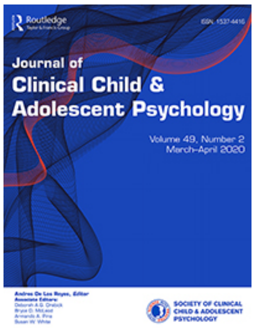

Journal of Clinical Child & Adolescent Psychology

ISSN: 1537-4416 (Print) 1537-4424 (Online) Journal homepage: https://www.tandfonline.com/loi/hcap20

Assessing Callous-Unemotional Traits Using the Total Score from the Inventory of Callous-Unemotional Traits: A Meta-Analysis

James V. Ray & Paul J. Frick

Journal of Clinical Child & Adolescent Psychology, 49(2), 190–199, 2020

ISSN: 1537-4416 print/1537-4424 online
DOI: https://doi.org/10.1080/15374416.2018.1504297

Routledge
Taylor & Francis Group

Assessing Callous-Unemotional Traits Using the Total Score from the Inventory of Callous-Unemotional Traits: A Meta-Analysis

James V. Ray
Department of Criminal Justice, University of Central Florida

Paul J. Frick
Department of Psychology, Louisiana State University and Institute for Learning Sciences and Teacher Education, Australian Catholic University

The Inventory of Callous-Unemotional Traits (ICU) is a widely used rating scale measure of callous-unemotional (CU) traits. Most studies have used a unit-weighted total score that sums all of the items into a composite, despite the consistent finding that items from this measure can be described using a bifactor model (1 general factor and 3 bifactors). We conducted a meta-analysis using the results of past published bifactor tests of the ICU, using indices to estimate the sources of variance in the total and subscale scores. We included studies in our review that tested and found adequate fit for the bifactor model of the ICU resulting in 12 published studies (*M* age = 9–26 from both community and justice-involved samples) using either the self-report (n = 9) or parent-report (n = 4) versions of the ICU scale (one published study conducted separate factor analyses for each version). We reported model-based reliability estimates that break down unit-weighted ICU scores into the variance due to individual differences in the general factor and the variance due to individual differences in the bifactors. The reliable variance in the unit-weighted total scale score of the ICU is largely determined by the general factor. Findings support the ICU total score as a measure of the general construct of CU traits, despite support for a bifactor model in multiple samples. Concerns about using the unit-weighted subscale scores from the ICU are raised since their reliable variance was strongly influenced by the general factor.

Callous-unemotional (CU) traits are defined by a lack of empathy, a lack of guilt, a failure to put forth effort in important activities, and shallow emotions (Kimonis et al., 2015). These traits have been used to define the affective components of psychopathy in adult samples (Hare & Neumann, 2008) and the affective components of conscience in child samples (Frick, Ray, Thornton, & Kahn, 2014a). Further, there is now rather substantial evidence to support the importance of CU traits for designating a clinically important subgroup of antisocial youth. That is, recent reviews of the available literature have shown that the presence of elevated CU traits in children and adolescents with serious behavior problems designates a group that is especially severe, violent, and difficult to treat (Blair, Leibenluft, & Pine, 2014; Frick, Ray, Thornton, & Kahn, 2014b; Herpers, Rommelse, Bons, Buitelaar, & Scheepers, 2012). These reviews have also indicated that children and adolescents with elevated CU traits show a number of distinct genetic, biological, emotional, cognitive, and social characteristics when compared to antisocial youth who are not elevated on these traits, suggesting that the causal processes underlying the behavior problems of these two groups are different. As a result, the most recent edition of the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; American Psychiatric Association, 2013) has included CU traits in its specifier for the diagnosis of conduct disorder labeled “with limited prosocial emotions.”

Correspondence should be addressed to James V. Ray, Department of Criminal Justice, University of Central Florida, Orlando, FL 32816. E-mail: james.ray@ucf.edu

Based on this extensive research, there is a clear need to measure CU traits for both clinical and research purposes. To date, one of the most commonly used measures to assess CU traits in research is the Inventory of Callous-Unemotional Traits (ICU; Kimonis et al., 2008). The ICU is a 24-item behavior rating scale that includes forms for self-report, as well as parent and teacher ratings. It was designed to explicitly capture the affective dimension of psychopathy in a way that overcomes limitations in other scales in which CU traits are assessed as one component of the broader construct of psychopathy (Kotler & McMahon, 2005; Sharp & Kine, 2008). Specifically, the ICU items were developed to (a) provide a focused and comprehensive assessment of CU traits only (and not other dimensions of psychopathy), (b) include a rating format that allows for sufficient variability in responses but does not include a central tendency point (i.e., items are anchored on a 4-point Likert scale from 0 [*not at all true*] to 3 [*definitely true*]), and (c) include equal numbers of items rated in the positive (i.e., higher rating indicating higher levels of CU traits) and negative (i.e., higher ratings indicating lower levels of CU traits) directions (Frick & Ray, 2015). The ICU has been translated into more than 20 languages and has been used widely in research, with more than 150 published studies using the ICU in samples ranging in age from 3 years to young adulthood (https://sites01.lsu.edu/faculty/pfricklab/icu/). Across these studies, the unit-weighting of the ICU items has led to an internally consistent total score, although Items 2 ("What I think is right and wrong is different from what other people think") and 10 ("I do not let my feelings control me") often show very low item-total correlations, and overall internal consistency is often increased by removing them from the composite (Kimonis et al., 2008; Ray, Frick, Thornton, Steinberg, & Cauffman, 2016). Further, this total score has been shown to designate a distinct group of antisocial youth who show more severe antisocial behavior and who show distinct correlates to their conduct problems, supporting the clinical validity of this measure of CU traits (Frick & Ray, 2015).

