

Modeling maladaptive personality traits with unipolar item response theory: The case of Callousness

Fabia Morales-Vives, Pere J. Ferrando & Ana Hernández-Dorado

To cite this article: Fabia Morales-Vives, Pere J. Ferrando & Ana Hernández-Dorado (2025) Modeling maladaptive personality traits with unipolar item response theory: The case of Callousness, *The Journal of General Psychology*, 152:3, 375-402, DOI: [10.1080/00221309.2024.2404398](https://doi.org/10.1080/00221309.2024.2404398)

To link to this article: <https://doi.org/10.1080/00221309.2024.2404398>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC



Published online: 18 Sep 2024.



Submit your article to this journal [↗](#)



Article views: 904



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 3 View citing articles [↗](#)

Modeling maladaptive personality traits with unipolar item response theory: The case of Callousness

Fabia Morales-Vives , Pere J. Ferrando , and Ana Hernández-Dorado 

Psychology Department, Research Center for Behavior Assessment, Universitat Rovira i Virgili, Tarragona, Spain

ABSTRACT

Most IRT applications in personality assume that the measured trait is a bipolar dimension, normally distributed in the population. These assumptions, however, could be questionable for maladaptive, (quasi) pathological traits that still fall in the normal range. This study focuses on one such trait, Callousness, and uses two different instruments and samples to determine whether there is a basis for modeling it as a unipolar trait instead of a bipolar one. More specifically, the following community samples were used, recruited in several Spanish high schools: a) 719 adolescents (13-19 years old, 55.8% girls), b) 681 adolescents (13-19 years old, 44.9% girls). Callousness was assessed with the Inventory of Callous-unemotional traits and Antisocial behavior in the first sample and with the Inventory of Callous Unemotional traits in the second sample. We compared the outcomes of fitting the Graded-Response model (a bipolar-trait model) and the Log-Logistic model (a unipolar trait model) in these community samples and found that they differed considerably at the scoring level. In terms of accuracy, the conditional reliability functions had opposite patterns: it was maximum at high levels in the Graded-Response model and at low levels in the Log-Logistic model. In terms of validity, the models showed different results regarding the prediction of indirect aggressiveness and non-planning impulsiveness.

ARTICLE HISTORY

Received 6 March 2024
Accepted 9 September 2024

KEYWORDS

Psychopathy; callous-unemotional traits; unipolar variables; graded-response model; log-logistic model

Introduction

Item Response Theory (IRT) models began to be used in personality measurement in the mid-1980s and early 1990s, mainly by psychometricians (e.g. Reise & Waller, 1990; Ferrando, 1994). Since then, they have been increasingly used by applied researchers and practitioners to the point that now they are an accepted and usual choice in this domain (e.g. Bryant-Lees &

CONTACT Fabia Morales-Vives  fabia.morales@urv.cat  Department of Psychology, Universitat Rovira i Virgili, Crta. de Valls s/n., 43007, Tarragona, Spain

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

LaHuis, 2022). Initial implementations, however, were very naïve because standard models that had been designed for measuring abilities and cognitive traits were directly transported into the new domain in the hope that they would also be suitable (e.g., Reise & Moore, 2012). In our view, the overall balance 40 years later is positive: the use of IRT in personality has generally increased rigor and the precision of measurement, and has provided a better understanding of how personality measures function. At the same time, however, this use is not free from problems, it raises new challenges, and it allows us to safely state that there is ample room for improvements. This article aims to explore one of these.

Explicitly or implicitly, most unidimensional IRT applications in personality assume that the trait under study is (a) a bipolar dimension, and (b) (approximately) normally distributed in the population of interest (Reise et al., 2018, 2021). By bipolar trait, we understand one that shows substantial and meaningful variation across all the levels of the continuum (Lucke, 2015; Reise et al., 2018). Therefore, a bipolar trait is equally informative and interpretable at both poles or ends of the continuum, with substantial variation at both. A clear example is extraversion, which has two meaningful poles that are given specific names (extraversion at one pole and introversion at the other).

For most instruments intended to measure normal-range traits in community populations, the assumptions of bipolarity and normality are plausible (Reise et al., 2018, 2021; van der Maas et al., 2011). In other scenarios, however, they fit very uneasily. A typical illustration is a psychiatric symptoms checklist administered in a community sample (see e.g. Morales-Vives et al., 2023). Bipolarity is highly questionable here, as the low trait end is likely to reflect merely the absence of symptoms. And, as for normality, most individuals who do not suffer from clinical disorders or have very low levels are expected to be grouped at the lower end of the trait continuum. The far fewer individuals who do have disorders to a nonnegligible extent, on the other hand, will extend over the upper tail with a higher degree of heterogeneity. In turn, these assumptions suggest that the latent trait distribution in our example has, intrinsically, a low mean and high variance both leading to a pronounced right skewness (e.g., Magnus & Liu, 2018; Morales-Vives et al., 2023).

In psychometric modeling, one possible approach for addressing the two problems above is to change the assumptions about the trait distribution, and, so far, two main approaches have been considered within this line. The first (unipolar modeling) models the trait as a variable that only adopts positive values and which has a skewed distribution in the target population (Lucke, 2013, 2015; Reise et al., 2018, 2021). The second (mixture modeling) assumes a bimodal distribution that reflects a mixture of two populations: (a) asymptomatic and (b) symptomatic with different degrees

of severity (e.g., Magnus & Liu, 2018). In our view, both approaches hold great promise for assessing clinical or psychiatric constructs.

