THE JOURNAL OF CHILD PSYCHOLOGY AND PSYCHIATRY

ACAMH THE ASSOCIATION FOR CHILD AND ADOLESCENT MENTAL HEALTH

Journal of Child Psychology and Psychiatry 56:11 (2015), pp 1249–1261 doi:10.1111/jcpp.12357

Using self-reported callous-unemotional traits to cross-nationally assess the DSM-5 'With Limited Prosocial Emotions' specifier

Eva R. Kimonis,[1] Kostas A. Fanti,[2] Paul J. Frick,[3,4] Terrie E. Moffitt,[5,6] Cecilia Essau,[7] Patricia Bijttebier,[8] and Monica A. Marsee[3]

[1]The School of Psychology, The University of New South Wales, Sydney, NSW, Australia; [2]Department of Psychology, University of Cyprus, Nicosia, Cyprus; [3]Department of Psychology, University of New Orleans, New Orleans, LA, USA; [4]Learning Sciences Institute of Australia, Australian Catholic University, Fitzroy, Vic., Australia; [5]Duke Psychology and Neuroscience Psychiatry & Behavioral Sciences, Duke Institute for Genome Sciences and Policy, Durham, NC, USA; [6]Institute of Psychiatry, King's College London, London; [7]School of Human and Life Sciences, Roehampton University, London, UK; [8]School Psychology and Child and Adolescent Development, KU Leuven, Belgium

Background: The presence of callous-unemotional (CU) traits designates an important subgroup of antisocial youth at risk for severe, persistent, and impairing conduct problems. As a result, the fifth revision of the *Diagnostic and Statistical Manual* includes a specifier for youth meeting diagnostic criteria for Conduct Disorder who show elevated CU traits. The current study evaluated the DSM-5 criteria using Item Response Theory (IRT) analyses and evaluated two methods for using a self-report measure of CU traits to make this diagnosis. **Methods:** The sample included 2257 adolescent (*M* age = 15.64, *SD* = 1.69 years) boys (53%) and girls (47%) from community and incarcerated settings in the United States and the European countries of Belgium, Germany, and Cyprus. **Results:** IRT analyses suggested that four- or eight-item sets from the self-report measure (comparable to the symptoms used by the DSM-5 specifier) provided good model fit, suggesting that they assess a single underlying CU construct. Further, the most stringent method of scoring the self-report scale (i.e. taking only the most extreme responses) to approximate symptom presence provided the best discrimination in IRT analyses, showed reasonable prevalence rates of the specifier, and designated community adolescents who were highly antisocial, whereas the less stringent method best discriminated detained youth. **Conclusions:** Refined self-report scales developed on the basis of IRT findings provided good assessments of most of the symptoms used in the DSM-5 criteria. These scales may be used as one component of a multimethod assessment of the 'With Limited Prosocial Emotions' specifier for Conduct Disorder.
Keywords: Callous-unemotional traits, DSM-5, conduct disorder, with limited prosocial emotions, item response theory analysis.

Introduction

Youth exhibiting significant and impairing patterns of antisocial and aggressive behavior that meet diagnostic criteria for Conduct Disorder (CD) are heterogeneous with respect to their severity, prognosis, and presumed etiology (Frick & Viding, 2009). Further, a substantial body of research suggests that nonnormative levels of callous-unemotional (CU) traits are useful for identifying antisocial youth who show a distinct pattern of severe, chronic, and aggressive conduct problems that are resistant to traditional mental health interventions, and underpinned by distinct causal factors. The fifth edition of the *Diagnostic and Statistical Manual* (DSM-5) integrates CU traits into its diagnostic criteria for CD by including the specifier 'With Limited Prosocial Emotions' [American Psychiatric Association (APA) 2013]. The specifier is used when a person meeting diagnostic criteria for CD persistently (≥12 months) exhibits two or more of the following characteristics in multiple relationships or settings: (a) lack of remorse or guilt; (b) callous-lack of empathy; (c) unconcerned about performance; or (d) shallow or deficient affect. These four criteria closely approximate the Affective Dimension of psychopathy that has been considered core to defining this construct in adult samples (Hare & Neumann, 2008), and also showed the most consistent loadings on a CU factor using parent and teacher ratings across community and clinical samples of children (Frick et al., 2000).

The addition of this CU specifier is expected to provide greater information about etiology, current and future impairment, and to aid in treatment planning for youth diagnosed with CD (Frick, Ray, Thornton, & Kahn, 2014), of whom an estimated 12–50% present with significant CU traits (Kahn, Frick, Youngstrom, Findling, & Youngstrom, 2012; Rowe et al., 2010). For example, nonnormative CU traits differentiate groups of children and adolescents with CD showing more premeditated and instrumental (i.e. for gain) aggression (Frick & White 2008). They also designate those showing distinct brain activity patterns in response to emotional stimuli (Viding et al., 2012). Similarly, CU traits at school age predict criminal and antisocial behavior in adulthood, even after controlling for severity and onset of CD (McMahon, Witkiewitz, Kotler, & the Conduct Problems Prevention Research Group 2010).

Conflict of interest statement: No conflicts declared.

Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA 02148, USA

The inclusion of the specifier in the DSM-5 necessitates further research to test and refine the optimal indicators of the construct of CU traits. As this method for operationalizing nonnormative CU traits in the DSM-5 is relatively new and untested, it is crucial to rigorously evaluate it so that modifications can be made in subsequent revisions of the manual. It is also important to evaluate commonly used methods for assessing CU traits to determine which best capture the overall construct and best identify youth with serious antisocial behavior. In the current study, we begin to advance these two goals using the self-report version of the *Inventory of Callous-Unemotional Traits* (ICU; Kimonis et al., 2008) in a large multinational adolescent sample. The ICU was chosen because it was systematically developed over two decades, providing one of the most comprehensive measures of CU traits currently available. The 24-item ICU shows acceptable internal consistency and correlates with important outcomes, such as reduced emotional responding to distress cues and severe aggression, across a wide age range, gender, types of samples, and different language translations (e.g. Ezpeleta, Osa, Granero, Penelo, & Domènech, 2013; Fanti, Frick, & Georgiou, 2009).

The second reason for testing the ICU is that its items were used in the secondary data analyses that guided the DSM-5 specifier's formation (Frick & Moffitt, 2010). Specifically, the four DSM-5 criteria symptoms were based on the original four items that guided the item pool for developing the ICU, although the DSM criteria provide a more extended description of each symptom in comparison. When developing the DSM-5 specifier symptoms, a longer nine-symptom list was also considered based on confirmatory factor analyses of the ICU in four samples from different countries with different language translations (Essau, Sasagawa, & Frick, 2006; Fanti et al., 2009; Kimonis et al., 2008; Roose, Bijttebier, Decoene, Claes, & Frick, 2010). Items loading >0.40 on the overarching CU factor, and/or being one of the two highest loading items on a subfactor in two or more samples, were selected for the extended list. Three items overlapped with the four-item criteria set. Because associations with several external criteria revealed similar effect sizes for the four- and nine-item sets across samples, the shorter and more parsimonious set was eventually chosen to form the basis for the DSM-5 specifier criteria (Frick & Moffitt, 2010).

The current study sought to extend these tests in the same four cross-national samples used by Frick and Moffitt (2010) in several important ways. First, the study sought to provide a more rigorous comparison of the two-item sets (i.e. four- and nine-item sets) using a latent-variable statistical approach, item response theory (IRT) analysis. IRT assesses how well the item set as a whole and each individual item comprising the sets measure the overarching latent construct of CU traits across its continuum. It provides a method for evaluating the adequacy of the DSM-5 operationalization, especially in comparison to a viable alternative with more items. Second, the study compares two ways for translating the ordinal ICU item rating format into dichotomous decisions required by the DSM-5 specifier. Specifically, it evaluates more and less stringent methods for dichotomizing items into symptoms that are either present or absent by testing the utility of each for designating youth with severe antisocial and aggressive behavior.

Methods

Participants

Participants were 2,257 community and detained adolescents from five studies conducted in Belgium (Dutch-speaking part), Germany, United States, and Cyprus. Belgian participants were 455 14–20-year-old (M = 16.67, SD = 1.34) adolescents (44% girls) recruited from six high schools in rural and urban areas of Flanders (Roose et al., 2010). German participants were 1315 13–18-year-old (M = 15.59, SD = 1.56) adolescents (46% girls) recruited from three urban and three rural schools in Nordrhein Westfalia (Essau et al., 2006). American participants were 158 12–18-year-old (M = 15.29, SD = 1.30) youth (38% girls) housed in detention facilities in the Southeast (Kimonis et al., 2008; Marsee & Frick, 2007). Cypriot participants were 329 12–18-year-old (M = 14.63, SD = 2.05) youth (52% girls) recruited from middle (49.3%) and high (50.7%) schools (Fanti et al., 2009).