A number of studies have tested the factor structure of the ICU (see Table 1), and the most consistent finding is that the best-fitting model tends to be a bifactor model specifying a general CU factor and three independent subfactors: callousness (capturing a lack of empathy and remorse), uncaring (capturing an uncaring attitude about performance on tasks and other's feelings), and unemotional (capturing deficient emotional affect). This factor structure has been supported in preschool children (Ezpeleta, Osa, Granero, Penelo, & Domènech, 2013), middle school children (Ciucci, Baroncelli, Franchi, Golmaryami, & Frick, 2014), and adolescents (Essau, Sasagawa, & Frick, 2006); it has been supported in community samples (Essau et al., 2006) and samples of juvenile offenders (Kimonis et al., 2008); and it has been supported across multiple language translations (Ciucci et al., 2014; Essau et al., 2006; Ezpeleta et al., 2013; Fanti, Frick, & Georgiou, 2009). Further, this factor structure has been supported for both self and other (i.e., parent and teacher) report formats (Roose, Bijttebier, Decoene, Claes, & Frick, 2010), and it has been found to be invariant across gender (Ciucci et al., 2014; Essau et al., 2006).

Based on this research, it appears that this three-dimensional model of CU traits is fairly robust across age, language, rater, and gender (Cardinale & Marsh, 2017). This is not to say that there have not been a number of studies reporting alternate factor structures (Feilhauer, Cima, & Arntz, 2012). For example, Feilhauer et al. (2012) found that a five-factor model best fit the data in a forensic sample of youth (n = 383; 13–20 years old) that consisted of a Lack of Conscience and Lack of Empathy factors in addition to the Callousness, Uncaring, and Unemotional factors typically identified. Houghton, Hunter, and Crow (2013) found support for a two-factor structure that consisted of only a Callous and Uncaring factor in a community sample of 268 Australian children (8–13 years old). In a large (N = 1,078) community sample of children (M age = 7 years), Willoughby et al. (2014) found support for a two-factor model as well that consisted of Empathic-Prosocial and Callous-Unemotional factors. Thus, the three bifactor structure is not always found to be the best-fitting in some samples; in fact, some studies have failed to find adequate fit for the three bifactor structure (e.g., Hawes et al., 2014; Kimonis et al., 2016; Willoughby et al., 2014). However, no alternative model has been consistently supported, and these other studies all support an overarching general CU factor with multiple underlying dimensions. Thus, there are consistent findings to suggest that CU traits may be better considered as a multidimensional construct.

However, there are also a number of problems with this conceptualization. First, and most important, the three factors that have emerged from past research were not predicted from a strong theoretical model for how these factors might explain the construct of CU traits nor have consistent and distinct correlates to these factors emerged that would support such a theoretical model (Frick & Ray, 2015). Further, it is possible that the factors to some extent represent 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"; Hawes et al., 2014). Further, in an ethnically diverse sample of 1,190 first-time justice-involved adolescents, Ray et al. (2016) conducted an item response theory (IRT) analysis of the ICU items and suggested that this method variance could reflect different patterns of item endorsement and differences in item difficulty. That is, 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 and showed higher difficulty levels in IRT analyses (i.e., discriminated best at higher levels of CU traits). Thus, these findings suggest that the factors that have emerged in

TABLE 1
Model-Based Reliabilities for Included Studies

	Sample Description	*Factor*	*No. of Items Included*	*Items Removed*	*Omega*	*Omega Subscale*	*OmegaH*	*OmegaHS*
Self-Report ICU								
Byrd, Kahn, and Pardini (2013)	N = 425; M age = 25.78; 100% male;	General CU	24	None	.904		.607	
	community sample;	Callousness	11			.866		.725
	United States	Uncaring	8			.908		.106
		Unemotional	5			.572		.385
Ciucci et al. (2014)	N = 540; M age = 12.58;	General CU	22	2, 10	.878		.572	
	47.4% male;	Callousness	9			.810		.464
	community sample;	Uncaring	8			.849		.464
	Italy	Unemotional	5			.706		.542
Essau et al. (2006)	N = 1,443; M age = 15.59;	General CU	24	None	.790		.382	
	53.6% male;	Callousness	11			.694		.553
	community sample;	Uncaring	8			.789		.616
	Germany	Unemotional	5			.610		.334
Fanti et al. (2009)	N = 347; M age = 14.63;	General CU	24	None	.921		.616	
	50.7% male;	Callousness	11			.809		.528
	community sample;	Uncaring	8			.964		.454
	Greece	Unemotional	5			.725		.500
Gao and Zhang (2016)	N = 340; M age = 9.06;	General CU	16	2, 4, 8, 10, 5, 14, 6, 22	.735		.197	
	48.2% male;	Callousness	7			.661		.588
	community sample;	Uncaring	6			.788		.662
	United States	Unemotional	3			.514		.486
Kimonis et al. (2008)	N = 248; M age = 15.47;	General CU	22	2, 10	.735		.369	
	75.8% male; juvenile	Callousness	9			.737		.708
	justice involved	Uncaring	8			.841		.120
	sample; United States	Unemotional	5			.146		< .001
López-Romero et al. (2015)	N = 324; M age = 16.13;	General CU	24	None	.453	.521	.370	.130
	institutionalized	Callousness	11			.849		< .001
	sample; Spain	Uncaring	8			.193		.048
		Unemotional	5					
Mann et al. (2015)	N = 535; M age = 15.82;	General CU	24	None	.905		.807	
	community sample;	Callousness	11			.821		.036
	United States	Uncaring	8			.830		.030
		Unemotional	5			.833		.792
Paiva-Salisbury, Gill, and Stickle (2016)	N = 234; M ge = 15.38; 62.8% male; combined	General CU	22	2, 10	.882		.754	
	community and	Callousness	9			.737		.296
	juvenile justice	Uncaring	8			.863		.015
	involved sample; United States	Unemotional	5			.719		.619
M (SD)					.800 (.147)	.740 (.105)	.519 (.202)	.448 (.243)
95% CI					[.69, .92]	[.66, .82]	[.36, .67]	[.26, .63]
General CU						.853 (.055)		.274 (.271)
Callousness						[.81, .90]		[.07, .48]
Uncaring						.558 (.240)		.412 (.257)
Unemotional						[.37, .74]		[.21, .61]
Parent-Report ICU								
Gao and Zhang (2016)	N = 340; M age = 9.06;	General CU	19	2, 6, 8, 10, 21	.831		.478	
	48.2% male;	Callousness	7			.738		.556
	community sample;	Uncaring	8			.821		.504
	United States	Unemotional	4			.572		.353
Horan, Brown, Jones, and Aber (2015)	N = 355; M age = 15.09; 49.5% male;	General CU	24		.901		.749	
	community sample;	Callousness	11			.777		.418
	United States	Uncaring	8			.863		.124
		Unemotional	5			.773		.337
Moore et al. (2017)	N = 678; age range = 9–14; 48.4% male;	General CU	24		.926		.798	
		Callous/Uncaring	19			.918		.085
	community twin sample; United States	Unemotional	5			.854		.710

