Criterion validity: Peer Conflict Scale (PCS; Marsee & Frick, 2007).** The PCS is a 40-item instrument that was designed to provide a comprehensive measure of two forms of aggression: reactive (e.g., “If others make me mad, I hurt them”) and proactive (e.g., “I carefully plan out how to hurt others”). Items are rated on a four-point scale. In support of the validity of PCS subscales, separate factors for the two forms of aggression have emerged across multiple samples of both detained and community adolescents (Marsee et al., 2014), and scores on the aggression subscales have been associated with laboratory measures of aggressive behavior, with the reactive and proactive subtypes showing different

responses to provocation in a sample of detained boys (Muñoz, Frick, Kimonis, & Aucoin, 2008). Moreover, scores on the reactive and proactive scales have been positively associated with self-reported number of violent acts (Kimonis et al., 2008).

**Criterion validity: Self-Reported Offending (SRO; Denver Youth Survey, Huizinga, Esbensen, & Weiher, 1991).** The SRO consists of 24 dichotomous items asking participants if they have engaged in different types of crime (e.g., shoplifting, assault, burglary). At baseline, respondents were asked if they had “ever” engaged in these types of offenses, and at follow-up, they were asked about the past “6 months.” The responses were summed to obtain a variety of offending total score, with higher scores indicative of more diverse offending. The SRO has shown positive correlations with official reports of offending (Thornberry & Krohn, 2000).

**School misconduct.** A measure of school misconduct was derived from nine items asking participants about school conduct problems (e.g., received detention, cheated on test, skipped school without permission). At baseline, participants responded to questions that asked if they had “ever” engaged in different types of behaviors with either a “yes” (1) or a “no” (0), and the responses were summed for a total variety measure of school misconduct, with higher scores indicating higher levels of variety of school misconduct. At follow-up, respondents indicated on a scale ranging from 1 (*not at all*) to 4 (*often/many times*) how often they had engaged in each behavior over the past 6 months.

## Data Analyses

To address the study goals, we first calculated Cronbach's alphas and item-total correlations (ITCs) for the ICU at baseline. This information, mainly the corrected ITCs (i.e., ITCs eliminating the item of interest from the total score), was also used to determine if any items were providing poor discrimination. Although it is typically recommended that items with ITCs below .30 be regarded as discriminating poorly (Nunnally & Bernstein, 1994), we took a more inclusive approach so as to retain as many of the original items from the ICU as possible and only removed those that both (a) were below .10 and (b) had been shown to be poorly



functioning across different samples on the basis of prior research. To further assess the functioning of items on the ICU, we compared the patterns in responses between the negatively and positively worded items. This was done by totaling the frequencies of each response category for the negatively and positively worded items and comparing their distributions.

Next, we subjected ICU items to IRT analysis. Maximum-likelihood estimation was used, which assumes that any missing data are missing completely at random; however, in the current sample, missing data were minimal (only one ICU item from one participant). Because of the polytomous nature of the items on the ICU, they were examined using the graded response model, which allows for the application of IRT to items with ordered categorical response sets (Samejima, 1997), in Mplus Version 5 (Muthén & Muthén, 1998–2007). Using the two-parameter logistic model, we were able to examine item slope, or discrimination, parameters ( $as$ ) and threshold, or item difficulty, parameters ( $bs$ ). The steepness of a slope indicates the strength of the relation between item response categories and the latent trait. That is, greater slope parameters indicate that less change in the latent trait results in increased likelihood of responding in the next highest category and, thus, a more discriminating item. For  $k$  response categories, there are  $k - 1$  slope parameters. Item threshold values describe where along the latent trait continuum the likelihood of responding in a certain response category is highest. That is, higher threshold values indicate more difficult items (i.e., items requiring higher levels of the latent trait), and vice versa. The IRT analysis was used to examine whether the negatively and positively worded items were functioning differently on the basis of their difficulty parameters. In addition, the slope parameters were examined to identify poorly functioning items so as to develop a short form of the ICU. It has been suggested that items with slope parameters lower than .90 can be considered poorly functioning (de Ayala, 2009). Thus, the items that had slope parameters of .90 or higher were retained and included in a shortened version of the ICU.

Finally, we examined the psychometric properties of a shortened ICU measure based on these analyses. That is, we created a shortened version of the ICU by selecting those items that best assessed the latent trait of CU traits across the full range of potential levels according to the IRT analysis. We compared the psychometric properties of this shortened scale with those of the full ICU and with those of shortened versions using only positively worded items and only negatively worded items. We compared these four versions on their internal consistency, stability over 6 months, convergent validity (i.e., correlations with a measure of empathy), and criterion validity (i.e., correlations with measures of antisocial behavior and anxiety).

## Results

### Item Functioning of the ICU

Cronbach's alphas and corrected ITCs were examined to assess the internal consistency of the 24 items of the ICU at baseline. For all 24 items, Cronbach's alpha was .77. In addition, the corrected ITCs for all 24 items ranged from .03 (Item 2) to .48 (Item 16), with a mean of .31. Items 2 and 10 had the lowest ITCs (.03 and .10, respectively). These low ITCs were consistent with findings from prior research that these two items do not correlate highly

with the total score (e.g., S. W. Hawes et al., 2014; Kimonis et al., 2008), and, thus, reliability analyses were rerun after removal of the items. After Items 2 and 10 were removed, Cronbach's alpha increased to .78, and the corrected ITCs ranged from .13 (Item 13) to .50 (Item 16), with a mean of .33.