Procedures

School approval and parental written informed consent were obtained and the majority of European youth agreed to participate (94% Belgian; 92% German; 95% Cypriot). Community-based youth independently completed questionnaires in their classroom during regular school hours. Instruments were adapted and translated for non-English-speaking samples according to widely accepted guidelines for cross-cultural research (Brislin, 1970). For incarcerated participants, detention center directors approved the study and parents/legal guardians of residents provided informed consent via telephone recording. The youth who assented to participate (81% boys, 73% girls) completed questionnaires in small groups (3–8 participants) with questions read aloud to control for reading level.

Measures

Callous-unemotional traits. The 24-item ICU (Frick, 2004) was administered to all youth to assess CU traits. Items are rated on a 4-point Likert scale from 0 (Not at all true) to 3 (Definitely true). Alphas for total ICU scores were acceptable to good, ranging from 0.77 to 0.89 across samples (Essau et al., 2006; Roose et al., 2010).

Aggression/antisocial behavior. The antisocial behavior subscale of the Social and Health Assessment (SAHA; Schwab-Stone, Chen, Greenberger, Silver Lichtman, & Voyce, 1999) assessed antisocial behavior in the Belgian and German samples. Youth report on the frequency of engaging in a variety of antisocial acts (e.g. vandalism, weapon possession, theft, assault) during the past year using a 5-point scale (0, 1, 2, 3–4,

or 5 or more times) with items summed to form a total score (Belgium α = .86, German α = .84).

Externalizing/conduct problems. The Youth Self-Report (YSR, Achenbach, 1991) externalizing composite (aggressive, delinquent behavior subscales) assessed externalizing problems in the German sample. Externalizing scores have demonstrated good internal consistency and validity (e.g. Achenbach, 1991), with α = .86 in the present study. The Bremen Psychopathology Scale (Essau, 2000) assessed CD symptoms in the German sample. Participants rate symptoms on a 4-point scale (α = .77, current study), ranging from 0 (never) to 3 (very often).

Delinquency. The Self-Reported Delinquency Scale (SRD; Elliot & Ageton, 1980) assessed delinquency in the American sample. Youth indicate whether or not they have ever engaged in 36 illegal juvenile acts and endorsements are summed to create total, property (10-item) and violent (eight-item) delinquency scales (Krueger et al., 1994). Cronbach's alphas ranged from .61 (violent) to .88 (total).

Proactive aggression. Youth rated the 10 items of the proactive overt aggression scale from the Peer Conflict Scale (PCS; Marsee & Frick, 2007) on a 4-point scale from 0 ('Not at all true') to 3 ('Definitely true') to assess aggression in the American sample (α = .77). Youth rated the 12 items of the proactive aggression scale from the Reactive-Proactive Aggression Questionnaire (Raine et al., 2006) from 0 ('Never') to 2 ('Often') to assess aggression in the Cypriot sample (α = .81).

Bullying. The Student Survey of Bullying Behavior-Revised (SSBBR; Varjas, Meyers, & Hunt, 2006) assessed bullying in the Cypriot sample. Participants rated the frequency of engaging in physical, verbal, or relational bullying on an ordinal scale of: never, once or twice a year, monthly, weekly, or daily (α = .88).

Plan of analysis

To determine if a symptom is present or absent, as required by the DSM-5 criteria, ICU items were dichotomously coded using two different methods, which were compared. The more stringent method required extreme ratings on the item to be indicative of the symptom (coded as absent if rated 0 'not at all true' to 2 'very true', and present if rated equal to 3 'definitely true'), relative to the less stringent method (item coded as absent if rated 0 or 1, and present if rated either 2 or 3) used by Frick and Moffitt (2010) in developing the DSM-5 criteria. Hereafter, we refer to these as the 'extreme' and the 'split' coding methods, respectively.

Prior to carrying out IRT analyses, separate confirmatory factor analyses (CFA) were conducted using MPlus 6.1 statistical software (Muthén & Muthén, 2007) to formally test the unidimensionality assumption on which IRT relies (Reise & Henson, 2003). All CFAs with the dichotomous ICU items used weighted least squares means and variance adjusted (WLSMV) estimation, as recommended by Muthén and Muthén (2007). In addition to the Chi-square statistic, which indicates acceptable model fit with lack of significance, three standard fit indexes were used to evaluate model fit: The Root Mean-Square Error of Approximation (RMSEA), Standardized Root Mean Residual (SRMR), and the Comparative Fit Index (CFI). Cut-off values close to .06 for RMSEA, .08 for SRMR, and .95 for CFI were considered a good fit (Bollen, 1989).

Cronbach's coefficient alphas and item-to-total scale correlations were used to test the reliability of criteria set scores. Item-to-total scale correlations >.30 indicate good discrimination and Cronbach's alpha >.70 suggests that the item set is internally consistent (Nunnally & Bernstein, 1994). Frequency analyses were used to determine the prevalence of significant CU traits. Two-parameter IRT logistic models, based on available data from all countries, were applied to the two criteria sets to define the relationship between each criterion item and the underlying unobserved latent construct of interest (i.e. CU severity), performed separately for the extreme and split coding methods. IRT estimates two parameters for each item within each set: difficulty (threshold) and discrimination (slope). Item difficulty parameters represent the point along the CU latenttrait continuum at which 50% of the sample is likely to endorse an item, with higher threshold criteria being more severe and endorsed less frequently. Discrimination parameters indicate the strength of the relationship between the item and the underlying latent-trait, with higher values providing greater precision across the CU trait continuum. Item discrimination values >1.70 are considered very high discriminators, between 1.35 and 1.69 high, between 0.65 and 1.34 moderate, between 0.35 and 0.64 low, and <0.34 very low discriminators (de Ayala, 2009; Baker, 2001). Thus, item discriminations of 0.65 and above were considered acceptable.

Item characteristic curves (ICCs) were plotted and examined for each item within the two sets. The typical ICC has a well-defined S-shape indicating that the probability of endorsing a specific item increases monotonically as the latent-trait increases (Embretson & Reise, 2000). The difficulty parameter shifts the curve from left to right as the item criterion becomes more severe and the discrimination parameter is represented by the height of the curve's peak (higher curve = greater information and criterion discrimination). Item information curves (IICs) indicate the point along the CU latent-trait continuum that an item conveys the most information (i.e. reliability). All IRT models were analyzed using MPlus, which estimates item parameters via a maximum likelihood estimator with robust standard errors using a numerical integration algorithm. To assess overall model fit, likelihood ratio chi-square (G^2) statistics, Akaike's information criterion (AIC), and Bayesian information criterion (BIC) fit statistics were used (de Ayala, 2009). To refine item sets, items with high difficulty and low discrimination were removed and/or substituted with better functioning items assessing the same CU symptom category. In such cases, the AIC and BIC were used to compare the relative fit of the two nonnested IRT models, with lower values indicating better fit (Kang & Cohen, 2007; Rupp & Templin, 2010).

Finally, to test the discriminant validity of refined criteria sets, participants were categorized into CU specifier groups according to the number of criteria endorsed using extreme and split coding methods. The following two groups were formed for the four- and nine-item sets: those endorsing no symptoms or one symptom (i.e. not meeting CU specifier criteria) and ≥2 symptoms (i.e. meeting specifier criteria), reflecting the DSM-5 symptom threshold (APA, 2013). For the nine-item set, a symptom was considered endorsed if any of the items comprising that symptom's category were endorsed: shallow/deficient affect (item 1), unconcerned about performance (items 3,15), lack of remorse-guilt (5,13,16), callous-lack of empathy (items 8,17,24). The two groups were then compared on empirically supported external criterion measures using Analyses of Variance (ANOVA), controlling for age and gender, to test the validity of the two criteria sets.