Several studies show that clinical and psychiatric variables such as addictions, suicidal ideation or depressive symptomatology behave as a unipolar variable would be expected to behave (e.g., Lucke, 2013; Morales-Vives et al., 2023). However, whether (quasi) pathological, maladaptive or aversive personality traits have also a similar behavior has barely been considered so far (see Reise & Rodriguez, 2016). More specifically, this could be the case of socially aversive traits such as anger, sadism or psychopathy. Being these traits maladaptive, it is not unreasonable to assume for starting, that most people would have relatively low levels, which in turn would lead to skewed score distributions. Certainly, this result is not sufficient to determine whether a trait is unipolar, but it may be a first indication. If, in addition to this skewness other results that are also to be expected in the case of a unipolar variable were obtained, then, this possibility should be seriously considered. If maladaptive or aversive traits were not truly bipolar, then, treating them as if they were would have important implications, particularly when obtaining individual scores and taking important decisions based on these scores (for example, identifying individuals susceptible to having high trait levels, as previous studies have already shown in the case of clinical variables such as depression or suicidal ideation (e.g., Morales-Vives et al., 2023).

A conceptual review of unipolar modeling

Conceptually, the models of the type considered here can be viewed as transformed versions of the standard unidimensional IRT models in which the scale and distribution of the latent trait is changed. In more detail, in the standard models, the trait is modeled as a variable that is defined in the range minus/plus infinity and is normally distributed (e.g., Reise et al., 2021). In the unipolar models, on the other hand, the trait can only take positive values and its distribution is rightly skewed (Lucke, 2013, 2015). In principle, many theoretical distributions that fulfill these conditions can be considered for modeling a unipolar latent trait. Furthermore, in certain cases, a specific distribution of this type can be chosen for theoretical reasons (see Lucke, 2013, 2015 and Ferrando et al., 2024). However, such a strong foundation cannot be claimed here.

The model we shall consider in this paper is the unipolar graded response model with a lognormal latent distribution (LL-GRM) (e.g., Reise et al., 2018, 2021). This model is the unipolar counterpart of the standard logistic graded response model (GRM) (Samejima, 1969), and, as further discussed below, apart from for its plausibility, it has also been chosen for its psychometric properties. Technical details about the LL-GRM and its

relations to the GRM can be found in Reise et al. (2018, 2021) and in the UNIPOL-GC manual (see below).

At the item calibration level, the unipolar modeling changes the item response functions and transforms the item parameters. So, the response functions are no longer the familiar IRT ogives but power functions (Ferrando et al., 2024). As for the item parameters, they are monotone transformations of the standard IRT parameters, so they retain their interpretation (i.e., thresholds/locations and discriminations) and ordering: If items are to be selected or judged on the basis of their extremeness (locations) and quality (discriminating power), the standard and the unipolar model will make the same choice. Finally, since one is a transformation of the other, both models will have the same fit at the structural (item calibration) level when fitted to the same dataset. This last result implies that the greater suitability of one model or the other cannot be decided on the basis of model-data fit.

Where the two types of model differ more substantially is in the functioning of the derived scores, and this will be the focus of our study. In principle, however, the main differences do not lie in the individual scores themselves, but on the accuracy these scores are assumed to have across the trait levels. More specifically, if scores based on both models are obtained from the same dataset, they will be nonlinearly related but their rank order will be the same. So, if the purpose of the measurement is simply to rank individuals according to their levels, then one model will work just as well as the other (note the parallelism with the item selection scenario above).

To discuss the most important second issue, we shall again consider the clinical example above in which the instrument is administered in a community sample. Suppose first that the data is fitted with the standard IRT model. Because the proportions of endorsement are very low for most items, the test will be considered to be “difficult” and, according to this model (e.g., Hambleton & Swaminathan, 1985), the local accuracy or conditional reliability of the derived scores will be maximal at high trait levels. In other words, the test will be considered to be highly appropriate for differentiating between the (few) individuals who have very high levels. However, it will not be sensitive enough to differentiate between individuals at lower levels.

Assume next that the unipolar IRT model is fitted to this data. The conditional reliability curve will now be a decreasing function that has its highest values at very low trait levels but with considerably lower values at high trait levels (e.g., Ferrando et al., 2024; Huang & Bolt, 2023) (see Figures 2 and 6 below). So, according to this model, the test will accurately differentiate between those individuals who have no, or virtually no, symptoms and

those who clearly do have them. However, it will not be sensitive enough to make finer differentiations between those with high levels (that is, those in which the symptoms are substantial). So, in this respect, with the same data the two models are expected to function in diametrically opposite ways, which means that the choice of the most appropriate model for the data at hand is far more than a technical triviality. However, gathering evidence for making an informed judgment is difficult, because, as mentioned above, both models will fit the data equally.

So far, two main sources of evidence – (a) plausibility and theory, and (b) empirical evidence – have been proposed for deciding between a standard IRT model and its unipolar alternative. As for the first source (e.g., Hambleton & Swaminathan, 1985), we shall explain below why unipolarity is a plausible hypothesis in relation to callousness, which is the specific maladaptive trait on which the present study focuses.

Turning now to the empirical evidence, four main sources have been considered so far. The first is the observation that the distribution of most of the item scores is positively skewed. As mentioned above, however, this is only a useful initial check (Reise et al., 2018), because skewed item distributions can arise from many causes other than trait unipolarity (e.g., van der Maas et al., 2011). The second source is more related to the phases of test design and item selection. If the unipolar model is the most appropriate but the standard bipolar model is used instead, then, finding “easy” items to fill the lower tail of the distribution of the test scores is far more difficult than finding items that better measure at the (presumably more meaningful) upper pole of the dimension (Ferrando et al., 2024; Reise & Rodriguez, 2016).