The frequencies for the item response categories on the ICU are provided in Figure 1 (top), separately for positively and negatively worded items. These analyses were all conducted after the negatively worded items were reverse scored such that higher scores were indicative of higher CU traits for all items. From this graph, it is clear that positively worded items were much more likely to be endorsed with the lowest rating for CU traits. For example, a positively worded item (e.g., "I do not care who I hurt to get what I want") was much more likely to be rated 0 (*not at all true*) than a negatively worded item (e.g., "I try not to hurt others' feelings") was to be rated 3 (*definitely true*). In contrast, negatively worded items had a much higher frequency of endorsement at the two highest levels indicating CU traits. This was most evident at the second highest level. That is, a negatively worded item (e.g., "I feel bad or guilty when I do something wrong") was more likely to be rated 1 (*somewhat true*) than a positively worded item (e.g., "I do not care if I get in trouble") was to be rated 2 (*very true*).

To illustrate the strength of this difference, a repeated-measures analysis of variance was conducted comparing the average scores for negatively worded ( $M = 1.41$ ) and positively worded items ( $M = .78$ ; see Figure 1, bottom). These analyses indicated that the average score of the negatively worded items was significantly higher than the means for positive items,  $F(1, 1,189) = 1,349.55$ ,  $p < .001$ , and that this was a strong effect ( $\eta_p^2 = .532$ ). Further, Figure 1 (bottom) shows a comparison of skewnesses and kurtoses

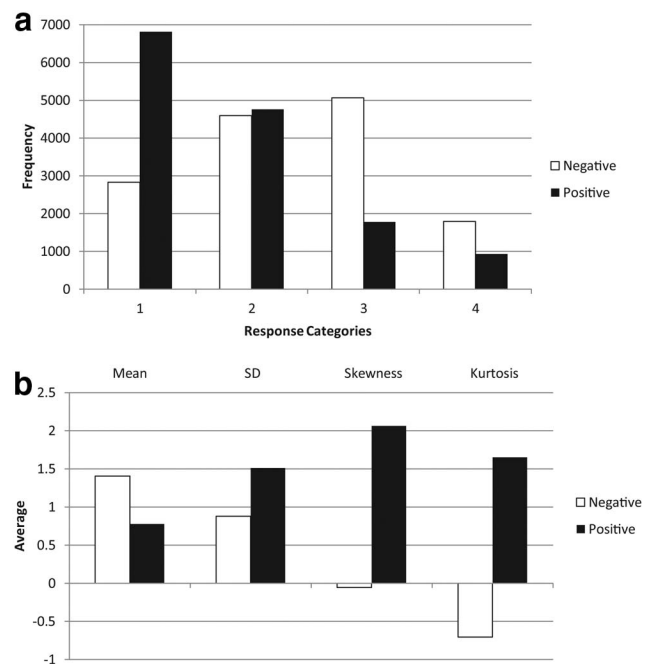


Figure 1. Top: Item response frequencies for positively and negatively worded items separately. 1 = *not at all true*; 2 = *somewhat true*; 3 = *very true*; 4 = *definitely true*. Bottom: Item means, standard deviations, skewnesses, and kurtoses for positively and negatively worded items.

of the mean item scores. Within this sample of detained adolescents, both of these distributions differed significantly from a normal distribution (Shapiro–Wilk scores = .94 and .99, respectively,  $ps < .001$ ). However, the negatively worded items showed a negative skew ( $-0.23$ ), whereas the positively worded items showed a positive skew ( $0.94$ ).

Next, IRT analyses of the 22 ICU items retained for further analysis were conducted; the item parameters are reported in Table 2. For comparison of potential differential item functioning of the negatively and positively worded items, the item threshold values are the most relevant parameters (i.e.,  $b_1$ ,  $b_2$ , and  $b_3$ ) because they represent item difficulty, or the level of the latent trait at which the next highest response category has the highest probability of being selected; thus, higher thresholds imply more difficult items. As is evident from Table 2, the negatively worded items had consistently lower threshold values, including mean threshold values, suggesting that negative worded items were less difficult, or required less of the latent CU trait to endorse with a higher response value compared with positively worded items. In Figure 2, the item characteristic curves (ICCs) for all 22 items are summarized such that the probabilities for item categories are summed across all response categories for each item. This provides a visual comparison of the item difficulties for the negatively and positively worded items. The ICCs for the positively worded items are located at higher levels of the latent CU continuum ( $x$ -axis) compared with the negatively worded items, again providing evidence that the positively worded items were consistently more difficult and were more accurately measuring CU traits at higher levels of the latent CU trait. In short, the positive and negative worded items appeared to be functioning quite differently in terms of item difficulty.

## Developing and Testing a Revised Total Score

Ten items were found to discriminate well (slope values above .90) across the latent CU trait continuum; these are presented in bold in Table 2. The reliability and the validity of a shortened scale (ICU-10) using these 10 items were compared with the reliability and validity of the full 22-item version of the ICU and with scales created from the positively worded items only ( $n = 10$ ) and the negatively worded items only ( $n = 12$ ).

The internal consistencies of these four scales in terms of coefficient alphas and mean corrected ITCs are provided in Table 3. There was minimal variation in alpha coefficient values across the four scales (.74–.78) or in mean corrected ITCs (.33–.41). Estimates of the stability of these scores over the 6 months between assessments are provided in Table 3. Again, all test–retest correlations were strong and positive across the four scales ( $rs = .50$ – $.63$ ,  $ps < .001$ ).

