Running Head: ICU FACTOR STRUCTURE IN A MULTINATIONAL SAMPLE

Inventory of Callous-Unemotional Traits (ICU) Factor Structure and Measurement Invariance in an Adolescent Multinational Sample

Emily C. Kemp, M.A.[a], James V. Ray, Ph.D.[b], Paul J. Frick, Ph.D.[a,c],

Emily L. Robertson, M.A.[a], Kostas A. Fanti, Ph.D.[d], Cecilia A. Essau, Ph.D.[e],

Andrea Baroncelli, Ph.D.[f], Enrica Ciucci, Ph.D. [g], Patricia Bijttebier, Ph.D.[h,i]

Accepted for publication in "Journal of Clinical Child and Adolescent Psychology"

Affiliations: [a] Department of Psychology, Louisiana State University, [b] Department of Criminal Justice, University of Central Florida, [c] Institute for Learning Sciences and Teacher Education, Australian Catholic University, [d] Department of Psychology, University of Cyprus, [e] Department of Psychology, Roehampton University, [g] Department of Education, Languages, Intercultures, Literatures, and Psychology, University of Florence, [h] School Psychology and Development in Context, KU Leuven, [i] KU Leuven Child and Youth Institute

Address correspondence to: Emily C. Kemp, M.A., Department of Psychology, Louisiana State University, 236 Audubon Hall Baton Rouge, LA 70803; ekemp4@lsu.edu

Disclosures: The authors have no financial disclosures or conflicts of interest to report.

Acknowledgements: We would like to thank our gracious collaborators who collected and provided the data to make this multinational ICU database possible.

Abstract

Objective: The Inventory of Callous-Unemotional Traits (ICU) is a widely used, comprehensive measure of callous-unemotional (CU) traits. While the ICU total score is used frequently in research, the factor structure of the scale remains highly debated. Inconsistencies in past factor structure research appear to be largely due to the use of small non-representative samples and failure to control for method variance (i.e., item wording direction). **Method:** The current study used a multitrait-multimethod (MTMM) confirmatory factor analytic (CFA) approach that considers both trait and method variance to test the factor structure of a 22-item version of the self-report ICU in a multinational community sample of 4,683 adolescents (ages 11-17). **Results:** Results showed that a hierarchical four-factor model (i.e., one overarching CU factor, four latent trait factors) that controlled for method variance (i.e., by allowing residuals from positively worded items to covary) provided the best fit (χ^2=2797.307, *df*=160, RMSEA=.059, CFI=.922, TLI=.888, SRMR=.045). **Conclusions:** Importantly, this factor structure is consistent with how the ICU was developed and corresponds to the four symptoms of Limited Prosocial Emotions (LPE) specifier in the DSM-5 criteria for Conduct Disorder (CD). In addition, measurement invariance of this factor structure across age (i.e., younger versus older adolescents) and sex were supported. As a result of these findings, mean differences in ICU total score across age and sex can be interpreted as reflecting true variations in these traits. Further, we documented that boys generally scored higher than girls on the ICU, and this sex difference was larger in later adolescence.

Keywords: callous-unemotional (CU) traits, Inventory of Callous-Unemotional Traits (ICU), confirmatory factor analysis (CFA), measurement invariance, method variance

Inventory of Callous-Unemotional Traits (ICU) Factor Structure and Measurement Invariance in an Adolescent Multinational Sample

Latest versions of major diagnostic systems, including the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (DSM-5; American Psychiatric Association, 2013) and the *International Classification of Disease, 11th Revision* (ICD-11; World Health Organization, 2018), have added the specifier "with Limited Prosocial Emotions (LPE)" to identify youth with disruptive and serious behavior problems who also show elevated callous-unemotional (CU) traits. CU traits are defined by lack of remorse or guilt, a callous lack of empathy, shallow or constricted affect, and lack of motivation to perform well in important activities (e.g., academics; Frick & Ray, 2015). This inclusion in diagnostic classification was based on substantial research reporting associations between CU traits and particularly severe and stable forms of antisocial behavior (Frick et al., 2014) that were not captured well by other indices of severity, such as number of conduct problems, co-morbid diagnoses, and age of onset of conduct problems (McMahon et al., 2010). Further, CU traits seem to designate an etiologically distinct group of children and adolescents with severe behavior problems who display very different emotional deficits underlying their conduct problems (Blair et al. 2014; Frick et al., 2014; Frick & Kemp, 2021).

Early measures of CU traits in children and adolescents that were used in research were largely subscales taken from broader measures of psychopathic traits (e.g., see Antisocial Process Screening Device; APSD; Frick & Hare, 2001; Child Psychopathy Scale; CPS; Lynam, 1997; Youth Psychopathic Traits Inventory; YPI; Andershed et al., 2002). As a result, these measures had a limited number of items assessing CU traits, leading to significant psychometric limitations (e.g., low internal consistency; see Poythress et al., 2006). To overcome these

problems, Frick (2004) developed the Inventory of Callous-Unemotional Traits (ICU) by taking four items from the CU subscale of the APSD that most consistently loaded onto a general factor of CU traits across samples (Frick et al., 2000) and developing three positively worded and three negatively worded items related to and including the original core item from the APSD. The new scale also expanded response options to include four, rather than three, which increased the variability in scores and did not allow for a central response. Thus, the ICU was constructed to include 24 items with a greatly expanded range of potential scores that includes equal numbers of positively and negatively worded items. Importantly, the four core items that formed the ICU (i.e., “I feel bad or guilty when I do something wrong”, “I am concerned about the feelings of other”, “I do not show my emotions to others”, “I care about how well I do at school or work”) correspond to the four symptoms of the DSM-5 criteria for LPE (American Psychiatric Association, 2013).