(*Continued*)

TABLE 1
(Continued)

	Sample Description	*Factor*	*No. of Items Included*	*Items Removed*	*Omega*	*Omega Subscale*	*OmegaH*	*OmegaHS*
Waller et al. (2015)	N = 540; age = 9.5; 62.8%	General CU	22	10, 23	.937		.822	
	male; high-risk	Callousness	10			.899		.315
	community sample;	Uncaring	7			.872		.012
	United States	Unemotional	5			.754		.401
M (*SD*)					.899 (.048)	.805 (.084)	.712 (.159)	.429 (.121)
95% CI					[.82, .97]	[.69, .97]	[.46, .97]	[.03, .66]
General CU						.852 (.027)		.213 (.258)
Callousness						[.78, .92]		[-43, .85]
Uncaring						.738 (.119)		.450 (.175)
Unemotional						[.55, .93]		[.17, .73]

Note: H = Hierarchical; HS = Hierarchical subscale; ICU = Inventory of Callous-Unemotional Traits; CU = callous-unemotional; CI = confidence interval.

bifactor models may be largely due to methodological artifacts and use of a total score that combines both positively and negatively worded items would include items that discriminate well across all levels of CU trait severity.

Thus, currently, there does not appear to be a strong justification for creating subscales from the ICU based on the findings from factor analyses. On the other hand, it does leave open the question as to how well unit-weighted total scores reflect a single latent construct of CU traits. That is, because the items that make up the total score reflect variance of both the general factor and the independent bifactors, it is unclear what proportion of variance in the total scale reflects these two sources of variance. Rodriguez, Reise, and Haviland (2016a) provided a compelling argument that even when confirmatory bifactor models fit the data well, this does not necessarily mean that unit-weighted total scores are uninterpretable and do not reflect substantial variance in the general overarching construct. More important, these authors provide a set of psychometric indices that can be derived from standardized factor loading matrices in confirmatory bifactor models and that can be used to determine how much of the variance in unit weighted total scores is due to common variance across dimensions (Rodriguez et al., 2016a). That is, they provide various statistics that can determine how well unit-weighted total scores reflect variance in the general factor as opposed to the bifactors and, as a result, how well these total scores can be interpreted as an indicator of this general latent construct, despite any multidimensionality reflected in the items.

In this article, we use these indices to provide a meta-analysis of the various bifactor tests of the ICU to determine how well a unit-weighted total score from this scale reflects the general factor and can be interpreted as a valid indicator of this single latent construct. We also include indices provided by Rodriguez, Reise, and Haviland (2016b) for testing how well unit-weighted subscale scores reflect variance specific to that subscale, as opposed to variance in the general latent trait of CU traits. However, these tests of the subscales from the ICU were clearly of secondary interest, given the lack of strong support for the importance of these subscales from past research.

METHODS

Study Inclusion

To obtain studies to be included in the meta-analysis, we conducted an exhaustive search for studies investigating the three-bifactor model of the ICU utilizing various electronic databases, reviewing the reference lists of published studies, and contacting researchers directly. Various search terms (e.g., *3-factor bifactor, Inventory of Callous Unemotional Traits, CU traits, validation, factor structure*, etc.) and combinations of these terms (e.g., *3-factor bifactor* and *CU traits*) were used when searching databases. Limiting the focus to bifactor models was necessary because the statistics used for this meta-analysis would be comparable across studies only using the same method for factor analysis. Further, this was the most common method for testing the factor structure of the ICU and, as result, provided the largest pool of studies using a common method for the meta-analyses. To maximize the number of studies in our analyses, we placed no restrictions on eligible participant populations, research designs, or geographical/cultural characteristics of samples. Likewise, we remained inclusive in terms of time period for publication of studies.