The correlations of all four (i.e., full ICU, ICU-10, positively worded items, and negatively worded items) ways of scoring the ICU with external criteria measured at baseline are provided in Table 4 along with the partial correlations of all four ways of scoring the ICU at baseline with criterion measures at the 6-month follow-up, controlling for baseline measures of external criteria. In terms of convergent validity, the four ways of scoring the ICU were all significantly negatively associated with the consideration of others subscale, a measure of empathy ( $rs$  ranged from  $-.21$  to  $-.41$ ). However, for the baseline associations between empathy and the four ways of scoring the ICU, the correlation with the positively worded items was significantly smaller than the correlations with the full ICU ( $z = 9.28$ ,  $p < .001$ ), the ICU-10 ( $z = 6.69$ ,  $p < .001$ ), and the negatively worded items ( $z = 5.13$ ,  $p < .001$ ).

Table 2  
Slopes and Thresholds for Items on the Inventory of Callous-Unemotional Traits (Excluding Items 2 and 10)

Item	$a$	$b_1$	$b_2$	$b_3$	$M(b)$
<b>16. I apologize (“say I am sorry”) to persons I hurt.<sup>a</sup></b>	1.65	-1.37	0.80	3.53	0.99
<b>17. I try not to hurt others’ feelings.<sup>a</sup></b>	1.58	-1.38	1.01	3.99	1.21
<b>15. I always try my best.<sup>a</sup></b>	1.42	-0.85	1.27	4.55	1.66
<b>3. I care about how well I do at school or work.<sup>a</sup></b>	1.39	-1.27	1.29	4.36	1.46
<b>23. I work hard on everything I do.<sup>a</sup></b>	1.24	-1.24	0.76	3.57	1.03
<b>8. I am concerned about the feelings of others.<sup>a</sup></b>	1.24	-2.13	0.29	2.79	0.32
<b>5. I feel bad or guilty when I do something wrong.<sup>a</sup></b>	1.16	-1.59	0.26	2.57	0.41
<b>24. I do things to make others feel good.<sup>a</sup></b>	1.13	-2.00	0.14	2.77	0.30
<b>11. I do not care about doing things well.</b>	1.08	1.01	0.09	3.42	1.51
<b>7. I do not care about being on time.</b>	1.01	0.45	2.27	3.66	2.13
21. The feelings of others are unimportant to me.	0.89	0.76	2.45	3.58	2.26
4. I do not care who I hurt to get what I want.	0.87	1.10	2.82	3.90	2.61
9. I do not care if I get into trouble.	0.83	0.26	2.18	3.42	1.95
12. I seem very cold and uncaring to others.	0.80	0.71	2.43	3.88	2.34
1. I express my feelings openly. <sup>a</sup>	0.76	-2.17	-0.47	2.11	-0.18
20. I do not like to put the time into doing things well.	0.72	0.24	2.30	4.13	2.22
18. I do not feel remorseful when I do something wrong.	0.66	-0.88	1.72	2.97	1.53
19. I am very expressive and emotional. <sup>a</sup>	0.57	-2.53	-1.18	0.70	-1.00
14. It is easy for others to tell how I am feeling. <sup>a</sup>	0.46	-2.53	-0.97	0.93	-0.86
22. I hide my feelings from others.	0.39	-0.70	1.13	2.10	0.84
13. I easily admit to being wrong. <sup>a</sup>	0.35	-2.16	-0.62	1.24	-1.54
6. I do not show my emotions to others.	0.19	-1.44	0.59	1.73	0.29

Note. Items are listed in order of their discrimination parameter ( $a$ ) values. Bolded items are those included in Inventory of Callous-Unemotional Traits short version.

<sup>a</sup> Reverse-coded (negatively worded) item.

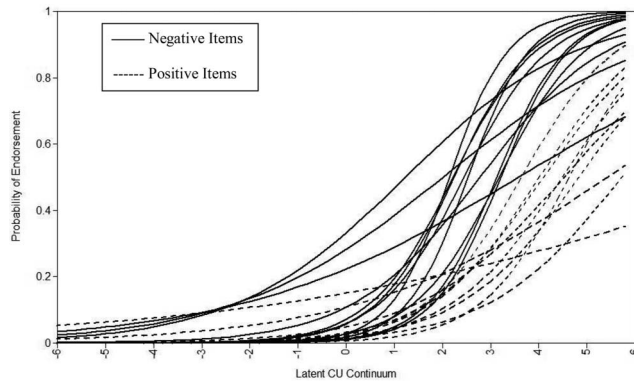


Figure 2. Item information curves for negatively and positively worded items. CU = callous and unemotional.

.001) when testing for significant differences between dependent correlations. A similar pattern was obtained with respect to partial correlations, demonstrating the association of baseline ICU scores with 6-month empathy scores, controlling for baseline levels of empathy.

In terms of criterion validity, all four ways for scoring the ICU were positively correlated with and predictive of self-reported delinquency, school misconduct, and aggression. The one exception was a nonsignificant partial correlation between the positively worded items scale and proactive aggression at 6 months. Further, the two total scores (i.e., full 22-item ICU and ICU-10) showed very similar correlations with all of these indicators of criterion validity. However, there were some differences in correlations with the positively worded and negatively worded items. Specifically, both scales were similarly correlated with school misconduct at baseline, but the correlations for positively worded items with delinquency ( $z = 3.35, p < .001$ ), proactive aggression ( $z = 3.85, p < .001$ ), and reactive aggression ( $z = 4.45, p < .001$ ) were significantly greater than those for negatively worded items<sup>1</sup> on the basis of tests for significant differences in correlations.