The ICU has separate forms for completion by parents, teachers, and youth self-report; it has separate forms for children under the age of five (i.e., preschool version); and it has been translated in over 28 languages (Ray & Frick, 2018). It has been used in over 300 peer-reviewed studies that have generally provided strong support for its reliability and construct validity, with the vast majority of these studies using the self-report version (Cardinale & Marsh, 2020; Deng et al., 2019; Ray & Frick, 2018). For example, in their meta-analysis of the reliability of the ICU, Deng and colleagues (2019) reported on 113 estimates for the self-report version, but only 23 and 4 estimates for the parent- and teacher-report versions, respectively. Further, Matlasz and colleagues (2021) reported that, when compared to informant-report versions, the ICU self-report was most consistently associated with relevant clinical validators across different developmental

stages (i.e., 3rd, 5th, and 8th grades) and was the only version to show strong validity in the oldest age group (i.e., 8th grade).

Thus, there is substantial research supporting the reliability and validity of the ICU self-report version for use in adolescent samples. However, the primary concern that has been raised about the ICU is its factor structure (Hawes et al., 2014). Ray and Frick (2018) provided a review of 23 factor analyses of the ICU using either the self or parent report. Of most concern, none of these factor analyses provided support for the theoretical model that guided the development of the ICU (i.e., four item clusters loading onto an overarching dimension; see for example, Kimonis et al., 2008). Instead, the most common best-fitting factor structure was a three-factor model that includes *callousness* (i.e., deficient empathy and remorse), *uncaring* (i.e., limited concern about behavioral performance and others' feelings), and *unemotional* (i.e., restricted or deficient affect) subdimensions (see for examples, Byrd et al., 2013; Essau et al., 2006; Kimonis et al., 2008; Roose et al., 2010). However, this three-factor model is not found to provide adequate fit in many samples (see for examples, Allen et al., 2020; Benesch et al., 2014; Feilhauer et al., 2012; Hawes et al., 2014; Houghton et al., 2013; Kimonis et al., 2016; López-Romero et al., 2015; Thøgersen et al., 2020; Wang et al., 2017; Willoughby et al., 2014; Zhang et al., 2019). When this factor structure does obtain adequate fit, it is usually only after the use of post-hoc modifications that do not replicate across samples (Hawes et al., 2014). Further, these three empirical subfactors or dimensions have not shown consistent correlates or been integrated into any theoretical model to explain the structure of CU traits (Frick & Ray, 2015; Ray & Frick, 2018).

With this, a critical issue that needs to be addressed to guide the use of the ICU as a measure of CU traits and an indicator of the LPE specifier is to explain these inconsistencies in

factor structure. We propose two primary explanations. First, studies conducting factor analyses have varied widely in sample size and type. For example, studies by Benesch and colleagues (2014) and Thøgersen and colleagues (2020) conducted factor analyses in relatively small samples of clinically referred youth (*N*s = 131 and 160, respectively) and found varying support for the three-factor model. Specifically, Benesch et al. (2014) reported a best-fitting three-factor structure that failed to provide adequate model fit, while Thøgersen et al. (2020) reported a best-fitting two-factor structure that provided marginal-to-good model fit with a shortened form of the ICU (i.e., 12-item). Additionally, studies in relatively small samples of juvenile offenders and/or institutionalized adolescents report varying support for the three-factor model (e.g., *N* = 324; López-Romero et al., 2015; *N* = 221; Pechorro et al., 2016). This is important because small samples in factor analyses can result in a model fit that is highly influenced by very few scores, leading to difficulty in replicating models across samples. However, problems in sample size and composition cannot be the sole explanation for the unstable factor structure across samples because some studies that have used large, representative samples have also not found consistent evidence for the proposed factor structure of the ICU (e.g., Carvalho et al., 2017; Ciucci et al., 2014; Pechorro et al., 2019; Roose et al., 2010; Ueno et al., 2019; Willoughby et al., 2014).

A more likely explanation for the inconsistent factor structure is the failure to consider the effects of method variance. As noted above, the ICU was designed to have equal numbers of positively and negatively worded items. However, item wording seems to have a significant effect on the range of responses and, as a result, item difficulty. Specifically, Ray and colleagues (2016) investigated the item functioning of the ICU in a sample of over 1,000 high-risk (i.e., justice-involved) adolescents and reported that positively worded items (i.e., in the callous-unemotional direction) that indicate higher levels of CU traits were much less likely to be rated

with higher response categories (i.e., “very true” and “definitely true”), whereas negatively worded items (i.e., reversal items written in the prosocial direction) that indicate lower levels of CU traits were much more likely to be rated with lower response categories (i.e., “not at all true” or “somewhat true”). That is, participants were more likely to report an absence of prosocial traits rather than the presence of CU traits (Ray et al., 2016). As a result, Ray et al. (2016) demonstrated that positively worded items showed higher levels of difficulty in item response theory (IRT) analyses, such that these items discriminated best at higher levels of CU traits and were more highly correlated with more severe behavioral manifestations of CU traits (e.g., measures of antisocial, aggressive behavior). In contrast, negatively worded items showed lower levels of difficulty and discriminated best at lower levels of CU traits.

Thus, the use of both positively and negatively worded items is important for distinguishing well at both high and low levels of CU traits. However, item wording direction could lead to significant method variance that could influence the factor structure of the scale. To further illustrate this, the seven-item *callousness* subscale consists mostly of (i.e., all but one) positively worded items, whereas the eight-item *uncaring* subscale consists exclusively of negatively worded items. These patterns of covariation may thusly be more related to different item endorsement patterns, rather than to theoretically meaningful clusters of items. Further, given that these items account for the vast majority of items and the most variance in ICU total score, they could influence the structure of the remaining items. Several recent studies have begun to use factor analytic methods that consider the potential influence of item wording and have tested the factor structure of the ICU after controlling for item wording direction. The findings from these studies have provided much more consistent support for the original structure of the scale. Specifically, two recent studies in large samples of German 9th grade

students (*N* = 3,878; Kliem et al., 2020) and Greek-Cypriot high school students (i.e., grades 7-9; *N* = 1,536; Koutsogiorgi et al., 2020) found that controlling for method variance improved the overall fit of models testing ICU factor structure. Further, each of these studies provided support for a model with a general CU factor and the four subfactors corresponding with original item clusters.