In addition, we included only published studies that found acceptable fit for the three-factor bifactor model. This was necessary because the statistics used in the meta-analysis are interpretable only when the factor model adequately accounts for the correlations among items in the sample (Rodriguez et al., 2016b). Studies in which the majority or at least half of these fit indices suggested acceptable fit for the three-factor bifactor model were included in the review. Acceptable fit was determined by the five most commonly reported fit indices: chi-square, χ^2/df, comparative fit index (Bentler, 1990), Tucker–Lewis index (Tucker & Lewis, 1973), root mean square error of approximation (Steiger, 1990). Specifically, small and nonsignificant chi-square, χ^2/df values less than 5 (West, Taylor, & Wu, 2012), comparative fit index and Tucker–Lewis index values greater than .90, and root mean square error of approximation values less than .10 indicated acceptable fit (Hu & Bentler, 1999). We included the study if the fit indices indicated that the three-bifactor model was acceptable, even if the study found better fit for a different solution (Ciucci et al., 2014; Gao & Zhang, 2016; Lopez-Romero, Gomez-Fraguela, & Romero, 2015; Mann, Briley, Tucker-Drob, & Harden, 2015). For example, Ciucci et al. (2014) found support for a hierarchical factor model that showed improved fit over the three-bifactor model; however, the three-bifactor model showed acceptable model fit on all fit indices. Also, whereas Mann et al. (2015) found support for a four-factor bifactor model consisting of a Callous, Careless, Uncaring, and Unemotional factors, they also found acceptable fit for the three-bifactor model.

We did, however, include studies that found support for *modified* three-factor bifactor models (e.g., elimination of items, two-factor bifactor model). For example, several studies eliminated items with low factor loadings (see Table 1 for number of items on each factor), and Moore et al. (2017) found support for a two-factor bifactor model. If studies reported factor loadings for multiple models (e.g., all 24 items retained and modified models with items removed), we included the model with the more complete version of the ICU, if it met acceptable model fit criteria (see the preceding). As noted in Table 1, seven effect sizes were based on all 24 items, whereas four were based on 22 items, and two were based on a further reduced item pool.

Our search resulted in a total of $k = 23$, but $k = 8$ of these studies did not indicate adequate fit for the bifactor model (Benesch, Görtz-Dorten, Breuer, & Döpfner, 2014; Colins, Andershed, Hawes, Bijttebier, & Pardini, 2016; Feilhauer et al., 2012; Hawes et al., 2014; Houghton et al., 2013; Kimonis, Branch, Hagman, Graham, & Miller, 2013; Kimonis et al., 2016; Willoughby et al., 2014). Of the remaining 15 studies, $k = 8$ did not include adequate information necessary to calculate model-based reliabilities. In turn, the authors of these eight studies were contacted via an e-mail that included a description of our study along with a form to be completed that requested the necessary information (e.g., standardized factor loadings) to compute model-based reliabilities. Of the eight authors contacted, five fulfilled our request. These procedures resulted in a sample of $k = 12$ studies with a combined total sample size of 6,008. Of these studies, $k = 9$ reported results for the self-report version of ICU ($n = 4{,}435$), and $k = 4$ reported on the parent-report version ($n = 1{,}913$); one study (Gao & Zhang, 2016) reported results for both parent and self-reports of the ICU. Of the 12 studies, four used a version of the ICU that was translated into a language other than English: Italian (Ciucci et al., 2014), German (Essau et al., 2006), Greek (Fanti et al., 2009), and Spanish (López-Romero et al., 2015).

Bifactor Statistical Indices

Using formulas provided by Rodriguez et al. (2016a, 2016b), we report four *omega* indices. First, we calculated the omega coefficient for the general CU factor, which is the "factor-analytic model-based reliability estimate of the proportion of variance in the unit-weighted total score attributable to all sources of common variance" (Rodriguez et al., 2016a, p. 224). The omega coefficient can be calculated as follows:

$$omega = \frac{(\sum \lambda genCU)^2 + (\sum \lambda groupCal)^2 + (\sum \lambda groupUnc)^2 + (\sum \lambda groupUne)^2}{(\sum \lambda genCU)^2 + (\sum \lambda groupCal)^2 + (\sum \lambda groupUnc)^2 + (\sum \lambda groupUne)^2 + \sum (1 - h^2)},$$

where $\lambda genCU$ represents the standardized factor loadings for each item on the general CU factor and $\lambda groupCal$, Unc, *and* Une are the factor loadings for the subscales on their respective factors. In addition, $1 - h^2$ represents the unique variance or error for each item. Thus, the numerator represents all of the common sources of unit weighted total score variance and the denominator is all the sources of variance including both the common and unique sources of variance (Rodriguez et al., 2016b).

Next, the omega hierarchical (omegaH) values for the CU total score was calculated. When data show an adequate fit with a bifactor structure, omegaH provides an estimate of the amount of *unique* total score variance that is attributable to the general factor (i.e., CU general factor) and, thereby, treats the subscale factors as error variance (Rodriguez et al., 2016b). Therefore, a larger omegaH is indicative that the general factor is the dominant source of systematic variance. OmegaH was computed as follows:

$$omegaH = \frac{\left(\sum \lambda genCU\right)^2}{\left(\sum \lambda genCU\right)^2 + \left(\sum \lambda groupCal\right)^2 + \left(\sum \lambda groupUnc\right)^2 + \left(\sum \lambda groupUne\right)^2 + \sum(1 - h^2)}.$$

Rodriguez et al. (2016b) recommended two ways to estimate the ability of the unit-weighted total score to reflect variance in the general factor from bifactor models. First, when omegaH is high, the total scores can be considered essentially unidimensional in the sense that the vast majority of reliable variance is attributable to the general factor. Second, the differences in the values of the omega and omegaH can be compared to each other. When there is not a substantial difference between omega (i.e., proportion of variance in unit-weighted total score that attributable to *all* source of common variance) and omegaH (i.e., proportion of variance in unit-weighted total score that attributable to individual differences in the general factor), this suggests that the general factor is the overwhelming determinant of the systematic variance underlying the unit-weighted total score.