In addition, the different ways of scoring the ICU were positively correlated with and predictive of anxiety, with the exception of the negative items. However, these correlations exhibited the biggest difference between positive and negatively worded items. That is, the scale using only negatively worded items was unrelated to anxiety at baseline ( $r = -.02, p = .43$ ), whereas the scale using only the positive items showed the highest correlation with anxiety ( $r = .25, p < .001$ ), and this difference was significant ( $z = 7.29, p < .001$ ). Importantly, these correlations with anxiety at baseline were repeated after controlling for delinquency to determine if the suppressor effect found in past studies was also evident in the current sample (Frick, Cornell, Bodin, Dane, Barry, & Loney, 2003; Frick, Lilienfeld, Ellis, Loney, & Silverthorn, 1999). That is, past studies have found that measures of CU traits are sometimes positively correlated with anxiety because of their shared association with antisocial behavior. However, this correlation was eliminated once the effects of antisocial behavior were controlled. Controlling for delinquency at baseline, the correlations between the two total ICU scores with anxiety became nonsignificant ( $r = .05, p = .09$ , and  $r = -.01, p = .84$ , for the full ICU and ICU-10, respectively), similar to past studies. Further,

the positive correlation between the ICU based solely on the positively worded items and anxiety ( $r = .19, p < .001$ ) remained, whereas the correlation between the negative items and anxiety became significant in the opposite direction ( $r = -.08, p < .01$ ).

## Discussion

An important goal of the current analyses of the ICU was to test possible explanations for why positively worded and negatively worded items often form separate subfactors in bifactor models using confirmatory factor analyses (Ciucci et al., 2014; Essau et al., 2006; Ezpeleta et al., 2013; Fanti et al., 2009; Kimonis et al., 2008; Roose et al., 2010; Waller et al., 2014). Several findings suggest that this may be a result of differences in item response patterns and corresponding differences in item difficulty (i.e., how well an item assesses the construct of CU traits at different levels of severity). Specifically, positively worded items are less likely to elicit responses indicating the highest level of CU traits than are negatively worded items (see Figure 1). Further, IRT analyses indicated that positively worded items displayed their best discrimination at much higher levels of CU traits (see Figure 2) than did negatively worded items. In short, it seems that it takes a higher level of CU traits to endorse the top ratings for positively worded items than to endorse the lowest ratings for negatively worded items.

This method variance could also account for the somewhat different patterns of correlations with external criteria found for scales formed using either positively or negatively worded items alone. That is, for measures indicative of less severe antisocial behavior (i.e., school misconduct) or prosocial behavior, the negatively worded items correlated the same as or better than the positively worded items. In contrast, for measures of more severe antisocial behavior that are likely to be largely present at higher levels of CU traits, the positively worded items exhibited stronger correlations. Further, given that anxiety is typically a marker of the severity of conduct problems (Frick, 2012), the fact that positively worded items are highly correlated with anxiety (unless conduct problem severity is controlled) is also consistent with this interpretation. In summary, the present results support the contention that prior bifactor analyses may reflect differences in item severity. Also, it appears that negative items that are worded to reflect prosocial features may be more helpful in assessing less severe manifestations of the construct, whereas positive items are more helpful in assessing more severe manifestations of CU traits.

These results also suggest that the total score of the ICU provides a reliable and valid continuous measure of CU traits, similar to the way they have been operationalized in the *DSM-5* specifier for conduct disorder called “with limited prosocial emotions” (Essau et al., 2006; Kimonis et al., 2008; Waller et al., 2014).

<sup>1</sup> To examine the unique associations with different types of aggression, we conducted partial correlations with each ICU version and each type of aggression (i.e., proactive [controlling for reactive] and reactive [controlling for proactive]) at baseline only. For proactive aggression, the ICU total ( $r = .22, p < .001$ ), ICU-10 ( $r = .18, p < .001$ ), negatively worded items scale ( $r = .12, p < .001$ ), and positively worded items scale ( $r = .24, p < .001$ ) were all positively correlated, controlling for reactive aggression. For reactive aggression, however, only the scale comprising the positively worded items ( $r = .12, p < .001$ ) showed a significant correlation, controlling for proactive aggression.



Table 3  
*Internal Consistency and Stability Estimates for Each of the Four Versions of the Inventory of Callous-Unemotional Traits (ICU)*

Variable	Full ICU	ICU-10	Negative items	Positive items
Coefficient alpha	.78	.78	.79	.74
Baseline mean corrected item-total correlation	.33	.45	.43	.41
6-month test-retest correlation <sup>a</sup>	.63	.59	.57	.50

Note. Items 2 and 10 were not included in analyses.

<sup>a</sup> All were significant at  $p < .001$ .