Importantly, Kliem and colleagues (2020) also reported finding strict measurement invariance across sex and ethnicity for its best fitting four-factor model. Such tests are important for use of the ICU across different groups to ensure that the scale measures the construct of CU traits similarly across groups, which allows for interpretation of absolute levels of these traits. That is, research has consistently shown that boys tend to score higher on CU traits than girls across multiple samples (see Cardinale & Marsh, 2020; Fragkaki et al., 2016; Kliem et al., 2020; Pechorro et al., 2019; Pihet et al., 2015; Ueno et al., 2019). However, until measurement invariance is established, it is not clear if these differences are due to true differences in how CU traits are expressed across sex or due to differences in how the items are measuring the traits across these groups. Further, establishing measurement invariance is particularly critical if one wishes to establish normative cutoff scores for the ICU to aid in making diagnostic decisions (Kemp et al., 2021).

Based on these findings, the current study attempted to replicate recent factor analyses that supported the theorized ICU factor structure when item wording is controlled. That is, we used CFA to compare the fit of a one-factor model to the fit of the three-factor model that has been found in a large number of past factor analyses and then compared the fit of this model to one that includes all four item clusters that led to scale development and that correspond to LPE DSM-5 criteria. More importantly, using a multitrait-multimethod (MTMM) confirmatory

factor analytic (CFA) approach that considers both trait and method variance, we compared the fit of each of the models that did and did not control for method variance related to item wording. To avoid problems related to using a small select sample, we combined several large community samples used in previous research, so that our tests used a large multinational sample (N = 4,683) of community adolescents ages 11-17. Further, we then tested whether the best-fitting model was invariant across boys and girls and between younger (i.e., ages 11-14) and older (i.e., ages 15-17) adolescents, which was largely consistent with the delineation of youth age groups by Deng et al. (2019). Finally, once establishing measurement invariance, we then tested for differences in mean scores across these different groups to provide guidance for whether normative cutoff scores should be based on sex- and age-specific norms.

Methods

Participants

The current study utilized a combined dataset that was pooled from five previously published studies (Baroncelli et al., 2018; Ciucci et al., 2014; Essau et al., 2006; Fanti et al., 2013; Roose et al., 2010). Authors were approached to provide their data for secondary data analysis if they collected the ICU on a large and representative sample of adolescents (i.e., the sample was not selected based on risk, placement, or clinic referral). This led to a pooled sample of 4,683 adolescents ages 11-17 years (M = 13.98; SD = 1.66) that was 49.3% (n = 2,308) female. Supplementary Table 1 describes the demographic characteristics separated by subsample. The majority of the sample was living in Cyprus and completed the Greek translation of the ICU (n = 1,919; 41%), followed by 26% living in Germany and completing the German translation (n = 1,232), 25% living in Italy and completing the Italian translation (n = 1,158; 24.7%), and 8% living in Belgium and completing the Dutch translation (n = 374). Across

samples that reported on parental education, a majority of participants' parents indicated they had a high school education or higher (see Baroncelli et al., 2018; Ciucci et al., 2014; Fanti et al., 2013; Roose et al., 2010 for more details).

Measures

Inventory of Callous-Unemotional Traits (ICU)

Callous-unemotional (CU) traits were assessed across all study samples using the 24-item Inventory of Callous-Unemotional Traits (ICU) self-report youth version. Items were rated with a 4-point Likert-type scale, and response options ranged from 0 (Not at all true) to 3 (Definitely true). Items 2 and 10 (i.e., "what I think is 'right' and 'wrong' is different from what other people think"; "I do not let my feelings control me," respectively) showed the weakest item-total correlations (i.e., ITCs = .21 and .19, respectively), whereas all other ITCs were .25-.59, with a mean of .43. This is consistent with past studies using the ICU (e.g., Ciucci et al., 2014; Kimonis et al., 2008). As a result, these two items were not included in the main study analyses. The overall internal item consistency of the resulting 22-item scale was acceptable ($\alpha = .79$). Subscales representing the four latent trait factors (i.e., *limited concern about performance, limited remorse, callous-lack of empathy, and restricted affect*) resulted in McDonald's Omega (ω) reliabilities ranging from .64-.75 and were positively correlated with one another (Pearson's rs = .17-.53, ps < .01).

Analytic Plan

All factor analyses were conducted using Mplus software (Muthén & Muthén, 1998-2011). Consistent with the primary goals of the study, factor analyses proceeded in three stages. First, a series of confirmatory factor analytic (CFA) models were conducted based on prior research and theory. The CFA were conducted to test each model, both with and without

controlling for method variance. Specifically, we first estimated a unidimensional model in which all 22 items loaded onto a single factor. In the second model, we controlled for method variance by allowing residuals from same method items (i.e., all positively worded items) to covary. We chose to model positively worded items, as opposed to negatively worded, to reduce model complexity given that deleted items (i.e., 2 and 10) are both positively worded items and, as a result, using positively worded items as the method factor resulted in fewer parameters estimated. The third model was the empirically derived three-factor model that included one general CU factor and three trait factors, (i.e., *callousness, uncaring*, and *unemotional* subdimensions), without accounting for method variance. The fourth model was an MTMM three-factor model that accounted for method variance (i.e., by allowing positively worded item residuals to covary). The fifth model was a four-factor model based on how the ICU was developed, which included one general CU factor and four trait factors (i.e., *limited concern about behavioral performance, lack of remorse, callous-lack of empathy,* and *restricted affect*), but that did not account for method variance. The final model was an MTMM four-factor model that accounted for method variance by allowing positively worded item residuals to covary. To determine the best-fitting model, we used the chi-square statistic and associated probability, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Standardized Root Mean Square Residual (SRMR), and Root Mean Square Error of Approximation (RMSEA). We used standard criteria to determine the goodness of fit of each model: CFI $\geq$ 0.95, TLI $\geq$ 0.95, SRMR $\leq$ 0.08, and RMSEA $\leq$ 0.06 were indicative of good model fit and were considered acceptable when CFI $\geq$ 0.90, TLI $\geq$ 0.90, SRMR $\leq$ 0.10, and RMSEA $\leq$ 0.08 (Cheung & Rensvold, 2002; van de Schoot et al., 2012).