The last two sources of empirical evidence are the properties of the scores derived from one model or the other. The third source, discussed by Reise and Moore (2012), is the empirical assessment (in applied settings) of whether the raw scores function more like the “decreasing accuracy” mechanism that the unipolar model predicts: i.e. (a) they are more sensitive to differentiating between individuals with very low trait levels (the majority) and those with substantial levels but (b) they are much less sensitive to making finer differentiations among the latter. Finally, some authors (Ferrando et al., 2024; Morales-Vives et al., 2023; Reise et al., 2021) claim that, in order to escape from internal or circular evidence, the scores derived from unipolar modeling should be checked to see whether they behave as theoretically expected when they predict relevant “external” measures (i.e., validity assessment). More in detail, if the trait is unipolar, and far more meaningful at its upper end, a differential validity effect (Ghiselli, 1956) in which the relation is stronger at higher trait levels should be observed.

Are unipolar models appropriate for measuring callousness?

According to Crego and Widiger (2015), psychopathy is perhaps the prototypic personality disorder. While different conceptualizations of psychopathy have been proposed in the literature, they generally include callous-unemotional (CU) traits (lack of empathy, absence of guilt and remorse, ability to manipulate others, irresponsible attitude, and poor emotional expression) as essential components of this construct (e.g., Albert et al., 1959; Cleckley, 1976; Hare, 2003). Furthermore, in recent years there has been an increasing interest in studying the CU traits in children and adolescents, conceptualizing the psychopathic traits as a developmental construct. In fact, high levels of these traits in youth are related to problems such as antisocial behavior and delinquency (e.g., Caputo et al., 1999; Frick et al., 2003; Preston et al., 2020), and premeditated violence (e.g., Frick et al., 2003).

Although psychopathy might be potentially unipolar, measures of this construct are usually fitted and scored using standard models that assume it is bipolar, in spite of the reported problems this practice appears to cause. In particular, this is the case of measures specifically developed to assess the CU components, explained below. And it is also the case of those measures that focus on the dark triad of personality (e.g., Jones & Paulhus, 2014; Webster & Jonason, 2013), a model that has received an increasing interest in recent years, and that includes psychopathy as one of the three maladaptive profiles it assesses. The considerable evidence collected on the dark triad shows that, in spite that these measures are routinely fitted by assuming that the traits are bipolar and normally distributed in community samples, the scores that measure them, particularly the psychopathy subscales (e.g., Kajonius et al., 2016) have strongly skewed distributions. This finding is congruent with what was found in related studies carried out before the dark triad model was proposed, particularly those concerned with psychoticism, which is to some extent a construct similar to psychopathy (Zuckerman, 2012). Initially, Eysenck and Eysenck (1976) considered that the problem of skewed distributions solely reflected a limitation of the measurement instruments. However, the fact that similar results have been repeatedly obtained with different questionnaires that assess psychopathy-related traits, while other measures that assess more clearly bipolar traits do not have this problem, suggests that the problem is not in the scales or in faulty item designs but in the intrinsic nature of the trait that is assessed. In fact, according to the results of the meta-analysis by Sanz-García et al. (2021), the prevalence of psychopathy in the general adult population is only 4.5%.

The most commonly used instrument for assessing the CU traits in adolescents is the Inventory of Callous-Unemotional traits (ICU) (Frick, 2004), which has three subscales: Callousness, Uncaring, and Unemotional.

According to Book and Power (2016), callousness encompasses interpersonal and affective traits that are considered to be markers of Factor 1 in Hare's (2003) conceptualization of psychopathy, including lack of empathy, lack of guilt and shallow affect. However, as Morales-Vives et al. (2019) pointed out, it seems that some items of ICU that focus on these issues are included in subscales other than callousness, while the callousness subscale includes some items very similar to others included in the uncaring subscale. Furthermore, the results of several studies also raise doubts about the content of the subscales, as some items seem to belong to subscales other than those originally proposed (e.g., Ciucci et al., 2014; López-Romero et al., 2015). For these reasons, Morales-Vives et al. (2019) developed the *INventory of Callous-unemotional traits and Antisocial behavior* (INCA), which contains four subscales: unemotional, callousness, uncaring, and antisocial behavior. The subscale callousness of this questionnaire includes items about lack of empathy, guilt and remorse, and manipulation of others, while the items about lack of effort and irresponsibility are included in the subscale uncaring. As occurs with other questionnaires that assess psychopathy-related traits, both ICU and INCA data have been so far fitted and scored assuming bipolarity.

Although each of the callousness-unemotional traits could potentially be a unipolar trait, the focus of the present study is specifically on callousness. As this trait refers to the interpersonal and affective facets of psychopathy, it may be an example of a maladaptive or aversive variable that behaves like a unipolar trait, considering what was discussed above about psychopathy. In fact, in the study by Morales-Vives et al. (2020) conducted in a community sample, more than 60% of the adolescents chose the two lowest Likert response options (or the two highest response options in the case of reversed items) in most of the callousness items, despite these items having 5 response options, thus giving rise to markedly rightly skewed item distributions. Similar results have been reported in other studies (e.g., Cardinale et al., 2021).