However, as in past studies (S. W. Hawes et al., 2014; Kimonis et al., 2008), two items (i.e., “What I think is ‘right’ and ‘wrong’ is different from what other people think” and “I do not let my feelings control me”) did not correlate highly with the total score. Given the poor performance of these items across samples and versions of the ICU, future users of the ICU should consider eliminating them when calculating total scores. Further, on the basis of our IRT analyses, a shortened version of the ICU using the 10 best discriminating items exhibited very similar levels of reliability and validity to the full version. This finding is consistent with past studies examining shortened versions of the ICU (S. W. Hawes et al., 2014; Waller et al., 2014), and it suggests that abbreviated versions of the ICU can be used without substantial reductions in reliability and validity. Interestingly, there was only modest overlap between the items that performed best in this adolescent sample (ages 13–17 years) using the self-report of the ICU and the items that performed best using parent report in the S. W. Hawes et al. (2014) sample of boys ages 6–12. It is not clear if these differences in item performance are attributable to the differences in age of the samples or the format of the ICU used. However, it is possible that the utility of the shortened version may not be a result of the relative importance of certain items but, rather, reflects that the items all measure a similar construct and that any substantial subset of items could adequately capture the construct. This possibility is supported by analyses in the current sample using the shortened version developed by S. W. Hawes et al. (2014).<sup>2</sup> This shortened version of the ICU, which was developed from IRT analyses in a separate sample, demonstrated similar levels of reliability and validity to the shortened scale based on IRT analyses from this sample, despite the modest overlap in items.

Although our findings support the use of the total score (excluding Items 2 and 10), the current findings call into question the theoretical importance of the subscales of the ICU that have emerged in prior factor analyses. That is, the current results suggest that the bifactor structure identifying three subfactors (i.e., callousness, uncaring, and unemotional) may be an artifact of different item-endorsement rates and item difficulties rather than reflecting theoretically important dimensions of the construct of CU traits. Specifically, the uncaring and callousness factors identified in previous research (e.g., Kimonis et al., 2008) overlap considerably with the negative and positively worded scales identified in the current study, respectively. Specifically, all eight of the ICU items on the uncaring factor are negatively worded items,

whereas eight of the nine items on the callousness factor are positively worded. Also supporting the limited theoretical utility of the different subscales is the failure of past research to show consistent differences in external correlates to the subscales that conform to a clear theoretical model for a multidimensional construct of CU traits (Frick & Ray, 2014; Waller et al., 2014). Thus, until other data emerge to provide more substantive meeting to these subfactors, use of a total score seems to be most defensible in providing a continuous measure of CU traits for research and practice.

There are some important limitations in the use of the ICU total score that need to be addressed in future research. First, although research supports the use of the ICU total score as a continuous measure of CU traits, there may be some situations in which decisions need to be made as to whether the levels of these traits are elevated. For example, the use of the specifier “with limited prosocial emotions” requires a decision as to whether significant levels of these traits are present or absent in people who meet criteria for conduct disorder. To date, there is no widely used and extensively validated cut score to make such decisions with the ICU, with much of the past published research using sample-dependent cutoffs when such dichotomous decisions are called for (e.g., Lawing, Frick, & Cruise, 2010). Based on IRT analysis, our findings suggest that the ICU comprises items that capture CU traits across a broad range of the latent trait continuum and, thus, is best used as a continuous measure. However, future research should focus on whether there are cut scores that can aid in determining when scores on the ICU are well outside a normative level or when they optimally predict important outcomes (e.g., risk for offending, poor response to treatment).

Second, there are items from the ICU that can be used to aid people in determining if the symptoms of the specifier are present, which has been done in several prior studies (Kimonis et al., 2014; McMahon, Witkiewitz, & Kotler, 2010; Pardini, Stepp, Hipwell, Stouthamer-Loeber, & Loeber, 2012). However, in many settings, symptom presence or absence is a clinical decision that requires consideration of a range of sources and types of information that do not rely solely on ratings on a questionnaire. The clinician-rated Psychopathy Checklist: Youth Version (PCL-YV; Forth et al., 2003) includes four items in its affective dimension that correspond closely to the symptoms of the specifier that can be used for this purpose. However, the PCL-YV also assesses a number of other items (21 in total) that are not directly related to the specifier, making it a very time-consuming method for clinical assessment. A clinician-rated measure of the specifier symptoms is currently

<sup>2</sup> Analyses were also run comparing the reliability and validity of the 12-item version of the ICU (ICU-12) developed by S. W. Hawes et al. (2014) on the basis of their IRT analyses of the parent report on the ICU. Six items (i.e., 5, 8, 11, 16, 17, and 24) on the ICU-12 overlapped with the 10 items selected from the IRT in the current sample (ICU-10). The internal consistency ( $\alpha = .72$ ) and the stability ( $r = .61, p < .001$ ) of the ICU-12 were very similar to those reported in Table 3 for the full ICU and ICU-10. Further, the correlations (concurrent and predictive, respectively) with the consideration of others subscale ( $r_s = -.41$  and  $-.19, p_s < .001$ ), school misconduct ( $r_s = .30$  and  $.22, p_s < .001$ ), delinquency ( $r_s = .32$  and  $.17, p_s < .001$ ), proactive aggression ( $r_s = .33$  and  $.11, p_s < .001$ ), reactive aggression ( $r_s = .34$  and  $.15, p_s < .001$ ), and anxiety ( $r = .12, p < .001$ , and  $r = .09, p < .01$ ) were also very similar to those reported for the full ICU and ICU-10.