In the second stage, we tested measurement invariance of the best-fitting factor structure across sex and age (i.e., 11-14 vs. 15-17 years). This choice of age groups led to a cut-off that resulted in relatively equal numbers in each age grouping. For each of these variables, we followed conventional steps for testing different levels of measurement invariance (e.g., Grimm et al., 2016). Each step of testing measurement invariance assumes that the level of measurement invariance in the previous step was established. We first established configural invariance by testing whether the factor structure was similar across the groups. This was done by specifying a multiple group CFA across groups (i.e., boys and girls, younger and older adolescents) to confirm that the same latent factor structure emerged across groups. The next step was to establish weak factorial invariance. In weak factorial variance, the factor loadings are set to be equal across groups. The third step tested for strong factorial invariance. In this step, the item thresholds are also held equal across groups. Finally, strict factorial invariance was established by also constraining item residuals to be equal across groups, which is a two-step process. First, the item residual variance is freed while the factor loadings and item thresholds are constrained to be equal across groups. Second, the item residual variances are constrained to be equal across groups. At each step, we used multiple fit indices to determine if measurement invariance was established (Cheung & Rensvold, 2002; Liu et al., 2017; Marsh et al., 2010). We also considered comparative fit indices. Specifically, we observed the change in Comparative Fit Indices (ΔCFI) along with the change in Root Mean Squared Error of Approximation (ΔRMSEA), Standardized Root Mean Square Residual (ΔSRMR), and change in Tucker-Lewis Index (ΔTLI).

Finally[1], once measurement invariance of the best-fitting structure of ICU items is established across age and sex, mean differences in total scores can be considered to reflect

[1] Tests of measurement invariance across ICU language were not included in the main analyses due to widely varying sample sizes. However, exploratory analyses were conducted using two different approaches to address

meaningful variations in the construct across groups. Thus, the final step in our analyses was to assess for differences in ICU scores across age and sex. To do this, univariate tests of analysis of variance (ANOVA) were conducted using IBM SPSS Statistics for Mac, Version 27.0 to test for main effects of sex and age and an interaction effect of sex and age (i.e., using factorial or two-way ANOVA) on ICU total scores.

Results

Factor Analysis of ICU

The fit indices for the four CFA models are presented in Table 1. The same models and resulting fit indices are reported for the full-scale ICU (i.e., including items 2 and 10) in Supplementary Table 2. For the first two unidimensional models, only the SRMR met for acceptable-to-good fit, while all other fit indices failed to meet acceptable criteria. For the three-factor model, the SRMR met criteria for good model fit, while the RMSEA suggested acceptable model fit; however, both the CFI and TLI did not meet criteria for acceptable model fit. For the MTMM three-factor model that accounted for method variance, all model fit indices improved over the three-factor model. For the four-factor model, only the SRMR met for acceptable model fit. Finally, for the MTMM four-factor model that accounted for method variance, all model fit indices improved from that of the four-factor and MTMM three-factor models. Further, for this final model, the RMSEA, SRMR, and CFI all suggested acceptable-to-good model fit. The TLI was the only index of model fit that did not quite meet the threshold of acceptable fit. This may be due to model complexity, as the TLI tends to decrease as the complexity (e.g., number of

sample size variability. The first was to randomly sample 400 youth from each subsample to be of approximately equal weight as the smallest (i.e., the Dutch) subsample. The second approach was to test invariance between the Greek subsample, which was approximately half of the entire sample, to all others. The results of these analyses can be found in Supplementary Table 5. As noted in this table, consistent support across fit indices was found for configural and weak factorial models, but support was not consistent for strong and strict factorial invariance.

items, number of factors) of the model increases (e.g., Cheung & Rensvold, 2002). Thus, the MTMM four-factor model was identified as the best-fitting model, and this was observed to also be true for the 24-item ICU (Supplementary Table 2). The standardized factor loadings for the 22-item MTMM four-factor model can be seen in Figure 1. All items loaded significantly and positively onto their respective trait factor (i.e., *limited concern about behavioral performance, lack of remorse*, *callous-lack of empathy*, and *restricted affect*). Further, all latent trait subfactors loaded positively onto the general or overarching CU factor, and all positively worded items were positively covaried with one another, ranging from .03 to .43 (see Figure 1 for standardized subfactor loadings onto the general factor and covariances between positively worded items for each subfactor).