The current research

The main aim of the current study is to determine if there is enough basis to model callousness as a unipolar trait, and the evidence is obtained by comparing, mainly at the score level, the results obtained by fitting two different models: the GRM (bipolar traits) and the LL-GRM (unipolar traits). More specifically, this research focuses on the following properties of score estimates: a) their distribution, (b) their quality and accuracy, at both the conditional and marginal levels, and (c) their relation to particular external variables. Properties (a) and (b) can be viewed as “internal” and are relevant to assessing which type of respondents will be accurately differentiated

on the basis of their scores. Property (c) is “external”, and can provide further evidence about callousness in terms of predictability at different levels and for different individuals (Ghiselli, 1956). In this study we have used two external variables for assessing property (c): indirect aggression and non-planning impulsivity. Indirect aggression comprises a wide range of behaviors aimed at damaging the victim’s social image – for example, revealing private information, spreading false rumors, excluding the victim from activities – and giving rise to social isolation (Leenaars & Rinaldi, 2015). As the study by Warren and Clarbour (2009) showed that psychopathy is related to indirect aggression, we expected to find a relationship between callousness and indirect aggression, which is more pronounced at the upper end of the callousness continuum. For this study, we specifically chose indirect aggression, excluding more severe forms of aggression such as physical aggression, because behaviors such as ignoring the other person or saying bad things behind the other person’s back are relatively common (e.g., Toldos, 2005). In contrast, behaviors associated with physical aggression (e.g., hitting or pushing someone to the ground) are not very frequent in adolescents (e.g., Toldos, 2005). So, because most people are not physically aggressive, this variable may be considered potentially unipolar. As indirect aggression does not appear to be potentially unipolar, we believe it to be an appropriate criterion for the purposes of the present study. Furthermore, as several studies show that impulsivity is related to CU traits (e.g., López-Romero et al., 2015; Roose et al., 2010), we also expected to find differential validity effects for non-planning impulsivity, which involves focusing on the present, without considering future consequences.

The expected results above will be compared by using two different instruments (ICU and INCA), administered in different samples. If the results are similar and agree with the expectations, this across-instruments and across-samples congruence can be seen as a further source of evidence about the unipolar nature of the trait.

Study 1: INCAS

In this study, we focus on the results obtained with the callousness subscale of the INCA questionnaire in a sample of 719 adolescents. We also use indirect aggression scores as an external variable.

Materials and methods

Participants

The sample consists of 719 adolescents aged between 13 and 19 years old ($M = 15.22$, $SD = 1.62$) who were recruited in four different high schools

from Catalonia (Spain). Regarding sex, 55.8% of participants reported that the sex assigned at birth was female, and 44.2% reported that it was male. All the participants answered the INCA questionnaire, and 150 of them also answered the *Indirect-Direct Aggression Questionnaire* (I-DAQ). The ages of this subsample ranged between 13 and 19 years old ($M = 15.47$, $SD = 1.66$), and 56.2% were females.

Instruments

The inventory of callous-unemotional traits and antisocial behavior (INCA) (Morales-Vives et al., 2019). We administered the callousness (CA) subscale which consists of 11 items with a 5-point Likert response format (1 = completely disagree; 5 = completely agree). The omega estimated score reliability in this data was $\omega = 0.84$.

The indirect-direct aggression questionnaire (I-DAQ) (Ruiz-Pamies et al., 2014). We administered the 8-item indirect aggression subscale with a 5-point Likert response format (1 = completely disagree; 5 = completely agree). The estimated reliability of the scores was $\omega = .72$.

Procedure

All procedures were performed in compliance with relevant laws and institutional guidelines. Furthermore, the Research and Innovation Ethics Committee (CEIPSA) of Universitat Rovira i Virgili assessed and approved both study 1 and study 2 of this research (reference number: CEIPSA-2021-PR-0028). Both studies were carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for studies involving humans. School approval and parental written informed consent were also obtained before the study. The questionnaires were administered collectively to groups of 15-30 students in their classroom. Participants were asked to volunteer to answer the questionnaires. Anonymity and confidentiality of the data were guaranteed.

Data Analysis. In both studies, the analyses were conducted in the same four stages. In the first stage, sample descriptive statistics were obtained for both the items and test (raw) scores. As discussed above, positively skewed distributions are expected at both levels when the trait behaves as a unipolar dimension, although we stress again that this evidence is inconclusive. The sample reliability of the raw scores (omega coefficient) was also estimated in this stage.

In the second stage (item calibration) both the unipolar LL-GRM and the standard GRM were fitted to the data at the structural level. The GRM

was fitted using the program FACTOR (Lorenzo-Seva & Ferrando, 2006) whereas the LL-GRM was fitted using UNIPOL-GC, which was developed by the authors for the present study (see below). In both cases, the structural solution was fitted by using a limited-information approach (e.g., McDonald, 1982) and robust unweighted least squares estimation with mean and variance corrections. So, because the same calibration procedures were used for both models, the goodness of fit results of the fitted solutions should be the same. Model-data fit was assessed with three measures that cover different facets of fit: (a) the Root Mean Square Error of Approximation (RMSEA), a measure of relative fit per degree of freedom; (b) the Comparative Fit Index (CFI), a measure of comparative fit with respect to the null independence model, and (c) the Correlation Root Mean Squared Residual (CRMSR) (see e.g. Bollen, 1989) as a measure of absolute fit.

The last two stages focused on the internal (third stage) and external (fourth stage) properties of the scores derived from the LL-GRM, which are the core of the present research.

At the beginning of the third stage, and for both models (GRM and LL-GRM), Bayes expected a posteriori (EAP) (Bock & Mislevy, 1982) score estimates were obtained for each individual. In agreement with the models' assumptions, the prior distributions for estimating these scores were: standard normal (for the GRM) and lognormal (for the LL-GRM). Scores derived from the GRM were again obtained using FACTOR (Lorenzo-Seva & Ferrando, 2006) while those derived from the LL-GRM were obtained with UNIPOL-GC. Next, for both the LL-GRM and the standard GRM, the model-based conditional reliability estimates were computed as a function of trait level. In order to obtain accurate EAP point estimates and conditional reliability estimates, 60 quadrature points were used in both cases.

So as to check the third source of empirical evidence above (i.e., the "decreasing accuracy" mechanism), we empirically assessed how the raw scores behave in terms of conditional accuracy without resorting to any specific model. This assessment was made to see which of the two model-based predictions (LL-GRM or GRM) is closest to the empirical functioning of the simple raw scores. To this end, we empirically obtained the conditional standard error of measurement for the raw scores using Thorndike's (1971) difference method.