Table 4

Correlations for Each of the Four Versions of the Inventory of Callous-Unemotional Traits (ICU) at Baseline, With External Criteria at 6-Month Follow-Up

Measure	Full ICU		ICU-10		Negative items		Positive items	
	<i>r</i>	Partial <i>r</i>	<i>r</i>	Partial <i>r</i>	<i>r</i>	Partial <i>r</i>	<i>r</i>	Partial <i>r</i>
Consideration of others ( <i>n</i> = 1,164)	-.41**a	-.19**	-.40**a	-.17**	-.39**a	-.17**	-.21**b	-.09*
School misconduct ( <i>n</i> = 1,048)	.33**a	.24**	.32**a	.24**	.25**b	.19**	.26**b	.17**
Delinquency ( <i>n</i> = 1,164)	.30**a	.17**	.30**a	.15**	.20**b	.10**	.32**a	.17**
Proactive aggression ( <i>n</i> = 1,164)	.27**a	.08*	.25**a	.09*	.15**b	.08*	.29**a	.04
Reactive aggression ( <i>n</i> = 1,164)	.31**a	.17**	.27**b	.16**	.17**c	.14**	.33**a	.11**
Anxiety ( <i>n</i> = 1,164)	.13**d	.10*	.08**c	.08*	-.02 <sup>d</sup>	.04	.25**a	.13**

Note. *N* = 1,190. Correlations reflect associations at baseline. Partial correlations reflect associations between the ICU at baseline and criterion variables at 6 months, controlling for baseline levels of criterion variables. Samples sizes are different because of missing values on criterion measures. Correlations with different superscripts are significantly different from one another (across rows) on the basis of tests for equality of dependent correlations. All analyses excluded items 2 and 10.

\* *p* < .01. \*\* *p* < .001.

under development (Frick, 2013), but the reliability and validity of this method of assessment has not been tested. In short, more work is needed to develop time-efficient, reliable, and valid measures of CU traits that can be used in a variety of clinical settings.

Third, the conceptualization of the factor structure of the ICU as being largely determined by item wording may explain the identification of the callous (i.e., positively worded items) and uncaring (i.e., negatively worded items) dimensions, but it does not provide a strong explanation for the consistent emergence of the unemotional factor. Past research suggests that the unemotional subscale of the ICU contributes to an overall CU factor in bifactor models but is weakly correlated with the other subscales and shows more divergent associations with some external criteria, such as being more weakly correlated with conduct problems and aggression (S. W. Hawes et al., 2014; Houghton et al., 2013; Kimonis et al., 2008; Waller et al., 2014). However, this scale does seem to be correlated (negatively) with measures of empathy and other aspects of prosocial behavior (Ciucci et al., 2014; Kimonis et al., 2008; Roose et al., 2010). Thus, it appears to assess a distinct aspect of CU traits that is more specific to prosocial emotions independent of aggression, and, as a result, its emergence in factor analysis does not appear to be largely a function of item wording<sup>3</sup> Future research needs to examine how this dimension is related to and differs from the rest of the ICU content to determine the best way to conceptualize it within the broader construct of CU traits (Latzman, Lilienfeld, Latzman, & Clark, 2013; Waller et al., 2014).

Although this study provides more clarity on the ICU measure, it has several limitations that need to be considered when drawing conclusions about the ICU as a measure of CU traits. The current sample consisted of first-time male juvenile offenders. Therefore, it is unclear whether the current findings generalize to the more general population of youths, including nonjustice-involved youths and girls, although the findings are fairly consistent with IRT analyses using the parent report of the ICU in a sample of boys ages 6–12. Further, all measures included in the current study were based on self-report, and, thus, associations may be inflated because of shared-method variance. As a result, future research should continue to test the ICU in relation to other criteria that are measured using other methods of assessment (see Kimonis et al., 2008).

In conclusion, it appears that the total score of the ICU (excluding Items 2 and 10) provides a psychometrically sound continuous measure of CU traits as they are conceptualized in the *DSM-5* specifier “with limited prosocial emotions.” Because inclusion of CU traits in diagnostic classification is also being considered for the upcoming revision of the *International Classification of Diseases* by the World Health Organization (Rutter, 2012), it is critically important to have sound measures of this construct for use in both research and practice. Our findings suggest that the design of the ICU to include equal numbers of positively and negatively worded items may have contributed to the factor structure that emerged in past studies. In fact, there is little evidence to date to support a conceptualization of CU traits as being composed of multiple theoretically important dimensions, although more research is needed to explore the specific role of the unemotional items in the overall construct of CU traits. Thus, it appears that CU traits are best conceptualized as a unidimensional construct and that the total ICU score should be used as a continuous measure of CU traits in which both positive and negative worded items are included, albeit eliminating Items 2 and 10 and recognizing that shortened item lists can be used with minimum reductions in reliability and validity.

<sup>3</sup> Correlations for the unemotional subscale in the current sample with consideration of others (*r* = -.17, *p* < .001), self-reported delinquency (*r* = .15, *p* < .001), and reactive (*r* = .08, *p* < .01) and proactive aggression (*r* = .08, *p* < .01) were consistent with prior research.

## References

- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.
- Chorpita, B. F., Moffitt, C. E., & Gray, J. (2005). Psychometric properties of the Revised Child Anxiety and Depression Scale in a clinical sample. *Behaviour Research and Therapy*, 43, 309–322. <http://dx.doi.org/10.1016/j.brat.2004.02.004>
- Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of *DSM-IV* anxiety and depression in