Measurement Invariance across Sex

Supplementary Table 3 provides fit indices for each CFA in girls and boys separately, with the MTMM four-factor model providing the best fit among the models for each sex. Table 2 shows the model fit indices for each step of testing measurement invariance across sex. All fit indices, apart from the TLI, showed acceptable-to-good model fit for the configural model, suggesting that the factor structure is the same for boys and girls. Good model fit was met for both the RMSEA and SRMR, along with acceptable fit for the CFI and TLI, in the weak factorial model, suggesting that not only is the factor structure consistent across sex, but the factor loadings are consistent as well. For the strong invariance model, good model fit was obtained for the RMSEA and SRMR, along with acceptable fit for the CFI. Based on these fit indices, we can also assume strong factorial invariance for the ICU across sex. It should be noted, however, that the change in TLI from weak factorial to strong models did not meet conservative criteria previously set at <-.01 (see Table 2). Again, failure of the TLI to meet this criterion may be due

to model complexity (see Sass et al., 2014). Finally, strict factorial invariance was assessed by constraining the item residuals to be equal across groups (steps 1 and 2 in Table 2). In both steps, all model fit indices, apart from the TLI, were acceptable-to-good. All model fit indices in step 2 only changed negligibly (i.e., within the recommended cutoff of -.01 for invariance; Cheung & Rensvold, 2002), providing evidence of strict factorial invariance across sex as well.

Measurement Invariance across Age

Supplementary Table 4 provides fit indices for each CFA in younger (ages 11-14) and older (ages 15-17) adolescents separately, with the MTMM four-factor model providing the best fit among the models for each age group. It is important to note that, while the MTMM four-factor model was the best fitting model for the older adolescents, the individual fit indices were generally below an acceptable fit in this age group. Table 3 shows the model fit indices for each stage of testing measurement invariance across age groups. All fit indices, apart from the TLI, showed acceptable-to-good model fit for the configural model, suggesting that the factor structure is the same for younger and older adolescents. For the weak factorial model, the RMSEA and SRMR met criteria for good model fit, along with acceptable model fit for the CFI and TLI, suggesting that the factor loadings are equal across age groups as well. It should be noted, however, that the change in TLI from weak factorial to strong models did not meet conservative criteria set at <-.01 (see Table 3). Again, failure of the TLI to meet this criterion may be due to model complexity (see Sass et al., 2014). For the strong model, good model fit was obtained for the RMSEA and SRMR, along with acceptable fit for the CFI. Based on these fit indices, we can also assume strong factorial invariance for the ICU across the age groups. Finally, both steps in strict factorial models resulted in acceptable-to-good model fit indices,

apart from the TLI. All model fit indices in step 2 were either unchanged or changed negligibly (<-.01). Thus, our findings also suggest strict factorial invariance across age groups.

Group Differences in ICU Total Score

Because measurement invariance was demonstrated for this ICU MTMM four-factor structure across sex and age, we tested for significant group differences in mean total scores. This total score showed acceptable internal consistency (Cronbach's α = .79), and its distribution did not deviate significantly from normality (skewness = .50; kurtosis = .24). Table 4 shows the significant main effects for sex and age (i.e., ages 11-14 vs. 15-17) and a significant interaction effect of sex and age on ICU total score. This interaction effect of sex and age showed that, in both the younger and older age groups, boys scored higher on the ICU than girls, but this difference in mean total scores was greater in the older (i.e., ages 15-17) adolescents.

Discussion

The current results provide important information to guide use of the ICU as a measure of CU traits and as an indicator of the diagnostic specifier "with Limited Prosocial Emotions (LPE)," that is now part of the criteria for Conduct Disorder (CD) in the DSM-5 (American Psychiatric Association, 2013). Of most importance, our results suggest that an inability to find a consistent factor structure in research is likely due to a failure to consider the effects of item wording. That is, consistent with two other recent factor analyses that control for different response patterns for negatively and positively worded items (Kliem et al., 2020; Koutsogiorgi et al., 2020), we found support for the structural model that led to the item content of the ICU, with four item clusters loading onto an overarching general factor (Kimonis et al., 2008). Further, our factor analysis was conducted on a large (N = 4,683) sample of non-referred youth, as were other studies finding support of this factor structure (Kliem et al., 2020; Koutsogiorgi et al., 2020).

Our results suggest that a hierarchical factor structure with one general CU factor and four trait factors or subscales seems to best represent the structure of the ICU. Importantly, our results showed that this factor structure demonstrated strict measurement invariance across boys and girls and across younger (ages 11-14) and older (15-17) adolescents. Similarly, Kliem and colleagues (2020) found support for strict measurement invariance for the four-factor structure that accounts for method variance across sex and ethnicity in their large sample of 9th grade students in Germany. With this, the ICU appears to capture the construct of CU traits similarly across several demographic groups.

Our final aim was to test potential differences in mean scores for boys and girls and across younger and older adolescents. Given that we demonstrated measurement invariance across these different groups, such differences can be interpreted as reflecting meaningful variations in CU traits, rather than differences in how the traits are measured across groups. While sex (e.g., Allen et al., 2020; Carvalho et al., 2017; Ciucci et al., 2014; Essau et al., 2006; Fanti et al., 2009; Fragkaki et al., 2016; Kliem et al., 2020; Pechorro et al., 2019; Pihet et al., 2015; Thøgersen et al., 2020; Ueno et al., 2019) and age (Carvalho et al., 2017; Houghton et al., 2013; Thøgersen et al., 2020; Ueno et al., 2019; Zhang et al., 2019) trends in CU scores have been reported in past research, these have not been based on samples as large as our multinational sample, which allowed us to test sex by age interactions. While boys showed higher levels of CU traits than girls in both age groups, consistent with past research (e.g., Cardinale & Marsh, 2020; Fragkaki et al., 2016; Kliem et al., 2020; Pechorro et al., 2019; Pihet et al., 2015; Ueno et al., 2019), this gap was wider in older adolescents. Overall, these sex differences are consistent with a wealth of data across development documenting that girls tend to show greater levels of empathy and other types of affiliative emotions across the lifespan, with

these increasing into adolescence (see Christov-Moore et al., 2014 for a review). There is evidence to support both biological predispositions as well as cultural expectations to account for these differences that appear to be reflected in ICU scores (Christov-Moore et al., 2014).