Turning finally to the fourth (external properties) stage, the scores on the external variable indirect aggression were regressed on the LL-GRM estimated trait scores for both questionnaires. We also carried out the same procedure with the external variable non-planning impulsiveness for the ICU questionnaire. The main focus was not on the validity estimates per se, but rather on whether the relations were differentially stronger at

different regions of the estimated trait levels (i.e., differential validity) as would be expected if callousness behaved as a unipolar trait. Graphically, the differential validity hypothesis considered here implies a type of negative heteroskedastic relation that varies regularly with the estimated trait level (i.e., as the trait level increases the conditional errors of prediction tend to decrease). So, the scatter of points around the regression line is more dispersed at the lower end of the trait continuum than at the upper end. A formal test of this hypothesis (e.g., Cohen & Cohen, 1983) can be obtained by regressing the absolute regression residuals on the estimated trait levels: a significant negative slope (or correlation) would provide evidence for the type of differential validity expected here.

Results

Descriptive item statistics

Table 1 shows the descriptive statistics for the INCA items, after scoring all the items in the same direction of the trait. As predicted, most items have skewness coefficients higher than 1 in absolute value. Furthermore, in all but one item, more than 70% of participants endorsed the response options 1 and 2, which suggests that most of the sample had very low levels on most behaviors or feelings that are indicators of callousness. In fact, for items 2, 3, 4, 8, 10 and 11, between 50% and 65% endorsed the lowest option.

Table 1. Descriptive statistics of the INCA items.

	Mean	Standard deviation	Skewness	Percentage of responses				
				1	2	3	4	5
1. I do what I like even if it might be detrimental to other people	2.02	1.10	0.95	40.9	30.6	17.4	7.5	3.6
2. I take advantage of others	1.56	0.93	1.88	65.4	20.7	8.9	2.6	2.4
3. I feel bad about inconveniencing others*	1.79	1.11	1.47	55.1	25.5	9.3	5.6	4.6
4. I feel bad when I hurt someone*	1.74	1.10	1.59	58.8	21.8	11.3	2.9	5.1
5. Seeing other people's misfortunes upsets me*	1.87	1.10	1.32	48.8	28.9	13.2	4.3	4.7
6. I seldom apologize when I make a mistake	2.61	1.50	0.41	32.1	22.8	14.0	14.3	16.7
7. I care about others*	1.76	0.89	1.37	45.9	38.9	10.4	2.9	1.8
8. It is logical that clumsy people are made fun of	1.81	1.06	1.32	52.6	25.7	13.6	4.7	3.3
9. I feel bad for people who are worse off than I am*	1.95	1.10	1.07	44.6	29.9	14.6	7.4	3.5
10. I am often very pleasant with people I do not like so that I can get something out of them	1.86	1.12	1.22	52.7	22.7	14.6	6.0	4.0
11. I sometimes use others to get what I want	1.94	1.17	1.10	50.1	22.4	15.7	7.1	4.7
Raw sum scores	20.90	6.60	0.85					

*Reverse-scored item in the original presentation.

Item calibration

After freeing two doublets, the pairs 2-11 and 3-4, the goodness-of-fit results (which were the same for both the GRM and the LL-GRM), were: RMSEA (90% confidence interval) = .077 (.067 – .087), CFI = .929, and CRMSR = .065. Overall, all the indicators suggest that the model-data fit is acceptable. As for the identified doublets, the content of items 2 (*I take advantage of others*) and 11 (*I sometimes use others to get what I want*) is very similar, as both items refer to taking advantage of others for one's own benefit. Likewise, the content of items 3 (*I feel bad about inconveniencing others*) and 4 (*I feel bad when I hurt someone*) is very similar, referring to feeling bad about hurting others. So, there was a clear substantive foundation for freeing these residuals.

Internal properties of individual scoring

The histogram of the INCA raw scores is shown in [Figure 1](#). Clearly, the distribution is rightly skewed, although, as expected, its skewness coefficient is lower than the average item skewness (see [Table 1](#)). For illustrative purposes, the best fitting normal distribution (thin-dashed-line), lognormal distribution (thick-solid-line) and kernel density estimate (thin-solid-line) are superimposed. The kernel density estimate is, essentially, a nonparametric estimate of an unknown probability density function (e.g., Silverman, 1986). So, it does not make parametric assumptions about the shape of the distribution and can be taken as a reference for the other two distributions.

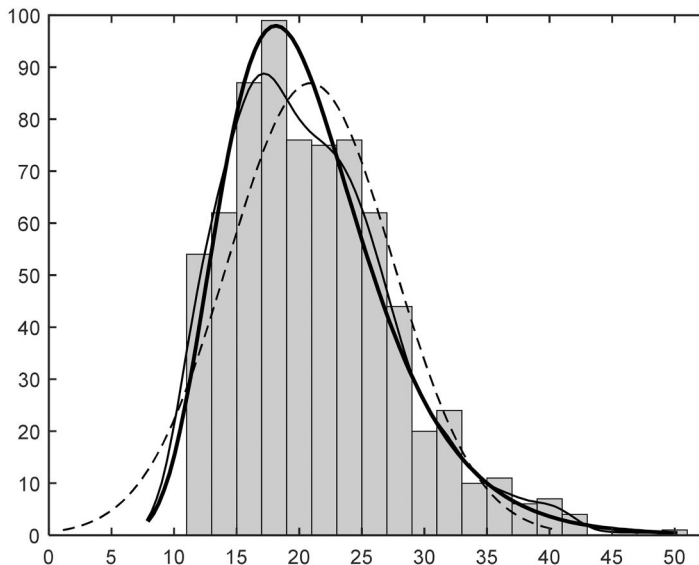


Figure 1. INCA questionnaire: Histogram of the raw scores with the best fitting normal distribution (thin-dashed-line), lognormal distribution (thick-solid-line), and kernel density estimate (thin-solid-line).