Given our primary interest in the ability of the unit-weighted total score to reliably capture variance in the overarching CU factor from bifactor models of the ICU, omega and omegaH were the two indices of primary interest in the meta-analysis. However, as noted by Rodriguez et al. (2016b), analogous omega coefficients can be applied to the bifactors to compute model-based reliabilities of subscales derived from unit-weighted items. Thus, we calculated the omega subscales (omegaS) with the following formula presented for the Callousness subscale:

$$omega\_Cal = \frac{\left(\sum \lambda genCU\right)^2 + \left(\sum \lambda groupCal\right)^2}{\left(\sum \lambda genCU\right)^2 + \left(\sum \lambda groupCal\right)^2 + \sum(1 - h^2)}.$$

This was computed for each subscale (Callousness, Uncaring, and Unemotional). Further, omegaHS provides an estimate of the reliability of the subscale items after partialing out the general factor (i.e., CU factor). Again, one can interpret the overall level of omegaHS and can compare it to omegaS, such that when omegaHS is low relative to omegaS, much of the reliable variance of the subscale scores can be attributable to the general factor and not what is unique to bifactor. The following formula was used to compute the omegaHS:

$$OmegaHS\_Cal = \frac{\left(\sum \lambda groupCal\right)^2}{\left(\sum \lambda genCU\right)^2 + \left(\sum \lambda groupCal\right)^2 + \sum(1 - h^2)}.$$

RESULTS

The model-based reliabilities that were computed for each study are reported in Table 1 along with descriptions of the study sample, number of items included in the model, and the items excluded if applicable. The bifactor statistics of interest are reported separately for the self-report version of the ICU and the parent-report version. That is, the top portion of the table presents omega values for studies using the self-report version of the ICU (n = 9). The bottom portion presents these statistics for the model-based reliabilities for studies using the parent-report version (n = 4). We were unable to obtain information to compute model-based reliabilities from any studies using the teacher-report version of the ICU. Also, one study (i.e., Gao & Zhang, 2016) included both the parent- and self-report versions of the ICU, and these are included in both sections.

Self-Report ICU

With regard to the self-report ICU, the omega values for the general CU factor ranged from .453 to .921 (M = .800, SD = .147), suggesting strong reliability for the total score when considering all sources of variance. Thus, on average, 80% of the variance in the total ICU score is attributable to all "modeled" sources of common variance" (p. 140; Rodriguez et al., 2016b). However, the omegaH values for the general CU factor for the self-report version of the ICU ranged from .197 to .807 (M = .519, SD = .202), suggesting that on average 51.9% of the variance in the unit-weighted total scores were due to individual differences on the general CU factor in the bifactor models (Rodriguez et al., 2016b). Further, when comparing the mean omega (.800) to the mean omegaH (.519) for all studies, it suggests that on average the general factor was the substantial determinant of the systematic variance of the unit-weighted total score. Specifically, the majority of the reliable variance in total scores (.519/.800 = .649) is due to the CU general factor.

The omegaS values for the subscales ranged from .521 to .866 (M = .740, SD = .105) for the Callousness scale, .788 to .964 (M = .853, SD = .055) for the Uncaring scale, and .146 to .833 (M = .558, SD = .240) for the Unemotional scale. Thus, the reliable variance in the subscales tend to be acceptable for the Callousness and Uncaring subscales but low for the

Unemotional subscale. More important, the omegaHS values ranged from .036 to .708 (M = .448, SD = .243) for the Callousness scale, less than .001 to .662 (M = .274, SD = .271) for the Uncaring scale, and less than .001 to .792 (M = .412, SD = .257) for the Unemotional scale. These indices suggest that the variance in the unit-weighted subscales largely reflected variance from the general factor, with only 27.2% (Uncaring) to 44.8% (Callousness) of the reliable variance reflecting individual differences in the specific bifactor (partialing out variance associated with the general CU factor). Further, when comparing the omegaS to the omegaHS, individual differences on the respective bifactor were a substantial determinant of the unit-weighted subscale score for the Unemotional subscale only.

Parent-Report ICU

Similar estimates of the reliable variance for the studies finding an acceptable bifactor structure for parent-reported ICU items are shown at the bottom portion of Table 1. The omega values for the general CU factor ranged from .831 to .937 (M = .899, SD = .048), suggesting quite strong overall reliability for the unit-weighted total score considering all sources of variance. The omegaH values for the general CU factor for the parent-report version ranged from .478 to .822 (M = .712, SD = .159), suggesting that 71.2% of the variance in the unit-weighted total scores were due to individual differences on the general CU factor. Further, when comparing the mean omega (.899) to the mean omegaH (.712), it is apparent that the overwhelming determinant of the systematic variance underlying unit-weighted total scores using the parent-rated items was the general factor. The majority of the reliable variance in total scores (.712/.899 = .791) for the parent-reported ICU is due to the CU general factor.