Our findings need to be interpreted in light of several limitations. First, while we conducted our study on a large multinational sample, the sample was somewhat homogenous in terms of cultural diversity, largely being of European descent. Thus, the measurement invariance of the ICU needs to be tested across different racial, ethnic, and cultural groups. Further, the youngest age in our sample was 11 but due to limited sample size, especially in the youngest age groups, we were forced to use relatively large age ranges to test measurement invariance across development. Thus, future research should consider more discrete age ranges when testing measurement invariance. Of note, the self-report version of the ICU has been used in children as young as 9 (Cardinale & Marsh, 2020) and has proven to be related to important variables (e.g., conduct problems, negative peer perceptions) in preadolescent children (Matlasz et al., 2022). As a result, measurement invariance should specifically be tested for pre-adolescent children as well. Similarly, there are other versions of the ICU for informant-report (i.e., parent and teacher) that have not been subject to as much research as the self-report. Roose and colleagues (2010) reported measurement invariance across the different report formats in the Belgium sample included in our analyses, and Ueno and colleagues (2019) reported measurement invariance across self-, teacher-, and parent-report in non-referred German children ages 6 to 18 years. However, neither of these studies tested factor models controlling for method variance. As a result, the model tested in the current study needs to be tested in large samples using the informant versions of the ICU. Further, we collapsed across several different ICU language translations for our main tests, and the widely varying sample sizes made testing MI across these

translations difficult. Supplementary analyses provided strong support for configural and weak factorial invariance across translations but not strong or strict invariance (see Footnote 1). These findings suggest that the factor structure and factor loadings were similar across these language translations, but the item thresholds and item residuals varied across languages. Thus, further tests of the measurement invariance of the ICU across cultures and language translations are needed. Finally, given the finding that method variance seems to be due to differences in item endorsement patterns, it will be important to replicate the findings in other samples that may have different distributions of ICU scores. That is, the current community sample did not have as many adolescents with high scores on the ICU as would likely be found in clinic-referred or forensic samples, in which a larger number of youth with elevated levels of antisocial behavior is likely to be found.

With these limitations in mind, our analyses strongly suggest that future studies testing the factor structure of the ICU need to use methods that control for item wording direction. When this is done, a factor structure that is consistent with how the scale was theoretically developed has emerged. That is, past factor models appear to have resulted in subscales on the ICU that likely reflect item wording and the resulting differences in response rates across the negatively and positively worded items, rather than theoretically meaningful dimensions of CU traits. These findings strongly support the recommendation that most of interpretations from the ICU should be made from this total score (see also Ray & Frick, 2018). For most research purposes this can be done using a continuous score but Kemp et al. (2021) provide a number of ways in which either empirically derived or normative cutoff scores can be used to define clinically important groups based on ICU total scores. Importantly, the ICU appears to measure the construct of CU traits similarly across many demographic groups, which supports the use of

mean levels of these traits to form norm-referenced cutoff scores for diagnostic purposes. However, our findings suggest that these cutoffs should reflect differences in the levels of the traits across sex and across younger and older adolescents.

Finally, our findings provide guidance for using the ICU to approximate the DSM-5 LPE specifier. That is, some past research selected individual ICU items that most closely match each of the four symptoms of the specifier and used a minimum rating on each item to indicate symptom presence. An individual was considered to meet the LPE criteria if at least two of the four items reached this symptom threshold. A review of nine studies using this approach indicated that this method resulted in widely varying prevalence rates and poor evidence for its validity (Colins et al., 2020). These negative findings are likely the result of using a very limited item pool (i.e., four items), using a very restricted range of scores (i.e., 0 to 4), and a failure to consider differences in item difficulty associated with positively and negatively worded items. Instead, our findings support a) the clustering of the six items relating to each of the four LPE symptoms and b) the need to control for item wording, which was the approach tested and validated by Kemp et al (2021). This approach uses all six items to assess each symptom and uses different scoring criteria for positively (ratings of “very true” or “definitely true” were scored as indicating symptom presence) and negatively (only “extreme” responses of “not at all true” were scored as indicating symptom presence) worded items, in order to adjust for symptom difficulty. A symptom is considered present if two or more of the six items representing that symptom reach the threshold, and the LPE criteria is considered present if two or more of the symptoms are present. While this method requires further testing, it was associated with risk for later delinquency and aggression in adolescents and for experiencing conduct problems and peer relations difficulties in school children (Kemp et al., 2021).

References

Allen, J. L., Shou, Y., Wang, M. C., & Bird, E. (2020). Assessing the measurement invariance of the Inventory of Callous-Unemotional Traits in school students in China and the United Kingdom. *Child Psychiatry & Human Development*, 1-12. https://doi.org/10.1007/s10578-020-01018-0

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

Andershed, H. A., Kerr, M., Stattin, H., & Levander, S. (2002). Psychopathic traits in non-referred youths: A new assessment tool.

Baroncelli, A., Roti, B., & Ciucci, E. (2018). The associations between callous-unemotional traits and emotional awareness in youth. *Personality and Individual Differences*, *120*, 247-252. https://doi.org/10.1016/j.paid.2017.07.036

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. https://doi.org/10.1007/s10862-014-9420-7

Blair, R. J. R., Leibenluft, E., & Pine, D. S. (2014). Conduct disorder and callous–unemotional traits in youth. *New England Journal of Medicine*, *371*(23), 2207-2216. https://dx.doi.org/10.1056%2FNEJMra1315612

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*(1), 20-34. https://doi.org/10.1007/s10862-012-9315-4

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

Carvalho, M., Faria, M., Conceição, A., de Matos, M. G., & Essau, C. A. (2017). Callous-unemotional traits in children and adolescents. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000449

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M., & Ferrari, P. F. (2014). Empathy: gender effects in brain and behavior. *Neuroscience & Biobehavioral Reviews*, *46*, 604-627. https://doi.org/10.1016/j.neubiorev.2014.09.001

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

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

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

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, *13*(4), 454-469. https://doi.org/10.1177/1073191106287354

Fanti, K. A., Demetriou, C. A., & Kimonis, E. R. (2013). Variants of callous-unemotional conduct problems in a community sample of adolescents. *Journal of Youth and Adolescence*, *42*(7), 964-979. https://doi.org/10.1007/s10964-013-9958-9

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*(4), 285. https://doi.org/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*(4), 251-262. https://doi.org/10.1016/j.ijlp.2012.04.002

Fragkaki, I., Cima, M., & Meesters, C. (2016). The association between callous–unemotional traits, externalizing problems, and gender in predicting cognitive and affective morality judgments in adolescence. *Journal of Youth and Adolescence*, *45*(9), 1917-1930. https://doi.org/10.1007/s10964-016-0527-x

Frick, P. J. (2004). The inventory of callous-unemotional traits. *Unpublished rating scale*.