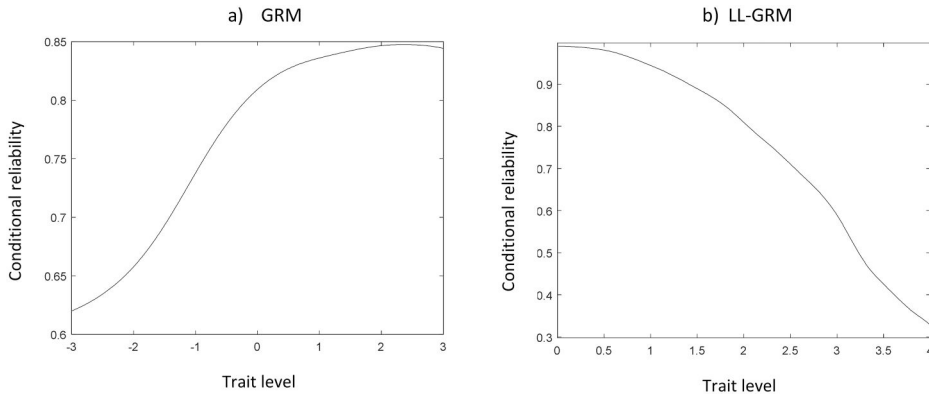


Figure 2. INCA questionnaire: Conditional reliability of the Callousness trait estimates at different trait levels for the GRM and LL-GRM models.

As a summary, (a) the lognormal and the kernel estimates fit better the histogram than the normal, and (b) they are quite similar to each other. With all the limitations discussed above, this result is consistent with what would be expected if the trait was unipolar.

Figure 2 shows the conditional reliability as a function of the trait level for both the GRM and the LL-GRM. As can be seen, the conditional reliability curves follow opposite patterns in each model: In the GRM, accuracy is maximum at high levels, while in the LL-GRM, it is maximum at low levels. Therefore, as has been previously found (Morales-Vives et al., 2023; Reise et al., 2018, 2021), if the GRM was correct, the GRM-based scores would enable us to distinguish between different levels of callousness severity, but not between those participants who are clearly characterized by this trait and those who are not. In contrast, the conditional reliability curve in the LL-GRM peaks at very low levels of the trait, although it is still somewhat accurate up to about half a standard deviation ($SD = 1.78$) above the mean ($M = 1.58$). Beyond this point, it is considerably less so. Therefore, if the LL-GRM was correct, the LL-GRM-based score estimates would allow us to differentiate the individuals who are callous from those who are not, but do not allow us to differentiate between different levels of callousness severity (see e.g Magnus & Liu, 2018; Morales-Vives et al., 2023; or Reise et al., 2018, 2021 for related results).

Figure 3 shows the empirical standard errors plotted against the raw INCA scores. As can be seen, the measurement errors are widely spread, which can be explained by the moderate reliability of the raw scores in this short scale. Even so, it seems clear ($r = .35, p < .05$) that the measurement errors tend to increase with the scores, which, as discussed above, is consistent with the expected functioning when the trait is unipolar.

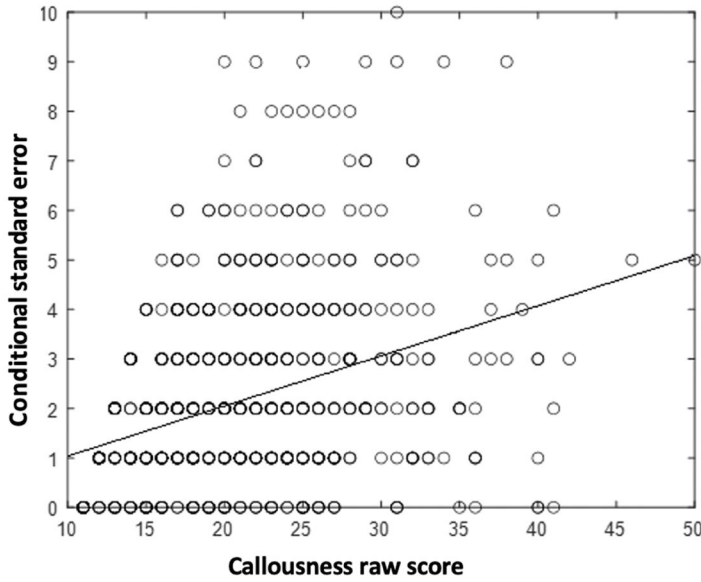


Figure 3. INCA questionnaire: Empirical conditional standard errors of measurement against raw scores.
Note. Darker shades in the circles mean a larger number of observations.

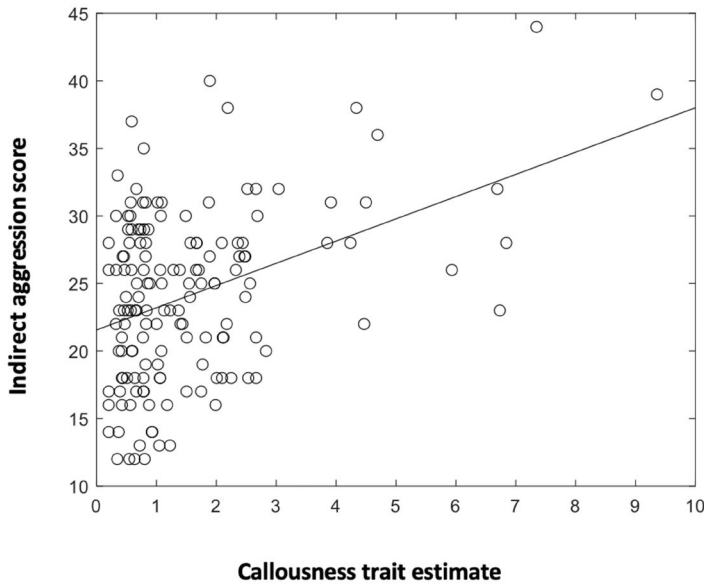


Figure 4. INCA questionnaire: Bivariate scatterplot between callousness and indirect aggression using the LL-GRM model.