Results

Confirmatory factor analysis

Four separate single factor CFA models in which all items loaded onto one overall factor representing CU traits were conducted. Using the extreme coding method, the four-item set provided good model fit, $\chi^2_{(2,\ N\ =\ 2{,}257)} = .83$, $p = .66$, $RMSEA = .01$ (RMSEA CI: 0.00|0.03), $SRMR = .01$, $CFI = .99$. However,

model fit was also acceptable using the split method with the four-item set ($\chi^2_{(2,\ N\ =\ 2,257)} = 14.28$, $p < .05$, *RMSEA* = .05(RMSEA CI: 0.03|0.07), *SRMR* = .04, *CFI* = .94). Model fit was acceptable for the nine-item set using the extreme coding method ($\chi^2_{(27,\ N\ =\ 2,257)} = 73.51$, $p < 001$, *RMSEA* = .025(RMSEA CI: 0.02|0.03), *SRMR* = .06, *CFI* = .93) or the split coding method ($\chi^2_{(27,\ N\ =\ 2,257)} = 230.32$, $p < .001$, *RMSEA* = .05(RMSEA CI: 0.04|0.06), *SRMR* = .06, *CFI* = .92). These findings provide support for model fit at the instrument level (de Ayala, 2009). Table 1 shows the resulting factor loadings of the four CFA models.

Extreme coding method

Prevalence and reliability of criteria sets. Cronbach's alphas were .40 for the four-item and .70 for the nine-item set (see Table 1 for ranges across samples). All items comprising the four- and nine-item sets had item-to-total scale correlations >.37, indicative of good discrimination (Table 2). The probability of endorsing each item ranged from 3.7% to 9.2% for the four-item set and from 2.7% to 15.3% for the nine-item set. Across samples, the prevalence of endorsing ≥2 CU symptoms from the four-item set was 3.8%, with significant gender differences (2.6% for boys, 1.2% for girls), $\chi^2_{(2,\ N\ =\ 2,257)} = 19.86$, $p < .001$. The prevalence rate for the nine-item set was 11.7%, with significant gender differences (8.2% for boys, 3.5% for girls), $\chi^2_{(1,\ N\ =\ 2,257)} = 29.89$, $p < .001$.

IRT analyses – four-item set. The four-item criteria set fit the data well, as suggested by a nonsignificant log-likelihood chi-square statistic, $G^2_{(8,\ N\ =\ 2,257)} = 17.05$, $p = .052$, *AIC* = 4543.20, *BIC* = 4583.76. Difficulty parameters indicated that higher levels of the latent CU trait were necessary to endorse item 6 (shallow/deficient affect symptom category). Discrimination parameters indicated that items 8 (callous-lack of empathy category) and 3 (unconcerned about performance category) best discriminated adolescents along the CU continuum. Figure 1A and 1B depicting plotted ICCs and IICs illustrate that items provided the greatest amount of information toward the higher end (i.e. more severe range) of the CU continuum. Items were more likely to be endorsed (i.e. higher reliability) among those possessing higher levels of the underlying CU trait, but had a low probability of endorsement across the sample. Item 3 provided the most information and item 6 the least, which corresponded to CFA findings.

A post hoc IRT analysis substituting item 6 with the other shallow/deficient affect item (item 1) from the nine-item set, showed a better fit to the data than the original model, as suggested by lower BIC and AIC values ($G^2_{(8,\ N\ =\ 2,257)} = 13.46$, $p = .10$, AIC changed from 4543.20 to 4534.32 and BIC changed from 4583.76 to 4574.88). Item 1 had lower difficulty (*Dif.* = 2.765, *SE* = .33) and higher discrimination within the acceptable range (*Dis* = .723, *SE* = .14) compared with item 6. This change decreased the item difficulty (*Dif.* = 2.536, *SE* = .34) and increased the discrimination (*Dis* = .639, *SE* = .11) values for item 5. Moreover, in the CFA model substituting item 6 with item 1, the factor loading for item 1 was .54 (*SE* = .06), $\chi^2_{(2,\ N\ =\ 2,257)} = .55$, $p = .76$, *RMSEA* = .01(RMSEA CI: 0.00|0.03), *SRMR* = .01, *CFI* = .99. The prevalence rate for the new nine-item set was 3.6%, with significant gender differences (5.3% for boys, 2% for girls), $\chi^2_{(1,\ N\ =\ 2,257)} = 17.17$, $p < .001$.

IRT – nine-item set. The nine-item criteria set also fit the data well, $G^2_{(494,\ N\ =\ 2,257)} = 416.07$, $p = .99$, *AIC* = 9681.07, *BIC* = 9779.56. Item 24 (callous-lack of empathy category) had the lowest difficulty parameter and item 8 had the highest, necessitating higher levels of the latent CU trait to endorse the latter. Item 13 (lack of remorse-guilt category) had the lowest discrimination parameter, whereas

Table 1 Factor loadings (standard errors) for ICU items from CFA analyses

	Extreme coding method		Split coding method	
Questionnaire items	Four-item	Nine-item	Four-item	Nine-item
1) I express my feelings openly		.51 (.05)		.26 (.03)
3) I care about how well I do at school or work	.69 (.08)	.61 (.06)	.68 (.07)	.53 (.03)
5) I feel bad or guilty when I do something wrong	.47 (.06)	.51 (.05)	.41 (.05)	.51 (.03)
6) I do not show my emotions to others	.36 (.07)		.33 (.05)	
8) I am concerned about the feelings of others	.65 (.08)	.57 (.06)	.42 (.06)	.37 (.04)
13) I easily admit to being wrong		.37 (.05)		.31 (.03)
15) I always try my best		.68 (.06)		.63 (.03)
16) I apologize to persons I hurt		.73 (.05)		.78 (.02)
17) I try not to hurt others' feelings		.75 (.05)		.76 (.03)
24) I do things to make others feel good		.59 (.04)		.57 (.03)
Cronbach's Alpha	.40 [.39–.41]	.70 [.56–.79]	.45 [.43–.47]	.72 [.61–.80]

All loadings statistically significant at the $p < .001$ level. All items except item 6 reverse-scored prior to analyses. Values in square brackets represent the range of alphas across samples.

doi:10.1111/jcpp.12357

Table 2 Item response theory parameters for four- and nine-item criteria sets

	Item-total scale correlation		% Endorsement		Difficulty		Discrimination	
	M1	M2	M1	M2	M1	M2	M1	M2
Four-item set								
3) I care about how well I do at school or work	.517 [.48–.62]	.608 [.59–.67]	4.8 [3.6–8.1]	20.9 [16.9–33.8]	2.439 (.25) [2.118–3.168]	1.263 (.11) [.542–2.055]	1.011 (.19) [.590–1.146]	0.834 (.11) [.468–1.307]
5) I feel bad or guilty when I do something wrong	.611 [.54–.68]	.621 [.60–.66]	8.7 [5.9–10]	35.8 [21.6–39.5]	2.905 (.40) [1.868–3.965]	0.884 (.11) [.619–1.421]	0.543 (.11) [.356–1.001]	0.436 (.06) [.153–.514]
6) I do not show my emotions to others	.617 [.53–.67]	.574 [.47–.61]	9.2 [6.7–17.3]	26.5 [23.2–33.1]	3.421 (.66) [2.289–3.804]	1.991 (.31) [1.194–1.714]	0.434 (.09) [.429–.717]	0.321 (.06) [.290–.492]
8) I am concerned about the feelings of others	.446 [.38–.63]	.492 [.40–.63]	3.7 [2.1–8.8]	12.9 [6.7–32.7]	2.769 (.29) [1.804–4.481]	2.676 (.36) [.620–3.644]	0.926 (.17) [.570–1.182]	0.467 (.07) [.462–1.066]
Nine-item set								
1) I express my feelings openly	.499 [.42–.54]	.411 [.38–.43]	9.5 [6.2–20.6]	52.5 [48.1–62.9]	2.629 (.24) [1.276–3.523]	1.581 (.06) [1.087–1.870]	0.623 (.08) [.438–.946]	0.257 (.03) [.204–.362]
3) I care about how well I do at school or work	.452 [.37–.54]	.474 [.45–.52]	4.8 [3.6–8.1]	20.9 [16.9–33.8]	2.741 (.10) [2.081–4.204]	1.608 (.13) [.801–2.097]	0.809 (.10) [.505–2.253]	0.569 (.05) [.478–.998]
5) I feel bad or guilty when I do something wrong	.478 [.45–.52]	.528 [.51–.61]	8.7 [5.9–10]	35.8 [21.6–39.5]	2.648 (.26) [2.270–3.051]	0.693 (.06) [.358–1.031]	0.614 (.08) [.487–.731]	0.596 (.04) [.554–.613]
8) I am concerned about the feelings of others	.377 [.20–.60]	.347 [.21–.58]	3.7 [2.1–8.8]	12.9 [6.7–32.7]	3.268 (.37) [1.677–4.064]	2.726 (.27) [.849–3.197]	0.701 (.11) [.049–1.455]	0.454 (.05) [.103–1.924]
13) I easily admit to being wrong	.502 [.51–.56]	.423 [.38–.56]	15.3 [6.1–19.8]	55.5 [27.4–61.4]	2.661 (.34) [1.641–4.029]	0.423 (.09) [.060–.758]	0.413 (.06) [.265–.535]	0.327 (.04) [.207–.641]
15) I always try my best	.399 [.21–.55]	.522 [.49–.60]	2.7 [2.1–6]	17.9 [15.7–26.2]	2.911 (.26) [2.262–4.882]	1.47 9 (.09) [1.106–1.927]	0.954 (.14) [.528–1.126]	0.757 (.06) [.616–1.035]
16) I apologize to persons I hurt	.467 [.35–.66]	.604 [.55–.68]	3.8 [3.1–5.6]	22.1 [19.9–38.3]	2.427 (.16) [2.003–3.043]	0.961 (.05) [0.406–1.268]	1.137 (.15) [.843–1.380]	1.29 9 (.11) [.913–1.876]
17) I try not to hurt others' feelings	.449 [.26–.64]	.576 [.53–.69]	3.1 [1.4–7]	20.8 [17.9–35.9]	2.570 (.19) [1.734–4.624]	1.062 (.05) [.419–1.327]	1.130 (.15) [.583–1.698]	1.165 (.10) [1.002–1.524]
24) I do things to make others feel good	.507 [.44–.55]	.560 [.55–.60]	8 [2.9–10.4]	36.4 [24.3–46.4]	2.332 (.18) [1.566–2.965]	0.606 (.05) [.123–.808]	1.008 (.09) [.814–1.465]	0.680 (.05) [.525–.832]