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*(4), 382-393. https://doi.org/10.1037/1040-3590.12.4.382

Frick, P.J., & Hare, R.D. (2001). *The Antisocial Process Screening Device (APSD)*. Toronto: Multi-Health Systems.

Frick, P. J., & Kemp, E. C. (2021). Conduct Disorders and Empathy Development. *Annual Review of Clinical Psychology, 17*, 391-416. https://doi.org/10.1146/annurev-clinpsy-081219-105809

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

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

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth Modeling: Structural Equation and Multilevel Modeling Approaches*. Guilford Publications.

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*(1), 256. https://doi.org/10.1037/a0034718

Houghton, S., Hunter, S. C., & Crow, J. (2013). Assessing callous unemotional traits in children aged 7-to 12-years: a confirmatory factor analysis of the inventory of callous unemotional traits. *Journal of Psychopathology and Behavioral Assessment, 35*(2), 215-222. https://doi.org/10.1007/s10862-012-9324-3

Kemp, E.C., Frick, P.J., Matlasz, T.M., Clark, J.E., Robertson, E.L., Ray, J.V., Thornton, L.C., Myers, T.D.W., Steinberg, L., & Cauffman, E. (2021). Establishing cut-off scores for the Inventory of Callous-Unemotional Traits (ICU). *Journal of Clinical Child and Adolescent Psychology*. https://doi.org/10.1080/15374416.2021.1955371

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*(4), 625-638. https://doi.org/10.1007/s10802-015-0075-y

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

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

Koutsogiorgi, C. C., Lordos, A., Fanti, K. A., & Michaelides, M. P. (2020). Factorial structure and nomological network of the Inventory of Callous-Unemotional Traits accounting for item keying variance. *Journal of Personality Assessment*, 1-12. https://doi.org/10.1080/00223891.2020.1769112

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486. https://doi.org/10.1037/met0000075

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*(3), 392-406. https://doi.org/10.1007/s10862-014-9469-3

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

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*(3), 471. https://doi.org/10.1037/a0019227

Matlasz, T. M., Frick, P. J., & Clark, J. E. (2021). A Comparison of Parent, Teacher, and Youth Ratings on the Inventory of Callous–Unemotional Traits. *Assessment*, https://doi.org/10.1177%2F10731911211047893

Matlasz, T. M., Frick, P.J., & Clark, J.E. (2022). Understanding the social relationships of youth with callous-unemotional traits using peer nominations. *Journal of Clinical Child and Adolescent Psychology.* E-pub prior to print publication.

McMahon, R. J., Witkiewitz, K., & Kotler, J. S. (2010). Predictive validity of callous–unemotional traits measured in early adolescence with respect to multiple antisocial outcomes. *Journal of Abnormal Psychology*, *119*(4), 752. https://doi.org/10.1037/a0020796

Pechorro, P., Braga, T., Hawes, S. W., Gonçalves, R. A., Simões, M. R., & Ray, J. V. (2019). The Portuguese Version of the Inventory of Callous-Unemotional Traits Self-Report and

its Short Form Among a Normative Sample of Community Youths. *Child Psychiatry & Human Development*, *50*(2), 245-256. https://doi.org/10.1007/s10578-018-0835-3

Pechorro, P., Ray, J. V., Barroso, R., Maroco, J., & Gonçalves, R. A. (2016). Validation of the Inventory of Callous-Unemotional Traits among a Portuguese sample of detained juvenile offenders. *International Journal of Offender Therapy and Comparative Criminology*, *60*(3), 349-365. https://doi.org/10.1177/0306624X14551256

Pihet, S., Etter, S., Schmid, M., & Kimonis, E. R. (2015). Assessing callous-unemotional traits in adolescents: Validity of the inventory of callous-unemotional traits across gender, age, and community/institutionalized status. *Journal of Psychopathology and Behavioral Assessment*, *37*(3), 407-421. https://doi.org/10.1007/s10862-014-9472-8

Poythress, N. G., Dembo, R., Wareham, J., & Greenbaum, P. E. (2006). Construct validity of the Youth Psychopathic Traits Inventory (YPI) and the Antisocial Process Screening Device (APSD) with justice-involved adolescents. *Criminal Justice and Behavior*, *33*(1), 26-55. https://doi.org/10.1177/0093854805282518

Ray, J. V., & Frick, P. J. (2018). Assessing callous-unemotional traits using the total score from the inventory of callous-unemotional traits: A meta-analysis. *Journal of Clinical Child & Adolescent Psychology*. https://doi.org/10.1080/15374416.2018.1504297

Ray, J. V., Frick, P. J., Thornton, L. C., Steinberg, L., & Cauffman, E. (2016). Positive and negative item wording and its influence on the assessment of callous-unemotional traits. *Psychological Assessment*, *28*(4), 394. https://doi.org/10.1037/pas0000183