For the subscales, the omegaS values ranged from .738 to .899 (M = .805, SD = .084) for the Callousness subscale, .821 to .872 (M = .852, SD = .027) for the Uncaring subscale, and .572 to .854 (M = .738, SD = .119) for the Unemotional subscale. These values generally support the overall reliability of the subscales. However, the omegaHS values ranged from .315 to .556 (M = .429, SD = .121) for the Callousness scale, .012 to .504 (M = .213, SD = .258) for the Uncaring scale, and .337 to .710 (M = .450, SD = .175) for the Unemotional scale. As was the case for the self-report ICU, these indices largely suggest that the variance in the unit-weighted subscales largely reflect variance from the general factor and not bifactor specific variance.

DISCUSSION

The ICU has become a widely used measure for assessing CU traits in research, and much of this research has used the total score as an overall measure of this construct, despite the relatively consistent finding of a three-bifactor structure across many samples. This practice has been justified by the lack of a clear theoretical model to explain the three factors and the failure to consistently find different correlates to these dimensions (Frick & Ray, 2015). In this meta-analyses, we provide one more reason to support this practice. That is, the unit-weighting (i.e., simple summing of item scores) of items to form a total score shows strong overall model reliability, and the majority of this reliable variance in the total score was determined by the general CU factor, when using model based reliability estimates from past bifactor tests of the ICU. As a result, this score seems to be a good indicator of the general latent trait of ICU, despite any multidimensionality suggested by the bifactor models (Rodriguez et al., 2016a).

These results also need to be interpreted in light of two IRT analyses of the ICU that have suggested that the emergence of subfactors from the ICU may partly be an artifact of different rates of item endorsement and subsequent differences in item difficulty, with positively worded items showing higher difficulty levels in IRT analyses and negatively worded items discriminating best at lower levels of the construct (Hawes et al., 2014; Ray et al., 2016). As a result, what emerges as dimension specific variance in bifactor models may reflect these different patterns of endorsement rather than variance from a theoretically important subdimension of CU traits. As a result, the variance in unit-weighted total score that are attributable to individual differences in the individual bifactors may also reflect in part this method variance in item difficulty. Rather than hurting the interpretation of the total score from the ICU, such method specific variance may reflect the use of items that discriminate well across all levels of CU trait severity.

These results provide additional support for relying on the total score when assessing CU traits using the ICU and the model-based statistics provide another reason to discourage use of unit-weighted subscale scores from the ICU. That is, with the exception of the Unemotional scale from the self-report version, the subscale scores largely reflect individual differences in the general CU factor and not individual differences in the specific bifactor, which makes it hard to unambiguously interpret these scores (Rodriguez et al., 2016b). Even the self-report Unemotional scale in which the reliable variance is heavily determined by variance due to the specific bifactor, the overall reliability falls below an acceptable level. These findings are consistent with the meta-analysis by Cardinale and Marsh (2017), which reported that the intercorrelations among the subscales of the ICU tend to be lowest with the Unemotional subscale and the internal consistency of the Unemotional scale is much more modest than the other subscales. Also, the fact that unit-weighted subscale scores are heavily determined by the general CU factor may explain the failure of past research to consistently find unique correlates to the subscales of the ICU across multiple samples (Frick & Ray, 2015) and again calls into question the use of these subscales in research.

These interpretations need to be made in the context of several limitations in the methodology used in this meta-analysis. The most important limitation is the few studies included in this meta-analysis, especially for the parent-report version of the ICU ($n = 4$). As a result, single studies could have a large influence on the findings. For example, the mean omegaH value for the self-report version of the ICU was .519 ($SD = .202$), and this was influenced by one very low value of .197 reported by Gao and Zhang (2016). Important to note, this study was the only study to eliminate more than two items from the calculation of the total score.

A second limitation of note is that the model reliability indices provide an overall estimate only of the reliable variance in the ICU items and how this variance can be partitioned into variance determined by the general factor and variance determined by the bifactors. These data do not address whether the three-factor bifactor model is the most appropriate way to describe the structure of the ICU items. It is important to note that of the 23 studies testing the bifactor structure and considered for inclusion, eight did not find adequate fit for this model. Also, we found four studies that directly compared the three-bifactor model to another factor model and found better fit for a different solution. However, it is also important to note that no other factor model has garnered consistent support in past work, although they consistently support a general CU dimension, which would again support the use of a total score and caution against using subscales based on any specific factor model (Frick & Ray, 2015).

Also, our results do not provide a way of determining if there are specific weaknesses in the item pool of the ICU. That is, it is still possible that there are weaknesses in the ICU item content that influences its validity for certain purposes. For example, one criticism of the ICU is that it focuses on only one dimension of psychopathic traits and, as a result, it is not an adequate measure of the construct of psychopathy (Salekin, 2016). Also, some studies have recommended the use of fewer items to form an abbreviated total score from the ICU (Hawes et al., 2014; Ray et al., 2016), and this could reduce the overall reliable variance in the total score due to the smaller number of items (Rodriguez et al., 2016b). Similarly, our results do not address the concern that certain item domains (e.g., the unemotional items) may not be appropriate for the overall construct of CU traits (Hawes et al., 2014; Kimonis et al., 2016).