M1 = Extreme coding method 1 (0 = item rated below 3; 1 = item rated equal to 3); M2 = Split coding method 2 (0 = item rated either 0 or 1; 1 = item rated either 2 or 3); All items except item 6 reverse-scored prior to analyses; Values in parentheses represent standard errors (); Values in square brackets represent ranges of values across the different samples [ ]. For difficulty and discrimination parameters, ranges were obtained using differential item functioning (DIF) analyses.

items 16 (lack of remorse-guilt category) and 17 (callous-lack of empathy category) had the highest, suggesting that these latter items best discriminated adolescents along the CU trait continuum. This corresponded to CFA results in which item 13 had the lowest factor loading and items 16 and 17 the highest. Figure 2A and 2B displaying ICCs and IICs illustrate that items provided the greatest amount of information toward the higher end of the CU latent-trait continuum similar to the four-item set. Items 16 and 17 provided the greatest amount of information, whereas item 13 provided the least.

Given the poor functioning of item 13, the model was rerun with this item deleted. This new eight-item set also fit the data well ($G^2_{(240,\ N\ =\ 2{,}257)}$ = 277.03, p = .05, AIC = 7693.19, BIC = 7740.85). For the eight-item model, the discrimination values for item 1 (*Dis* = .671, *SE* = .08) and item 5 increased (*Dis* = .652, *SE* = .08), although their difficulty scores remained relatively unchanged. Moreover, the CFA model for the eight-item set, $\chi^2_{(20,\ N\ =\ 2{,}257)}$ = 56.38, p < .001, *RMSEA* = .025(RMSEA CI: 0.02|0.03), *SRMR* = .06, *CFI* = .94, fit the data equally well to the nine-item set, according to model fit indices. The prevalence rate for the eight-item set was 8%, with significant gender differences (11.4% for boys, 4.3% for girls), $\chi^2_{(1,\ N\ =\ 2{,}257)}$ = 39.89, p < .001.

Split coding method

Prevalence and reliability of criteria sets. Cronbach's alphas were .45 for the four-item set and .72 for the nine-item set (see Table 1 for ranges). Item total scale correlations were good for all items (all >.34, Table 2). The probability of endorsing each item ranged from 12.9% to 35.8% for the four-item set and 12.9–55.5% for the nine-item set. The prevalence of endorsing ≥2 CU symptoms from the four-item set was 24.8%, with a higher percentage of boys than girls (16.2% vs. 8.6%), $\chi^2_{(2,\ N\ =\ 2{,}257)}$ = 83.27, p < .001. The prevalence of the CU specifier for the nine-item set was 67.9%, with a higher percentage for boys than girls (40.3% vs. 27.6%), $\chi^2_{(2,\ N\ =\ 2{,}257)}$ = 81.61, p < .001.

IRT analyses – four-item set. Table 2 displays IRT results for the four-item set, $G^2_{(8,\ N\ =\ 2{,}257)}$ = 36.28, p < .001, *AIC* = 10742.07, *BIC* = 10782.62. Difficulty scores were higher when using the extreme method, consistent with expectations. All items coded using the extreme method also had higher

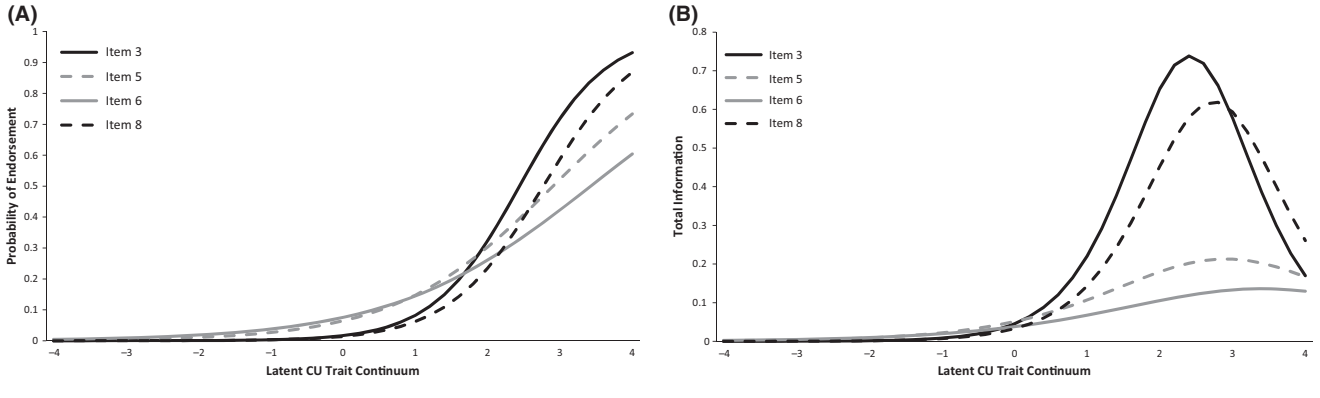

Figure 1 (A) Item characteristic curves (ICCs) and (B) Item information curves (IICs) for the four-item set using the extreme coding method

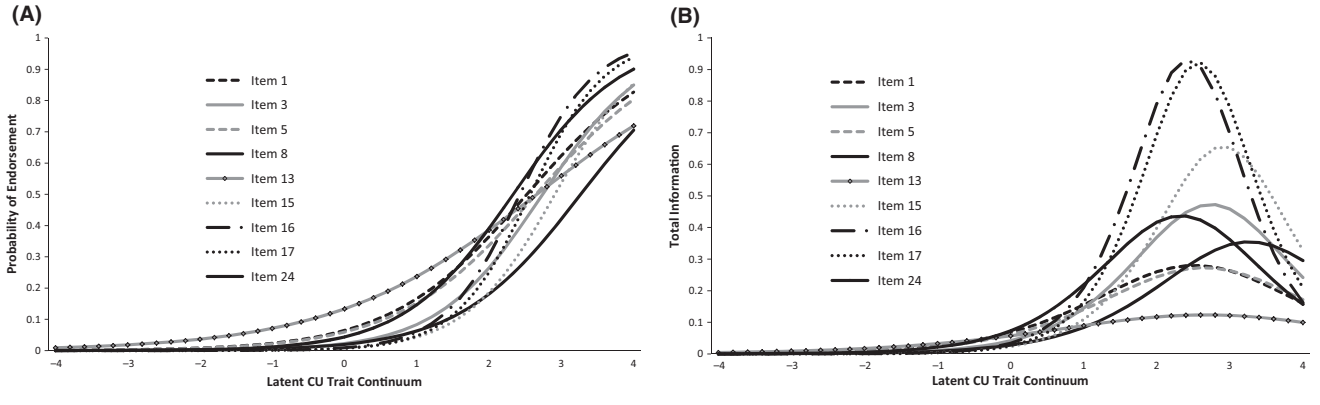

Figure 2 (A) ICCs and (B) IICs for the nine-item set using the extreme coding method