Roose, A., Bijttebier, P., Decoene, S., Claes, L., & Frick, P. J. (2010). Assessing the affective features of psychopathy in adolescence: A further validation of the inventory of callous

and unemotional traits. *Assessment*, *17*(1), 44-57. https://doi.org/10.1177/1073191109344153

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180. https://doi.org/10.1080/10705511.2014.882658

Thøgersen, D. M., Andersen, M. E., & Bjørnebekk, G. (2020). A multi-informant study of the validity of the inventory of Callous-Unemotional Traits in a sample of Norwegian adolescents with behavior problems. *Journal of Psychopathology and Behavioral Assessment*, 1-13. https://doi.org/10.1007/s10862-020-09788-6

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

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486-492. https://doi.org/10.1080/17405629.2012.686740

Wang, M. C., Gao, Y., Deng, J., Lai, H., Deng, Q., & Armour, C. (2017). The factor structure and construct validity of the inventory of callous-unemotional traits in Chinese undergraduate students. *Plos One*, *12*(12), e0189003. https://doi.org/10.1371/journal.pone.0189003

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. https://doi.org/10.1007/s10862-013-9379-9

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

Zhang, X., Shou, Y., Wang, M., Zhong, C., Luo, J., Gao, Y., & Wendeng, Y. (2019). Assessing callous-unemotional traits in Chinese detained boys: factor structure and construct validity of the inventory of callous-unemotional traits. *Frontiers in Psychology*, *10,* 1841. https://doi.org/10.3389/fpsyg.2019.01841

Table 1

Model Fit Statistics for Confirmatory Factor Analysis (CFA) Models of the Inventory of Callous Unemotional Traits (ICU) Self-Report Version (22-item)

	χ^2	*df*	RMSEA [90%-CI]	CFI	TLI	SRMR
Unidimensional Model	12133.721*	209	.110 [.109 - .112]	.648	.611	.094
Unidimensional Model with Method Modelled	6658.953*	164	.092 [.090 - .094]	.808	.730	.064
Three-Factor Model	5374.872*	206	.073 [.072 - .075]	.847	.829	.064
MTMM Three-Factor Model	3803.795*	161	.070 [.068 - .071]	.892	.846	.051
Four-Factor Model	7595.931*	205	.088 [.086 - .089]	.782	.754	.081
MTMM Four-Factor Model	2797.307*	160	.059 [.057 - .061]	.922	.888	.045

Note. MTMM = multitrait-multimethod approach to CFA denoting modelled method variance; χ^2 from chi-square test; *df* = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Square Residual.
* $p < .001$

Table 2

Tests of Measurement Invariance of MTMM Four-Factor Model across Boys and Girls

Model	χ^2	***df***	**RMSEA [90%-CI]**	**ΔRMSEA**	**CFI**	**ΔCFI**	**TLI**	**ΔTLI**	**SRMR**	**ΔSRMR**
Configural Invariance	2990.765*	320	.060 [.058 – .062]		.914		.876		.049	
Weak Factorial Invariance	2589.097*	342	.053 [.051 – .055]	-.007	.928	.014	.903	.027	.051	.002
Strong Invariance	3189.747*	404	.054 [.053 – .056]	.001	.911	-.017	.898	-.005	.052	.001
Strict Factorial Step 1	3136.656*	382	.055 [.054 – .057]		.912		.893		.051	
Strict Factorial Step 2	3189.747*	404	.054 [.053 – .056]	-.001	.911	-.001	.898	.005	.052	.001

Note. χ^2 from chi-square test*; df* = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; SRMR = Standardized Root Mean Square Residual.
* = $p < .001$

Table 3

Tests of Measurement Invariance of MTMM Four-Factor Model across Age (11-14 vs. 15-17)

Model	χ^2	***df***	**RMSEA [90%-CI]**	**ΔRMSEA**	**CFI**	**ΔCFI**	**TLI**	**ΔTLI**	**SRMR**	**ΔSRMR**
Configural Invariance	3072.212 *	321	.061 [.059 – .062]		.911		.872		.049	
Weak Factorial Invariance	2537.119*	341	.052 [.051 – .054]	-.009	.929	.018	.904	.032	.050	.001
Strong Invariance	3259.477*	403	.055 [.053 – .057]	.003	.908	-.021	.894	-.010	.051	.001
Strict Factorial Step 1	3188.585*	381	.056 [.054 – .058]		.909		.890		.051	
Strict Factorial Step 2	3259.477*	403	.055 [.053 – .057]	-.001	.908	-.001	.894	.004	.051	.000

Note. χ^2 from chi-square test; *df* = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; SRMR = Standardized Root Mean Square Residual.
* = $p < .001$

Table 4

*Group Differences by Sex, Age, and Sex*Age on ICU Self-Report Total Score*

	***n* (%)**	**Mean (*SD*)**	***F* (df_N, df_D)**	***p*-value**	η^2
Boys	2375 (51%)	21.90 (8.45)			
Girls	2308 (49%)	17.18 (7.65)			
Sex Main Effect			395.51 (1, 4681)	<.001	.078
Ages 11-14	2919 (62%)	18.38 (8.18)			
Ages 15-17	1764 (38%)	21.54 (8.41)			
Age Main Effect			155.93 (1, 4681)	<.001	.032
Boys Ages 11-14	1444 (31%)	20.53 (8.25)			
Boys Ages 15-17	931 (20%)	24.03 (8.31)			
Girls Ages 11-14	1475 (31%)	16.28 (7.53)			
Girls Ages 15-17	833 (18%)	18.76 (7.62)			
Sex*Age Interaction			4.53 (1, 4679)	.033	.001

Note. The total ICU score is based on 22 items, with items 2 and 10 excluded. *SD* = standard deviation; *F*-value from ANOVA test; df_N = numerator degrees of freedom; df_D = denominator degrees of freedom; η^2 = partial eta-squared.

Figure 1. *Loading Paths from the MTMM Four-Factor Model (22-item).*