External validity evidence

The bivariate scatterplot between the LL-GRM callousness scores and indirect aggression scores is shown in [Figure 4](#). It is noted that the trend of the regression is nonlinear, which was expected (see Ferrando et al., [2024](#)).

When the relation was linearized, the LL-GRM-based validity coefficient estimate was $r = .38$ ($p < .05$). The validity estimates obtained with the GRM scores and the raw scores were $r = .36$ and $r = .25$ respectively; somewhat smaller, although the differences between the three estimates were not statistically significant. We should stress again that this result is expected, because the rank order of the scores is virtually the same in the three schemas. So, the differences due to the use of the different models should not be looked for here.

The most important trend in [Figure 4](#), instead, is that the expected heteroscedasticity relation seems to be evident in the graphic, and was supported by the results of the formal test described above: The correlation between the absolute residuals and the estimated trait levels was $r = -.20$ ($p < .05$), negative and significant as expected. Furthermore, this negative trend was also observed with the scores derived from the GRM ($r = -.11$ $p < .05$), and also with the raw scores ($r = -.09$). In this last case, however, statistical significance was not attained. As a final source of evidence regarding an expected stronger relation at high trait levels we used that proposed in [Ferrando et al. \(2024\)](#): Based on the scatterplot in [Figure 4](#), we computed the validity coefficient estimate for the cases with θ_U estimates below and above 3. In the first sub-group, the validity estimate was $r = .16$. In the second sub-group, it was $r = .41$. The difference (obtained using Bootstrap resampling) was statistically significant.

A plausible substantive interpretation, for the result above is that at low trait levels, in which callousness is less meaningful, whether or not individuals engage in indirect aggressive behaviors will probably depend more on other types of variables (e.g., their regulation of emotions, their ability to resolve conflicts without resorting to aggressive behavior, etc.). However, at higher levels, the trait is more relevant, so that people high in callousness are more likely to engage in indirect aggressive behavior. In fact, the literature shows that adolescents with high levels of CU traits tend to be more aggressive, which seems related to their lack of empathy and guilt, while aggressiveness and antisocial behavior in adolescents with low CU traits tend to be related with problems such as a poor emotional regulation, with higher levels of anxiety and reactivity at negative stimuli (see for example the review by [Frick, 2006](#)).

Study 2: ICU

The aim of study 2 was to determine whether the results previously obtained with the INCA questionnaire were replicated in a new sample with a different instrument: the *Inventory of Callous Unemotional traits*

(ICU). Furthermore, in this study we use indirect aggression and non-planning impulsivity as external variables.

Materials and methods

Participants

The sample consists of 681 adolescents aged between 13 and 19 years old ($M = 15.86$, $SD = 2.03$), who were recruited in three different high schools from Catalonia (Spain). Regarding sex, 44.9% of participants reported that the sex assigned at birth was female, and 55.1% reported that it was male. All the participants answered the ICU and I-DAQ questionnaires, and 486 of them also answered the *Barratt Impulsiveness Scale – 11 for children* (BIS-11c). The ages of this subsample ranged between 13 and 19 years old ($M = 15.34$, $SD = 2.02$), and 50.2% were females.

Instruments

The inventory of callous unemotional traits (ICU) (Frick, 2004). We used the Spanish adaptation developed by López-Romero et al. (2015). As we specifically aimed to assess callousness, we selected the 9 items whose content univocally refer to lack of empathy, lack of guilt, shallow affect and manipulation of others (i.e., the core definition of callousness) regardless of the scale on which these items were included in the questionnaire. The items are shown in Table 2. Their response format is a 4-point Likert (1 = Not at all true; 4 = Definitely true). The score reliability estimates in the sample were $\omega = .79$.

The indirect-direct aggression questionnaire (I-DAQ) (Ruiz-Pamies et al., 2014). We administered the indirect aggression (IA) subscale as explained in Study 1. The estimated reliability of the scores was $\omega = .70$.

The Barratt impulsiveness scale - 11 for children (BIS-11c) (Chahin et al., 2010). We administered the non-planning impulsivity subscale which consists of 8 items with a 4-point Likert response format (0 = never/almost never; 3 = always/almost always). The estimated score reliability in this data was $\omega = .75$.

Procedure

As explained in Study 1, all procedures were performed in compliance with relevant laws and institutional guidelines. We obtained school approval and parental written informed consent before the study. The questionnaires were administered collectively to groups of 15-30 students, in their

Table 2. Descriptive statistics of the ICU items.

Item	Mean	Standard deviation	Skewness	Percentage of responses			
				1	2	3	4
1. I do not care who I hurt to get what I want	1.33	0.71	2.37	77.8	14.8	4.1	3.2
2. I feel bad or guilty when I do something wrong*	1.85	0.98	0.77	49.2	23.6	19.7	7.5
3. I am concerned about the feelings of others*	1.81	0.85	0.76	43.2	36.4	16.4	4.0
4. I easily admit to being wrong*	2.28	0.95	0.10	25.6	30.7	33.9	9.8
5. I apologize ("say I am sorry") to persons I hurt*	1.61	0.83	1.12	58.1	25.1	14.0	2.8
6. I try not to hurt others' feelings*	1.61	0.86	1.26	59.2	24.8	11.6	4.4
7. I do not feel remorseful when I do something wrong	1.70	0.95	1.21	56.4	25.6	9.8	8.2
8. The feelings of others are important to me*	1.40	0.72	1.94	71.1	20.9	5.3	2.8
9. I do things to make others feel good*	2.06	0.84	0.38	27.9	43.5	23.8	4.8
Raw sum scores	15.65	4.22	0.75				

*Reverse-scored item in the original presentation.

classroom. Participation was voluntary, and we guaranteed anonymity and confidentiality of the data.