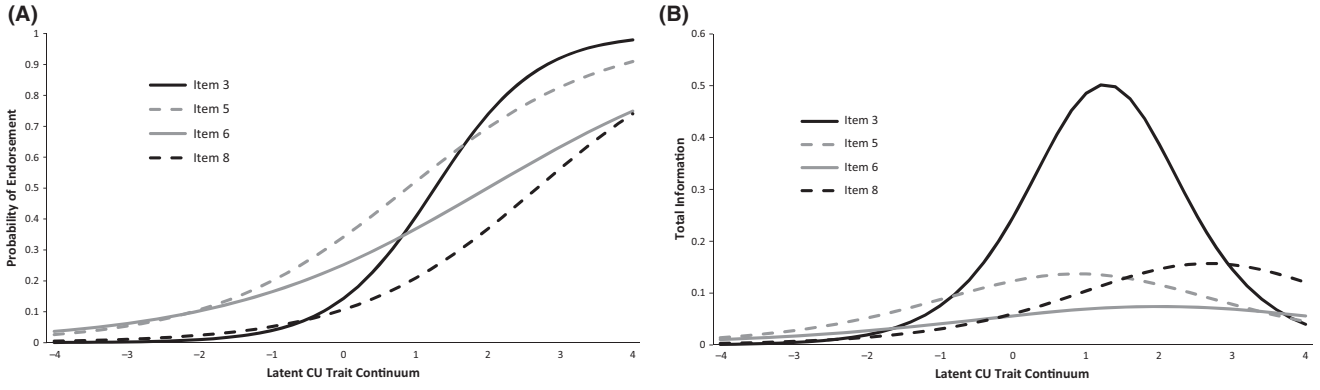

Figure 3 (A) ICCs and (B) IICs for the four-item set using the split coding method

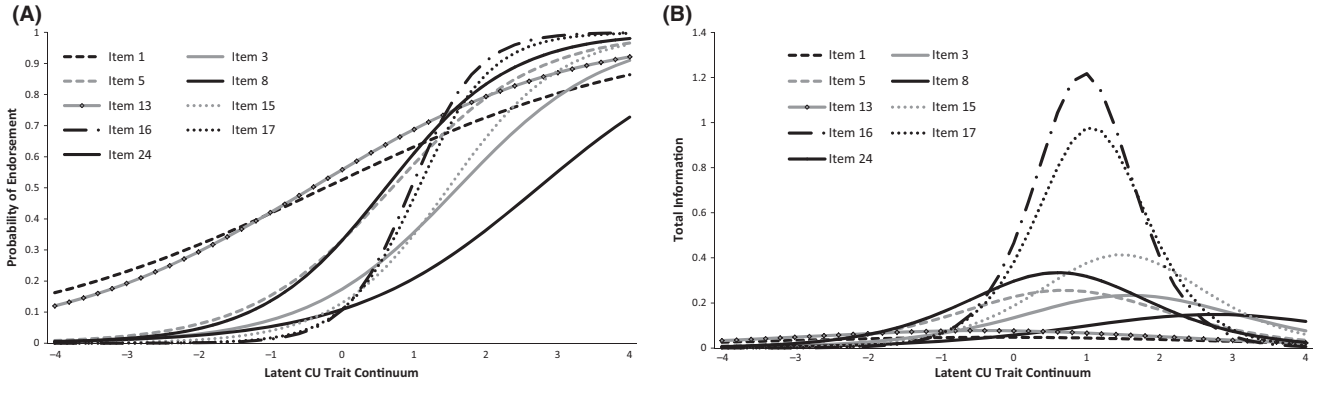

Figure 4 (A) ICCs and (B) IICs for the nine-item set using the split coding method

discrimination values compared to the split method, indicating that the former was more precise across the CU latent-trait continuum. Figure 3A and 3B display ICCs and IICs. Again, items provided the greatest amount of information toward the higher end of the CU continuum, albeit within a less severe range of the continuum than items coded using the extreme method. The steeper lines depicted in Figure 1A relative to Figure 3A reflect the higher discrimination values for the extreme over the split coding method. Regardless of coding method, item 3 provided the most information and item 6 the least. The greater peak height for all lines in Figure 1B compared with Figure 3B suggests that items dichotomized using the extreme method provided the most information.

A post hoc IRT analysis substituting item 6 with item 1 ($G^2_{(8,\ N\ =\ 2{,}257)}$ = 25.92, $p < .001$), showed a small change in BIC (changed from 10782.62 to 10235.89) and AIC (changed from 10742.07 to 10195.34). Item 1 had similar difficulty (*Dif.* = 2.079, *SE* = .29), but higher and acceptable discrimination (*Dis* = .723, *SE* = .14) values compared with item 6. In a CFA model substituting item 6 with item 1, the factor loading for item 1 was .34 (*SE* = .05), which was similar to the factor loading for item 6, $\chi^2_{(2,\ N\ =\ 2{,}257)} = 8.21$, $p = .02$, *RMSEA* = .04(RMSEA CI: 0.01|0.06), *SRMR* = .03, *CFI* = .97. The prevalence rate for the new four-item set was 35.9%, with significant gender differences (46.7% for boys, 23.8% for girls), $\chi^2_{(1,\ N\ =\ 2{,}257)} = 132.36$, $p < .001$.

IRT analyses – nine-item set. Item Response Theory (IRT) results for the nine-item set are also displayed in Table 2, $G^2_{(494,\ N\ =\ 2{,}257)}$ = 1283.50, $p < .001$, *AIC* = 23426.16, *BIC* = 23524.66. Results were consistent with analyses using the extreme coding method, although difficulty scores were lower. The majority of items coded using the extreme method had higher discrimination values than for the split method, with the exception of items 16 and 17. Figure 4A and 4B display ICCs and IICs, indicating that items provided the greatest amount of information toward the higher end of the CU latent-trait continuum; however, all items contributed information within a less severe range of the continuum than items dichotomized using the extreme coding method. Moreover, items in Figure 2A were characterized by steeper lines than the majority of items in Figure 4A, except items 16 and 17 that provided the greatest amount of information using the split coding method. Items 1 and 13 provided the least amount of information. The peak of the curves

for the majority of the items, except 16 and 17, were higher in Figure 2B compared with Figure 4B, providing partial support for using the extreme coding method.

Similar to the extreme coding method, we ran an IRT analysis deleting the poorest functioning item 13 ($G^2_{(240,\ N = 2,257)} = 698.71$, $p < .001$, AIC = 20108.20, *BIC* = 20195.11). The discrimination and difficulty values for the remaining items remained relatively unchanged. The CFA model for the eight-item set, $\chi^2_{(20,\ N = 2,257)} = 179.57$, $p < 001$, *RMSEA* = .05 (RMSEA CI: 0.04|0.06), *SRMR* = .06, *CFI* = .94, fit the data similarly to the nine-item set. The prevalence rate for the eight-item set was 54.5%, with significant gender differences (66.3% for boys, 41.5% for girls), $\chi^2_{(1,\ N = 2,257)} = 156.21$, $p < .001$.

Overlap between subtyping methods

The associations within the same scoring method but across item sets (e.g. four-item extreme with eight-item extreme) were fairly substantial, with both $r = .67$ ($p < .001$), whereas the associations within the same item set but using different scoring methods (e.g. the four-item extreme with the four-item split) were significant but modest, both $r = .27$ ($p < .001$). As noted above, the different methods varied in the prevalence of those identified with the specifier. The eight-item split method identified the largest group with the specifier ($n = 1341$; 55% of the sample). Of these, 872 (65%) were also identified by the four-item split method, 199 (15%) by the eight item extreme method; and 94 (7%) by the most conservative four-item extreme method. The four item split method identified the second largest group with the specifier ($n = 872$; 36% of sample). Of these, 159 (18%) were also identified by the eight-item extreme method; and 94 (11%) by the four-item extreme method. The eight-item extreme method identified the third largest group with the specifier ($n = 199$; 8% of sample). Of these, 94 (47%) were also identified by the four-item extreme method. The four-item extreme method identified the smallest group with the specifier ($n = 94$; 4% of sample).