In short, the model reliability indices provided in this meta-analysis indicate only how well the unit-weighted total score reflects the general CU construct, as it is measured by the specific item pool of ICU. It does not address the adequacy of the total score if a very different item pool is used. For example, some authors have recommended that, given some of the limitations in the unemotional items (e.g., low internal consistency, modest correlations with the other subscales, low correlations with aggression and antisocial behavior), total scores from the ICU should be determined eliminating these items (Cardinale & Marsh, 2017; Hawes et al., 2014). However, these findings from past research on the Unemotional subscale support our argument that the unemotional items alone are not good indicators of the construct of CU traits but they do not address the question of whether the variance from these items contribute to the overarching construct of CU traits in important ways. To illustrate this point, Cardinale and Marsh's (2017) meta-analysis reported that the Unemotional subscale was only modestly correlated with proactive aggression (pooled $r = .06$) across 14 studies but was more highly associated with empathy (pooled $r = -.22$) across 12 studies. Thus, the limited association with aggression would call into question this subscale, if it was considered as an indicator of CU traits alone, given that an important aspect of CU traits is its association with proactive aggression (Frick & Ray, 2015). However, the correlation with empathy could suggest that the variance that these items contribute to the total score may be important for defining a construct that predicts important aspects of cognitive and emotional functioning that is independent of aggression and antisocial behavior or that helps to designate subgroups of antisocial individuals with important cognitive and emotional differences, both of which are critical for the usefulness of the construct of CU traits (Frick & Ray, 2015). Although the effects of eliminating the unemotional items from the total score has not been directly tested in many studies and should be the focus of future research, the findings from Colins et al. (2016) support the concern that their elimination may reduce the theoretical and clinical importance of the construct. Specifically, in a sample of detained adolescent girls, the correlations between the total ICU score that included the unemotional items and a version with items eliminated did not show very different correlations with aggression ($r = .43$ vs. $r = .41$, $p < .01$, respectively), but the version with the unemotional items showed stronger associations with the personality dimension of Conscientiousness ($r = -.39$ vs. $r = -.26$, $p < .01$). Thus, including the unemotional items did not reduce the correlations of the total score with a measure of aggression, but their elimination reduced its correlation with a measure of prosocial behaviors and attitudes that was separate from aggression. In summary, the results of this meta-analysis address only the sources of variance in unit-weighted total scores with the current item set of the ICU and do not address the issue of whether changes in the item set are warranted.

It is also important to note that, as in any meta-analysis, we had to make decisions as to what studies were similar enough to be included in our calculation of effects. For example, we included studies that removed certain items (e.g., Items 2 and 10), and we included studies in which the three-factor bifactor structure obtained only adequate fit when post hoc model specifications (e.g., correlated residuals) were made. These decisions highlight the fact that our results do not address the adequacy of the three-factor bifactor model over other potential factor structures. Also, the small number of studies available for inclusion prevented us from systematically examining if these decisions influenced our findings in substantive ways. Thus, future research should test for moderating effects of such methodological features, as the research base expands and more tests of the factor structure of the ICU are conducted.

Finally, the current study focused on the reliability estimates solely for the purpose of computing unit-weighted total scores based on summing scores on the individual items. We felt that this was an important question, given that this is by far the most common way that scores from the ICU have been used in past research to assess CU traits (Frick & Ray, 2015). However, using item or subscale scores to estimate the latent CU trait in statistical models (e.g., structural equation modeling) may provide a more appropriate method for some research purposes (Grice, 2001; Rodriguez et al., 2016a).

In summary, a significant amount of past research suggests that CU traits seem to be important for designating a subgroup of antisocial youth who show a more severe and impairing pattern of antisocial behavior and who show a number of distinct correlates to their behavior problems that would seem to implicate distinct causal processes leading to their behavior problems, relative to antisocial youth who do not show elevated CU traits (Blair et al., 2014; Frick et al., 2014b; Herpers et al., 2012). Thus, as noted in the introduction, this construct seems to have substantial support for its clinical and etiological validity. The ICU provides one way of measuring this important construct, and our results suggest that, if the ICU items capture the construct of CU traits in a way that fits with the research question of interest, the unit-weighted total score shows adequate reliable variance that is largely determined by a general factor shared by all of the items on the test. In contrast, the reliable variance in the unit-weighted subscales seems to be heavily determined by this general factor as well, which makes interpretations of these scores more ambiguous and makes finding consistent, unique, and theoretically important correlates to them unlikely.

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Washington, DC: American Psychiatric Publishing.

Benesch, C., Görtz-Dorten, A., Breuer, D., & Döpfner, M. (2014). Assessment of callous-unemotional traits in 6 to 12 year-old children with oppositional defiant disorder/conduct disorder by parent ratings. *Journal of Psychopathology and Behavioral Assessment*, *36*(4), 519–529. doi: 10.1007/s10862-014-9420-7

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. doi: 10.1037/0033-2909.107.2.238

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

*Byrd, A. L., Kahn, R. E., & Pardini, D. A. (2013). A validation of the Inventory of Callous-Unemotional Traits in a community sample of young adult males. *Journal of Psychopathology and Behavioral Assessment*, *35*, 20–34. doi: 10.1007/s10862-012-9315-4

Cardinale, E. M., & Marsh, A. A. (2017). The reliability and validity of the Inventory of Callous-Unemotional Traits: A meta-analytic review. *Assessment*. doi: 10.1177/1073191117747392

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

Colins, O. F., Andershed, H., Hawes, S. W., Bijttebier, P., & Pardini, D. A. (2016). Psychometric properties of the original and short form of the Inventory of Callous-Unemotional Traits in detained female adolescents. *Child Psychiatry & Human Development*, *47*(5), 679–690. doi: 10.1007/s10578-015-0601-8

*Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, *13*, 454–469. doi: 10.1177/1073191106287354

Ezpeleta, L., Osa, N. D. L., 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. doi: 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. doi: 10.1007/s10862-008-9111-3

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. doi: 10.1016/j.ijlp.2012.04.002

Frick, P. J., & Ray, J. V. (2015). Evaluating callous-unemotional traits as a personality construct. *Journal of Personality*, *83*, 710–722. doi: 10.1111/jopy.12114

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

Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014b). 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. doi: 10.1111/jcpp.12152

*Gao, Y., & Zhang, W. (2016). Confirmatory factor analyses of self-and parent-report inventory of callous-unemotional traits in 8-to 10-year-olds. *Journal of Psychopathology and Behavioral Assessment*, *38*, 331–340. doi: 10.1007/s10862-015-9527-5

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430–450. doi:10.1037/1082-989X.6.4.430

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

*References marked with asterisk indicate studies included in the meta-analysis.