- children: A revised child anxiety and depression scale. *Behaviour Research and Therapy*, 38, 835–855. [http://dx.doi.org/10.1016/S0005-7967\(99\)00130-8](http://dx.doi.org/10.1016/S0005-7967(99)00130-8)
- 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, 189–200. <http://dx.doi.org/10.1007/s10862-013-9384-z>
- de Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford Press.
- Denver Youth Survey. Huizinga, D., Esbensen, F.-A., & Weiher, A. W. (1991). Are there multiple paths to delinquency? *Journal of Criminal Law and Criminology*, 82, 83–118. <http://dx.doi.org/10.2307/1143790>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Espelage, D. L., Mebane, S. E., & Adams, R. S. (2004). Empathy, caring, and bullying: Toward an understanding of complex associations. In D. L. Espelage & S. M. Swearer (Eds.), *Bullying in American schools: A social-ecological perspective on prevention and intervention* (pp. 37–61). Mahwah, NJ: Erlbaum.
- Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, 13, 454–469. <http://dx.doi.org/10.1177/1073191106287354>
- Ezpeleta, L., de la Osa, N., Granero, R., Penelo, E., & Domènech, J. M. (2013). Inventory of callous-unemotional traits in a community sample of preschoolers. *Journal of Clinical Child and Adolescent Psychology*, 42, 91–105. <http://dx.doi.org/10.1080/15374416.2012.734221>
- Fanti, K. A., Frick, P. J., & Georgiou, S. (2009). Linking callous-unemotional traits to instrumental and non-instrumental forms of aggression. *Journal of Psychopathology and Behavioral Assessment*, 31, 285–298. <http://dx.doi.org/10.1007/s10862-008-9111-3>
- Farrell, A. D., & Sullivan, T. N. (2000). Structure of the Weinberger Adjustment Inventory Self-Restraint scale and its relation to problem behaviors in adolescence. *Psychological Assessment*, 12, 394–401. <http://dx.doi.org/10.1037/1040-3590.12.4.394>
- Feilhauer, J., Cima, M., & Arntz, A. (2012). Assessing callous-unemotional traits across different groups of youths: Further cross-cultural validation of the Inventory of Callous-Unemotional Traits. *International Journal of Law and Psychiatry*, 35, 251–262. <http://dx.doi.org/10.1016/j.ijlp.2012.04.002>
- Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *Hare Psychopathy Checklist: Youth Version (PCL: YV)*. North Tonawanda, NY: Multi-Health Systems.
- Frick, P. J. (2009). Extending the construct of psychopathy to youth: Implications for understanding, diagnosing, and treating antisocial children and adolescents. *Canadian Journal of Psychiatry/Revue Canadienne de psychiatrie*, 31, 803–812.
- Frick, P. J. (2012). Developmental pathways to conduct disorder: Implications for future directions in research, assessment, and treatment. *Journal of Clinical Child and Adolescent Psychology*, 41, 378–389. <http://dx.doi.org/10.1080/15374416.2012.664815>
- Frick, P. J. (2013, July 1). *Clinical Assessment of Prosocial Emotions: Version 1.1 (CAPE 1.1)*. Retrieved from [http://labs.uno.edu/developmental-psychopathology/documents/CAPE%20Manual\[1\].pdf](http://labs.uno.edu/developmental-psychopathology/documents/CAPE%20Manual[1].pdf)
- Frick, P. J., Bodin, S. D., & Barry, C. T. (2000). Psychopathic traits and conduct problems in community and clinic-referred samples of children: Further development of the psychopathy screening device. *Psychological Assessment*, 12, 382–393. <http://dx.doi.org/10.1037/1040-3590.12.4.382>
- Frick, P. J., Cornell, A. H., Bodin, S. D., Dane, H. E., Barry, C. T., & Loney, B. R. (2003). Callous-unemotional traits and developmental pathways to severe conduct problems. *Developmental Psychology*, 39, 246–260. <http://dx.doi.org/10.1037/0012-1649.39.2.246>
- Frick, P. J., & Hare, R. D. (2001). *Antisocial Process Screening Device: APSD*. Toronto, Ontario, Canada: Multi-Health Systems.
- Frick, P. J., Lilienfeld, S. O., Ellis, M., Loney, B., & Silverthorn, P. (1999). The association between anxiety and psychopathy dimensions in children. *Journal of Abnormal Child Psychology*, 27, 383–392. <http://dx.doi.org/10.1023/A:1021928018403>
- Frick, P. J., & Ray, J. V. (2014). Evaluating callous-unemotional traits as a personality construct. *Journal of Personality*. Advance online publication. <http://dx.doi.org/10.1111/jopy.12114>
- Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Annual research review: A developmental psychopathology approach to understanding callous-unemotional traits in children and adolescents with serious conduct problems. *Journal of Child Psychology and Psychiatry*, 55, 532–548. <http://dx.doi.org/10.1111/jcpp.12152>
- Hawes, D. J., & Dadds, M. R. (2005). The treatment of conduct problems in children with callous-unemotional traits. *Journal of Consulting and Clinical Psychology*, 73, 737–741. <http://dx.doi.org/10.1037/0022-006X.73.4.737>
- Hawes, S. W., Byrd, A. L., Henderson, C. E., Gazda, R. L., Burke, J. D., Loeber, R., & Pardini, D. A. (2014). Refining the parent-reported Inventory of Callous-Unemotional Traits in boys with conduct problems. *Psychological Assessment*, 26, 256–266. <http://dx.doi.org/10.1037/a0034718>
- Houghton, S., Hunter, S. C., & Crow, J. (2013). Assessing callous unemotional traits in children aged 7- to 12-years: A confirmatory factor analysis of the Inventory of Callous Unemotional Traits. *Journal of Psychopathology and Behavioral Assessment*, 35, 215–222. <http://dx.doi.org/10.1007/s10862-012-9324-3>
- Kimonis, E. R., Fanti, K. A., Frick, P. J., Moffitt, T. E., Essau, C., Bijttebier, P., & Marsee, M. A. (2014). Using self-reported callous-unemotional traits to cross-nationally assess the DSM-5 “with limited prosocial emotions” specifier. *Journal of Child Psychology and Psychiatry*. Advance online publication. <http://dx.doi.org/10.1111/jcpp.12357>
- 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, 241–252. <http://dx.doi.org/10.1016/j.ijlp.2008.04.002>
- Latzman, R. D., Lilienfeld, S. O., Latzman, N. E., & Clark, L. A. (2013). Exploring callous and unemotional traits in youth via general personality traits: An eye toward DSM-5. *Personality Disorders: Theory, Research, and Treatment*, 4, 191–202. <http://dx.doi.org/10.1037/a0000001>
- 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, 298–305. <http://dx.doi.org/10.1037/a0018707>
- 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, 969–981. <http://dx.doi.org/10.1007/s10802-007-9147-y>
- Marsee, M. A., Frick, P. J., Barry, C. T., Kimonis, E. R., Muñoz Centifanti, L. C., & Aucoin, K. J. (2014). Profiles of the forms and functions of self-reported aggression in three adolescent samples. *Development and Psychopathology*, 26, 705–720. <http://dx.doi.org/10.1017/S0954579414000339>
- McMahon, R. J., Witkiewitz, K., & Kotler, J. S. (2010). Predictive validity of callous-unemotional traits measured in early adolescence with respect to multiple antisocial outcomes. *Journal of Abnormal Psychology*, 119, 752–763. <http://dx.doi.org/10.1037/a0020796>
- Muñoz, L. C., Frick, P. J., Kimonis, E. R., & Aucoin, K. J. (2008). Types of aggression, responsiveness to provocation, and callous-unemotional traits in detained adolescents. *Journal of Abnormal Child Psychology*, 36, 15–28. <http://dx.doi.org/10.1007/s10802-007-9137-0>

- Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles: Author.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.
- Pardini, D. A., & Byrd, A. L. (2012). Perceptions of aggressive conflicts and others' distress in children with callous-unemotional traits: "I'll show you who's boss, even if you suffer and I get in trouble." *Journal of Child Psychology and Psychiatry*, *53*, 283–291. <http://dx.doi.org/10.1111/j.1469-7610.2011.02487.x>
- Pardini, D. A., & Fite, P. J. (2010). Symptoms of conduct disorder, oppositional defiant disorder, attention-deficit/hyperactivity disorder, and callous-unemotional traits as unique predictors of psychosocial maladjustment in boys: Advancing an evidence base for *DSM-V*. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 1134–1144.
- Pardini, D., Stepp, S., Hipwell, A., Stouthamer-Loeber, M., & Loeber, R. (2012). The clinical utility of the proposed *DSM-5* callous-unemotional subtype of conduct disorder in young girls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*, 62–73.e4. <http://dx.doi.org/10.1016/j.jaac.2011.10.005>
- Poythress, N. G., Dembo, R., Wareham, J., & Greenbaum, P. E. (2006). Construct validity of the Youth Psychopathic Traits Inventory (YPI) and the Antisocial Process Screening Device (APSD) with justice-involved adolescents. *Criminal Justice and Behavior*, *33*, 26–55. <http://dx.doi.org/10.1177/0093854805282518>
- Quay, H. C. (1993). The psychobiology of undersocialized aggressive conduct disorder: A theoretical perspective. *Development and Psychopathology*, *5*(1–2), 165–180. <http://dx.doi.org/10.1017/S0954579400004326>
- 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*, 44–57. <http://dx.doi.org/10.1177/1073191109344153>
- Rutter, M. (2012). Psychopathy in childhood: Is it a meaningful diagnosis? *British Journal of Psychiatry*, *200*, 175–176. <http://dx.doi.org/10.1192/bjp.bp.111.092072>
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer. [http://dx.doi.org/10.1007/978-1-4757-2691-6\\_5](http://dx.doi.org/10.1007/978-1-4757-2691-6_5)
- Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. *Criminal Justice*, *4*, 33–83.
- Waller, R., Wright, A. G., Shaw, D. S., Gardner, F., Dishion, T. J., Wilson, M. N., & Hyde, L. W. (2014). Factor structure and construct validity of the parent-reported Inventory of Callous-Unemotional Traits among high-risk 9-year-olds. *Assessment*. Advance online publication. <http://dx.doi.org/10.1177/1073191114556101>
- Weinberger, D. A., & Schwartz, G. E. (1990). Distress and restraint as superordinate dimensions of self-reported adjustment: A typological perspective. *Journal of Personality*, *58*, 381–417. <http://dx.doi.org/10.1111/j.1467-6494.1990.tb00235.x>

Received December 22, 2014

Revision received April 9, 2015

Accepted May 14, 2015 ■

### Correction to Ray et al. (2015)

In the article "Positive and Negative Item Wording and Its Influence on the Assessment of Callous-Unemotional Traits" by James V. Ray, Paul J. Frick, Laura C. Thornton, Laurence Steinberg, and Elizabeth Cauffman (*Psychological Assessment*, Advanced online publication, June 29, 2015. <http://dx.doi.org/10.1037/pas0000183>), the sixth sentence of the second full paragraph in the **Data Analyses** subsection of the **Method** section should read "For  $k$  response categories, there are  $k-1$  threshold parameters."

<http://dx.doi.org/10.1037/pas0000216>