External validity of the criteria sets

Tables 3 and 4 depict results of ANOVAs used to compare groups (formed using the two methods for coding symptom presence) on several variables assessing severity of antisocial behavior for the refined four-item (substituting item 6 for item 1) and eight-item (removing item 13) sets, respectively. Effect sizes were calculated using standard meta-analytic methods and were based on marginal means and *SE*s after controlling for age and gender (DerSimonian & Laird, 1986; Lipsey & Wilson, 2001). Weighted average effect sizes were computed to account for differences in the number of outcomes across studies (i.e. nonindependence). For the full sample, the weighted average effect size was greatest for the four-item CU specifier criteria set coded using the extreme method ($ES_w = .83$, $SE = .12$, 95% CI [0.61, 1.06]), and large in size according to effect size conventions (Cohen's $d = 0.20$ small, $d = .50$ medium, $d = .80$ large; Cohen, 1992). The remaining weighted average effect sizes were small to medium for the split method four-item set ($ES_w = .25$, $SE = 0.04$, 95% CI = 0.16, 0.34), the split method eight-item set ($ES_w = .35$, $SE = 0.04$, 95% CI = 0.26, 0.43), and the extreme method eight-item set ($ES_w = 0.51$, $SE = 0.08$, 95% CI = 0.36, 0.67).

Weighted average effect sizes were computed separately for community-based and incarcerated youth. For community youth, the four-item CU specifier criteria set coded using the extreme method yielded a large effect size [$ES_w = .98$, $SE = 0.13$, 95% CI (0.72, 1.22)]. The remaining effect sizes for community youth were small to medium for the split method four-item set ($ES_w = .23$, $SE = 0.05$, 95% CI = 0.14, 0.32), split method eight-item set ($ES_w = .33$, $SE = 0.04$, 95% CI = 0.24, 0.42), and extreme method eight-item set ($ES_w = .54$, $SE = 0.09$, 95% CI = 0.38, 0.71). For incarcerated youth, the eight-item CU criteria set coded using the split method yielded the largest effect, which was moderate in size ($ES_w = .64$, $SE = 0.18$, 95% CI [0.29, 0.99)]. The remaining effect sizes for incarcerated youth were small to medium for the extreme method four-item set ($ES_w = .20$, $SE = 0.27$, 95% CI = −0.33, 0.73), extreme method eight-item set ($ES_w = .32$, $SE = 0.22$, 95% CI = −0.11, 0.75), and split method four-item set ($ES_w = .47$, SE = 0.16, 95% CI = 0.15, 0.79).

Discussion

The current study tested methods for operationalizing nonnormative levels of CU traits for the DSM-5 'With Limited Prosocial Emotions' specifier for CD. Two ways of translating items from a self-report rating scale of CU traits (i.e. the ICU) into clinical decisions about symptom presence or absence, as required by the specifier, were compared. The results can be summarized by three key findings. First, original and refined (on the basis of IRT analyses) four- and eight-item sets from the ICU assessed a single underlying CU construct. Second, the refined four-item criteria set (substituting item 6 for item 1) provided good fit to the data. When coded using the more stringent 'extreme' method to assess the DSM-5 criteria (i.e. endorsement reflected by a rating of '3'), this set was superior in identifying *community* youth with severe antisocial and aggressive behavior than when using items from the refined eight-item set (removing item 13) to assess the four symptoms comprising the CU specifier. However, one limitation to using only four ICU items to assess the

doi:10.1111/jcpp.12357

Table 3 ANOVA results for four- and eight-item criteria sets using the extreme coding method

	Four-item				Eight-item			
	Low risk	CU specifier	*F*-value	Cohen's *d*	Low risk	CU specifier	*F*-value	Cohen's *d*
Belgian sample	(*n* = 439; 197 girls)	(*n* = 16; 2 girls)			(*n* = 428; 196 girls)	(*n* = 27; 3 girls)		
Antisocial behavior	7.99 (0.41)	20.82 (2.17)	33.63**	1.50	7.68 (0.41)	20.62 (1.64)	58.41**	1.53
German sample	(*n* = 1257; 626 girls)	(*n* = 27)			(*n* = 1194; 608 girls)	(*n* = 87; 16 girls)		
Externalizing pr.	15.43 (0.23)	18.93 (1.53)	4.36*	0.43	15.42 (0.24)	17.29 (0.75)	5.54**	0.23
Conduct problems	4.29 (0.10)	8.04 (0.69)	15.72**	1.06	4.02 (0.09)	4.62 (0.30)	4.83*	0.19
Aggression and antisocial behavior	7.73 (0.22)	17.66 (1.48)	44.06**	1.27	7.71 (0.23)	11.41 (0.88)	16.47**	0.46
American sample	(*n* = 143; 59 girls)	(*n* = 15; 7 girls)			(*n* = 133; 54 girls)	(*n* = 25; 12 girls)		
Proactive aggression	6.10 (0.61)	6.88 (1.87)	0.82	0.04	5.73 (0.61)	9.85 (1.40)	7.17**	0.59
Total delinquency	13.41 (0.56)	16.25 (1.73)	2.42	0.43	13.29 (0.58)	15.75 (1.34)	2.79	0.37
Property delinquency	4.63 (0.23)	4.77 (0.70)	0.01	0.05	4.60 (0.25)	4.87 (0.45)	.75	0.01
Violent delinquency	2.78 (0.14)	3.26 (0.42)	1.11	0.29	2.75 (0.14)	3.23 (0.32)	1.78	0.30
Cypriot sample	(*n* = 308; 172 girls)	(*n* = 21; 9 girls)			(*n* = 295; 165 girls)	(*n* = 34; 16 girls)		
Proactive aggression	2.78 (0.19)	4.74 (0.72)	5.89**	0.59	2.72 (0.19)	4.20 (0.56)	5.96**	0.45
Bullying	6.21 (0.48)	12.30 (1.88)	9.79**	0.72	6.08 (0.49)	10.91 (1.43)	10.01**	0.43

CU, callous-unemotional; Estimated marginal means (*SE*) controlling for gender and age; *$p \leq .05$; **$p \leq .01$. All $df = 1$. Results for the four-item model substitute item 1 for item 6; Results for the eight-item model eliminate item 13. There were minimal differences between the eight- and nine-item models.

Table 4 ANOVA results for four- and eight-item criteria sets using the split coding method

	Four-item				Eight-item			
	Low risk	CU specifier	*F*-value	Cohen's *d*	Low risk	CU specifier	*F*-value	Cohen's *d*
Belgian sample	(*n* = 256; 141 girls)	(*n* = 199; 58 girls)			(*n* = 198; 119 girls)	(*n* = 257; 80 girls)		
Antisocial Behavior	6.79 (0.56)	10.59 (0.63)	19.55**	0.43	6.35 (0.64)	10.07 (0.55)	18.21**	0.42
German sample	(*n* = 883; 524 girls)	(*n* = 401; 101 girls)			(*n* = 599; 400 girls)	(*n* = 682; 224 girls)		
Externalizing pr.	15.36 (0.28)	15.57 (0.42)	0.16	0.03	14.74 (0.34)	16.17 (0.32)	8.87**	0.17
Conduct problems	4.09 (0.11)	4.93 (0.18)	14.10**	0.25	3.69 (0.15)	4.91 (0.13)	33.34**	0.35
Aggression and antisocial behavior	7.58 (0.26)	8.87 (0.42)	6.51**	0.16	6.69 (0.32)	9.14 (0.31)	27.95**	0.31
American sample	(*n* = 69; 29 girls)	(*n* = 89; 37 girls)			(*n* = 45; 17 girls)	(*n* = 113; 49 girls)		
Proactive aggression	4.43 (0.84)	7.89 (0.74)	9.54**	0.50	3.10 (1.03)	7.69 (0.64)	14.10**	0.68
Total delinquency	11.72 (0.78)	15.20 (0.69)	10.93**	0.54	10.24 (0.95)	15.05 (0.60)	17.97**	0.76
Property delinquency	3.99 (0.32)	5.14 (0.28)	7.32**	0.44	3.39 (0.38)	5.13 (0.25)	14.21**	0.67
Violent delinquency	2.48 (0.19)	3.10 (0.17)	5.67*	0.39	2.32 (0.24)	3.03 (0.15)	6.11*	0.45
Cypriot sample	(*n* = 217; 126 girls)	(*n* = 112; 55 girls)			(*n* = 169; 106 girls)	(*n* = 160; 75 girls)		
Proactive aggression	2.45 (0.23)	3.68 (0.31)	10.15**	0.37	2.16 (0.26)	3.63 (0.26)	16.40**	0.44
Bullying	6.11 (0.58)	7.49 (0.81)	2.06	0.16	4.92 (0.64)	8.33 (0.66)	13.54**	0.41

CU, callous-unemotional; Estimated marginal means (*SE*) controlling for gender and age; *$p \leq .05$; **$p \leq .01$. All $df = 1$. Results for the four-item model substitute item 1 for item 6; Results for the eight-item model eliminate item 13. There were minimal differences between the eight- and nine-item models.

specifier was the very poor internal consistency of this criteria set, which is likely due to the small number of items and our transformation of the ordinal rating format to dichotomous scores of symptom presence. This is supported by the size of the item-total correlations, which ranged from .45 to .62 for the four-item set and from .35 to .60 for the nine-item set. Third, the lenient 'split' coding method (i.e. endorsement reflected by ratings of '2' or '3'), particularly when applied to the eight-item criteria set, most consistently discriminated detained youth with high levels of proactively aggressive and violent delinquent behavior (see Kimonis et al., 2014).