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. doi: 10.1037/a0034718

Herpers, P. C., Rommelse, N. N., Bons, D. M., Buitelaar, J. K., & Scheepers, F. E. (2012). Callous–Unemotional traits as a cross-disorders construct. *Social Psychiatry and Psychiatric Epidemiology*, *47*, 2045–2064. doi: 10.1007/s00127-012-0513-x

*Horan, J. M., Brown, J. L., Jones, S. M., & Aber, J. L. (2015). Assessing invariance across sex and race/ethnicity in measures of youth psychopathic characteristics. *Psychological Assessment*, *27*, 657. doi: 10.1037/pas0000043

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*(2), 215–222. doi: 10.1007/s10862-012-9324-3

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

Kimonis, E. R., Branch, J., Hagman, B., Graham, N., & Miller, C. (2013). The psychometric properties of the Inventory of Callous–Unemotional Traits in an undergraduate sample. *Psychological Assessment*, *25*(1), 84. doi: 10.1037/a0029024

Kimonis, E. R., Fanti, K. A., Anastassiou-Hadjicharalambous, X., Mertan, B., Goulter, N., & Katsimicha, E. (2016). Can callous-unemotional traits be reliably measured in preschoolers? *Journal of Abnormal Child Psychology*, *44*, 625–638. doi: 10.1007/s10802-015-0075-y

Kimonis, E. R., Fanti, K. A., Frick, P. J., Moffitt, T. E., Essau, C., Bijttebier, P., & Marsee, M. A. (2015). Using self-reported callous-unemotional traits to cross-nationally assess the DSM-5 'With Limited Prosocial Emotions' specifier. *Journal of Child Psychology and Psychiatry*, *56*, 1249–1261. doi: 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. doi: 10.1016/j.ijlp.2008.04.002

Kotler, J. S., & McMahon, R. J. (2005). Child psychopathy: Theories, measurement, and relations with the development and persistence of conduct problems. *Clinical Child and Family Psychology Review*, *8*, 291–325. doi: 10.1007/s10567-005-8810-5

*López-Romero, L., Gómez-Fraguela, J. A., & Romero, E. (2015). Assessing callous-unemotional traits in a Spanish sample of institutionalized youths: The inventory of callous-unemotional traits. *Journal of Psychopathology and Behavioral Assessment*, *37*, 392–406. doi: 10.1007/s10862-014-9469-3

*Mann, F. D., Briley, D. A., Tucker-Drob, E. M., & Harden, K. P. (2015). A behavioral genetic analysis of callous-unemotional traits and Big Five personality in adolescence. *Journal of Abnormal Psychology*, *124*, 982. doi: 10.1037/abn0000099

*Moore, A. A., Carney, D., Moroney, E., Machlin, L., Towbin, K. E., Brotman, M. A., … Hettema, J. M. (2017). The Inventory of Callous-Unemotional Traits (ICU) in children: Reliability and heritability. *Behavior Genetics*, *47*, 141–151. doi: 10.1007/s10519-016-9831-1

*Paiva-Salisbury, M. L., Gill, A. D., & Stickle, T. R. (2016). Isolating trait and method variance in the measurement of callous and unemotional traits. *Assessment*, *24*, 763–771. doi: 10.1177/1073191115624546

Ray, J. V., Frick, P. J., Thornton, L. C., Steinberg, L., & Cauffman, E. (2016). Positive and negative item wording and its influence on the assessment of callous-unemotional traits. *Psychological Assessment*, *28*, 394. doi: 10.1037/pas0000183

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*, 223–237. Doi: 10.1037/met0000045

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137–150. doi: 10.1037/met0000045

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. doi: 10.1177/1073191109344153

Salekin, R. T. (2016). Psychopathy in childhood: Toward better informing the DSM–5 and ICD-11 conduct disorder specifiers. *Personality Disorders: Theory, Research, and Treatment*, *7*, 180–191. doi: 10.1037/per0000150

Sharp, C., & Kine, S. (2008). The assessment of juvenile psychopathy: Strengths and weaknesses of currently used questionnaire measures. *Child and Adolescent Mental Health*, *13*, 85–95. doi: 10.1111/j.1475-3588.2008.00483.x

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180. doi: 10.1207/s15327906mbr2502_4

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. doi: 10.1007/BF02291170

*Waller, R., Wright, A. G., Shaw, D. S., Gardner, F., Dishion, T. J., Wilson, M. N., & Hyde, L. W. (2015). Factor structure and construct validity of the parent-reported Inventory of Callous-Unemotional Traits among high-risk 9-year-olds. *Assessment*, *22*, 561–580. doi: 10.1177/1073191114556101

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling, In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. New York, NY: Guilford Press.

Willoughby, M. T., Mills-Koonce, W. R., Gottfredson, N. C., & Wagner, N. J. (2014). Measuring callous unemotional behaviors in early childhood: Factor structure and the prediction of stable aggression in middle childhood. *Journal of Psychopathology and Behavioral Assessment*, *36*(1), 30–42. doi: 10.1007/s10862-013-9379-9