Across settings, the prevalence rate for those elevated on CU traits using the extreme coding method with the four-item criteria set was 3.8%, whereas it was 25% when using the less stringent split method with the four-item set. Boys were also roughly two times more likely to meet the CU specifier than girls. There is currently no clear consensus as to the most appropriate base rate for defining nonnormative and impairing levels of CU traits and this will likely depend on the setting (e.g. community, clinic-referred, incarcerated), number and types of informants, and the purpose for making this diagnosis (e.g. the importance of avoiding false positives vs. avoiding false negatives; Kahn et al., 2012). However, Frick and Viding (2009) estimated that 2–4% of all children show a joint CP+CU presentation, on the basis of CU prevalence estimates obtained within samples of children with conduct problems. Also, in their Fast Track sample, McMahon et al. (2010) reported a prevalence rate of 5% for the CU specifier, regardless of the presence or absence of CD. These estimates more closely approximate the 2–6% prevalence rate identified for community samples using the four-item set and extreme coding method (Tables 3 and 4) than the overly inclusive 29–34% rate identified using the split coding method.

Several possibilities might explain why the optimal method for coding CU symptom presence varied according to setting. The range of difficulty and discrimination parameters reported in Table 2 implicated that certain items (e.g. items 8 and 16) may not have functioned equivalently across samples, indexing different ranges of the CU continuum with varying levels of precision across separate populations of youth. A variety of sources might account for this differential item functioning, such as disparities in the severity of antisocial behavior across samples, language differences, or cultural factors, which in turn may account for the variability in effect sizes across samples. These findings highlight the difficulties when attempting to translate continuous measures into dichotomous diagnostic decisions. Future studies are needed to understand what factors contribute to this heterogeneity in ICU item functioning across samples.

As highlighted above, certain ICU items functioned better than others in assessing self-reported CU traits, which begins to build a database for determining indicators that might best identify those meeting the DSM-5 specifier criteria (APA, 2013). Items 3 and 8 that correspond to lack of concern over performance and callousness-lack of empathy symptoms in DSM-5, respectively (four-item set), and 16 and 17 that correspond to lack of remorse/guilt and callous-lack of empathy symptoms, respectively (eight and nine-item sets), best discriminated youth along the CU continuum. These items also provided the greatest amount of information toward the higher end of the CU continuum, suggesting that they have a low probability of endorsement across the sample, but are more likely to be endorsed by those who possess higher levels of the underlying CU trait. Item 6 ('I do not show my emotions to others') and item 13 ('I easily admit to being wrong') – corresponding to the shallow-deficient affect and lack of remorse-guilt symptoms of the DSM-5 specifier, respectively – functioned poorly in IRT analyses. These items poorly discriminated youth that similarly endorsed other items in the criteria set and were replaced (item 6 with item 1) or removed (item 13) to refine ICU criteria sets. In support of these refinements, Kimonis et al. (2014) also found that replacing item 6 and removing item 13 improved model fit in their sample of 643 incarcerated adolescents.

Youth endorsing item 6, specifically, tended to fall at the very high end of the CU latent-trait continuum. Factor analytic research finds that nonreverse-scored ICU items, such as item 6, largely load on a different factor that tends to correlate moderately with reverse-scored items, leaving open the possibility that the difference in wording might explain the poorer functioning of this item (see Hawes et al., 2013 for a discussion). Alternatively, these data suggest that more research is needed into the optimal methods for assessing the shallow and deficient affective style included in the DSM-5 specifier. For example, items tapping shallow/deficient affect may require a change in wording to clarify that the youth is capable of turning emotions on and off at will and/or using emotions (e.g. anger) to get what he/she wants from others, rather than failing to express emotions at all (American Psychiatric Association, 2013). Moreover, the affective deficit may be specific to experiencing certain emotions, such as sadness and fear (Pardini, Lochman, & Frick, 2003; Stevens, Charman, & Blair, 2001), and ICU shallow/deficient affect items may inadequately capture this level of specificity. It is also possible that shallow-deficient affect might be best tapped using a multimethod approach that incorporates laboratory measures of emotional processing (Kimonis, Frick, Muñoz, & Aucoin, 2007; Muñoz,

2009). In short, an important goal for future research is to test for the optimal indicators of the deficient affect component of CU traits.

In addition to several strengths, such as the use of a large and heterogeneous cross-national sample, testing the symptom criteria using IRT, the use of a well-validated inventory measuring CU traits, and evaluating a variety of antisocial behaviors, the findings must be interpreted in light of several study limitations. First, CU traits were assessed solely through self-report although the DSM-5 criteria explicitly recognize the importance of carefully considering multiple sources of information from people who have known the individual for extended periods of time and across relationships and settings (APA, 2013). In practice, it will be important to combine information gleaned from the self-report ICU with other sources (e.g. parents, teachers, coworkers, extended family members, peers, official records) when evaluating the limited prosocial emotions specifier. Furthermore, self-report measures that work well under research conditions of complete confidentiality may work less well when applied in other settings where self-reports bring actual consequences to the reporter. However, in many forensic settings, the youthful offender may be the only source of information about CU criteria available. Second, the external validators were all self-report as well, which could have inflated validity estimates due to shared method variance. Future research is needed to examine whether these findings generalize to parent and teacher reports of ICU, with IRT studies using the full scale serving as a possible starting point for this research (e.g. Hawes et al., 2013). Third, external validation was based solely on testing the utility of the CU symptoms sets for designating youth with severe antisocial and aggressive behavior, which may not be the optimal benchmark. Other correlates that are important to the construct of CU traits, such as cognitive (e.g. punishment insensitivity), emotional (e.g. distress insensitivity) and biological (e.g. reduced amygdala activation) variables should be considered when validating measures of this construct (Frick, Ray, Thornton, & Kahn, 2014). Particularly among incarcerated samples with high levels of antisocial behavior, it stands to reason that those meeting the CU specifier might show similar levels of self reported antisocial behavior to those not meeting the specifier despite showing different emotion processing deficits and higher rates of recidivism, thus supporting that the specifier is not simply indexing severity.

Fourth, the presence of CD was not assessed and the application of the specifier 'With Limited Prosocial Emotions' requires that the person meets full criteria for CD before it can be given. Thus, future research should test whether or not the best indicators of CU traits differ in those who do and who do not meet criteria for CD. Further research is also needed to test whether effects are moderated by gender, which was not possible in the current study due to sample size limitations. Finally, the different methods for assessing the construct of CU traits that were tested required that the 4-point response format on the ICU be converted into dichotomous ratings of symptom presence or absence, which does not take into account the full range of ratings on this scale. This was done to approximate the clinical decisions required by the DSM-5 criteria. Supporting this use of the ICU to identify youth at the severe and impairing range of the CU continuum, our IRT analyses indicated that the CU symptoms largely discriminated best at high levels of the trait. For example, item difficulty parameters for the four-item criteria set and the extreme coding method best discriminated youth in the upper 5% of the CU construct (range 2.4–3.4 *SD*). While appropriate for diagnostic purposes and applied research focused on placing children with severe conduct problems into groups high and low on CU traits, information provided by the complete ICU response format and using all 24 items on the scale may be more appropriate for examining the full range of the CU construct.

Within the context of these limitations, our results provide some support for the DSM-5 criteria developed to define significant levels of CU traits. As noted above, this designation could be critical for identifying unique developmental trajectories leading to the antisocial behavior of youth at high risk for future impairment (Frick et al., 2014). Antisocial youths with elevated CU traits may benefit less from traditional mental health approaches and require more intensive, comprehensive, and specialized interventions that are tailored to their unique emotional, cognitive, and motivational styles (Hawes, Price, & Dadds, 2014). For example, in a study of 177 clinic-referred children, those with CU traits who received an individualized and comprehensive modular intervention evinced similar rates of improvement to other children with CD (Kolko & Pardini, 2010). This encouraging intervention research clearly supports the importance of refining our methods for assessing CU traits to guide appropriate diagnosis and subsequent tailored treatment.

Acknowledgements

No external funding was received for this work. The authors have declared that they have no competing or potential conflicts of interest.

Correspondence

Eva R. Kimonis, The School of Psychology, Mathews Building, The University of New South Wales, Sydney, NSW 2052, Australia; Email: e.kimonis@unsw.edu.au

Key points

- Callous-unemotional (CU) traits are important for identifying a unique subgroup of antisocial youths at risk for severe, persistent, and impairing conduct problems that have been attributed to distinct etiological processes and require specialized intervention.
- The self-report Inventory of Callous-Unemotional Traits (ICU) may be used as one source of information within a multimethod assessment of clinically significant CU traits for the DSM-5 'With Limited Prosocial Emotions' specifier to conduct disorder.
- Youth endorsing ≥2 CU symptoms, on the basis of scores on IRT-refined four- and eight-item ICU criteria sets, had the highest levels of antisocial and proactively aggressive behavior. A stringent method of scoring four ICU items (i.e. taking only the most extreme responses) to approximate symptom presence best discriminated among community youth, whereas a less stringent 'split' coding method best discriminated among detained youth.
- Boys were approximately twice as likely as girls to meet the CU specifier using ICU criteria sets.

References

Achenbach, T.M. (1991). *Manual for the Youth Self-report and 1991 profile*. Burlington: University of Vermont.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th edn.). Washington, DC: American Psychiatric Association.

de Ayala, R.J. (2009). *The theory and practice of Item Response Theory*. New York, NY: The Guilford Press.

Baker, F.B. (2001). *The basics of Item Response Theory* (2nd edn). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Brislin, R.W. (1970). Back translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185–216.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*, 177–188.

Elliott, D.S., & Ageton, S. (1980). Reconciling ethnicity and class differences in self-reported and official estimates of delinquency. *American Sociological Review*, *45*, 95–110.

Embretson, S.E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Essau, C.A. (2000). *Angst und Depression bei Jugendlichen (Anxiety and depression in adolescents)*. Thesis for a post-doctoral degree. Bremen: University of Bremen.

Essau, C.A., Sasagawa, S., & Frick, P.J. (2006). Callous-unemotional traits in community sample of adolescents. *Assessment*, *13*, 454–469.

Ezpeleta, L., 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 & Adolescent Psychology*, *42*, 91–105.

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.

Frick, P.J. (2004). *The Inventory of callous-unemotional traits*. New Orleans, LA: The University of New Orleans.

Frick, P.J., & Moffitt, T.E. (2010). A proposal to the DSM-V childhood disorders and the ADHD and disruptive behavior disorders work groups to include a specifier to the diagnosis of conduct disorder based on the presence of callous-unemotional traits. APA. Washington, DC. Available from http://www.dsm5.org/Proposal%20Revision%20Attachments/Proposal%20for%20Callous%20and%20Unemotional%20Specifier%20of%20Conduct%20Disorder.pdf. [last accessed July 23, 2010].

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.

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

Frick, P.J., & Viding, E. (2009). Antisocial behavior from a developmental psychopathology perspective. *Development & Psychopathology*, *21*, 1111–1131.

Frick, P.J., & White, S.F. (2008). Research review: the importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. *Journal of Child Psychology & Psychiatry*, *49*, 359–375.

Hare, R.D., & Neumann, C.S. (2008). Psychopathy as a clinical and empirical construct. *Annual Review of Clinical Psychology*, *4*, 217–4246.

Hawes, S.W., Byrd, A.L., Henderson, C.E., Gazda, R.L., Burke, J.D., Loeber, R., & Pardini, D.A. (2013). Refining the parent-reported Inventory of Callous-Unemotional Traits in boys with conduct problems. *Psychological Assessment*, *26*, 256–266.

Hawes, D.J., Price, M.J., & Dadds, M.R. (2014). Callous-unemotional traits and the treatment of conduct problems in childhood and adolescence: A comprehensive review. *Clinical Child and Family Psychology Review*, *17*, 248–267.

Kahn, R.E., Frick, P.J., Youngstrom, E., Findling, R.L., & Youngstrom, J.K. (2012). The effects of including a callous–unemotional specifier for the diagnosis of conduct disorder. *Journal of Child Psychology and Psychiatry*, *53*, 271–282.

Kang, T., & Cohen, A.S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*, 331–358.

Kimonis, E.R., Fanti, K., Goldweber, A., Marsee, M.A., Frick, P.J., & Cauffman, E. (2014). Callous-unemotional traits in incarcerated adolescents. *Psychological Assessment*, *26*, 227–237.

Kimonis, E.R., Frick, P.J., Muñoz, L.C., & Aucoin, K.J. (2007). Can a laboratory measure of emotional processing enhance the statistical prediction of aggression and delinquency in detained adolescents with callous-unemotional traits? *Journal of Abnormal Child Psychology*, *35*, 773–785.

Kimonis, E.R., Frick, P.J., Skeem, J., 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. *Journal of the*

International Association of Psychiatry and Law, *31*, 241–251.

Kolko, D.J., & Pardini, D.A. (2010). ODD dimensions, ADHD, and callous-unemotional traits as predictors of treatment response in children with disruptive behavior disorders. *Journal of Abnormal Psychology*, *119*, 713–725.

Krueger, R.F., Schmutte, P.S., Caspi, A., Moffitt, T.E., Campbell, K., & Silva, P.A. (1994). Personality traits are linked to crime among men and women: Evidence from a birth cohort. *Journal of Abnormal Psychology*, *103*, 328–338.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*, Vol. *49*. Thousand Oaks, CA: Sage.

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.

McMahon, R. J., Witkiewitz, K., Kotler, J. S., & the Conduct Problems Prevention Research Group. (2010). Predictive validity of callous-unemotional traits measured in early adolescence with respect to multiple antisocial outcomes. *Journal of Abnormal Psychology*, *119*, 752–763.

Muñoz, L.C. (2009). Callous-unemotional traits are related to combined deficits in recognizing afraid faces and body poses. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*, 554–562.

Muthén, L.K., & Muthén, B.O. (2007). *Mplus user's guide*, 5th edn. Los Angeles, CA: Muthén & Muthén.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*, 3rd edn. New York, NY: McGraw-Hill.

Pardini, D.A., Lochman, J.E., & Frick, P.J. (2003). Callous/unemotional traits and social-cognitive processes in adjudicated youths. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*, 364–371.

Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Raynolds, C .... & Liu, J. (2006). The reactive-proactive aggression questionnaire: Differential correlated of reactive and proactive aggression in adolescent boys. *Aggressive Behavior*, *32*, 159–171.

Reise, S.P., & Henson, J.M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*, 93–103.

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.

Rowe, R., Maughan, B., Moran, P., Ford, T., Briskman, J., & Goodman, R. (2010). The role of callous and unemotional traits in the diagnosis of conduct disorder. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *51*, 688–695.

Rupp, A., & Templin, J. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Schwab-Stone, M., Chen, C., Greenberger, E., Silver Lichtman, J., & Voyce, C. (1999). No safe haven, II: The effects of violence exposure on urban youth. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*, 359–367.

Stevens, D., Charman, T., & Blair, R.J.R. (2001). Recognition of emotion in facial expressions and vocal tones in children with psychopathic tendencies. *Journal of Genetic Psychology*, *16*, 201–211.

Varjas, K., Meyers, J., & Hunt, M.H. (2006). *Student survey of bullying behavior– revised 2 (SSBB-R2)*. Atlanta, GA: Georgia State University, Center for Research on School Safety, School Climate and Classroom Management.

Viding, E., Sebastian, C.L., Dadds, M.R., Lockwood, P.L., Cecil, C.M., 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. *The American Journal of Psychiatry*, *169*, 1109–1116.

Accepted for publication: 29 September 2014
First published online: 31 October 2015