Data analysis. We used the same procedures as in Study 1, which are explained above.

Results

Descriptive statistics for the items

Table 2 shows the descriptive statistics for the ICU items, after scoring all the items in the same direction of the trait. As was found with the INCA, many items have positive skewness coefficients higher than 1, with the distributions of items 1 and 8 being particularly skewed. In fact, in all but one item, more than 71% of participants endorsed the response options 1 and 2. Moreover, in five of these items, more than 82% endorsed these response options. These results suggest that most of the sample had very low levels on most behaviors or feelings that are indicators of lack of empathy, lack of guilt, shallow affect and manipulation of others.

Item calibration

After freeing a doublet, the pair 3-8, the following goodness-of-fit results were obtained for both the GRM and the LL-GRM, suggesting an acceptable model-data fit: RMSEA (90% confidence interval) = .060 (.047 – .074), CFI = .945, and CRMSR = .058. As far as the doublet identified is concerned, the content of items 3 (*I am concerned about the feelings of others*) and 8 (*I don't care about other people's feelings*) is almost equivalent, referring to the

concern for the feelings of others, with the difference that they are written in the opposite direction, with item 8 being the reverse form of item 3. There was therefore a clear substantive basis for freeing this residual.

Internal properties of individual scoring

Figure 5 shows the histogram of the ICU raw scores with the same specifications as in Figure 1 above. The results obtained in the first study are replicated again: the distribution is rightly skewed, and the lognormal and the kernel estimates fit better the histogram than the normal, being rather similar to each other.

Figure 6 shows the conditional reliability of the trait estimates for the GRM and the LL-GRM. As was found with the INCA, the conditional reliability functions follow opposite patterns in each model: In the GRM, accuracy is maximum at high levels, while in the LL-GRM, it is maximum at low levels.

Figure 7 shows the empirical standard errors plotted against the raw ICU scores. Although the measurement errors are again widely spread, they tend to increase with the scores ($r = .26, p < .05$), which is more consistent with the expected functioning when the trait is unipolar, as explained above.

External validity evidence

Figure 8 shows the bivariate scatterplot between the LL-GRM callousness scores and indirect aggression scores, which, again, follows a non-linear

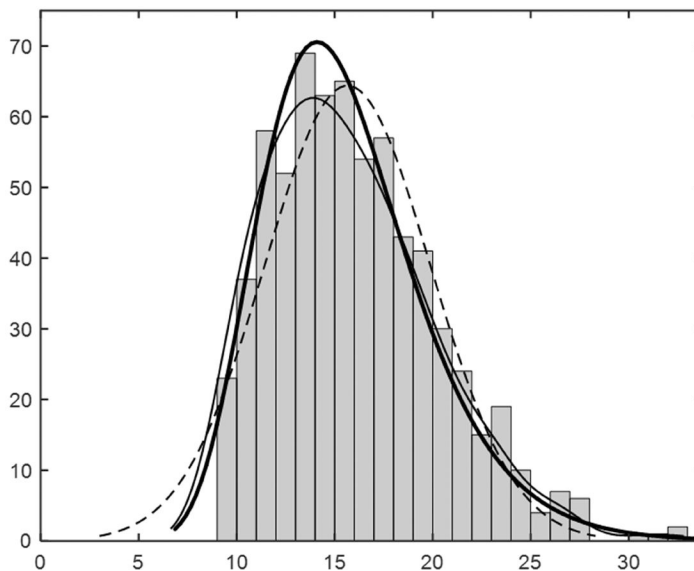


Figure 5. ICU questionnaire: Histogram of the raw scores with the best fitting normal distribution (thin-dashed-line), lognormal distribution (thick-solid-line), and kernel density estimate (thin-solid-line).

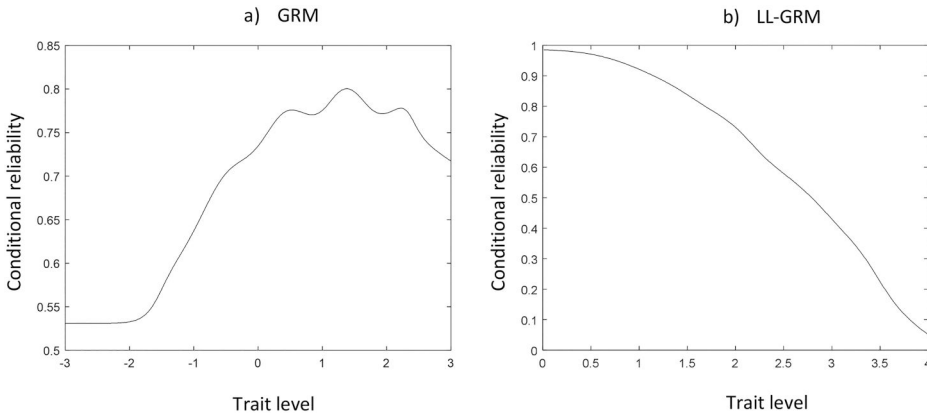


Figure 6. ICU questionnaire: Conditional reliability of the Callousness trait estimates at different trait levels for the GRM and LL-GRM models.

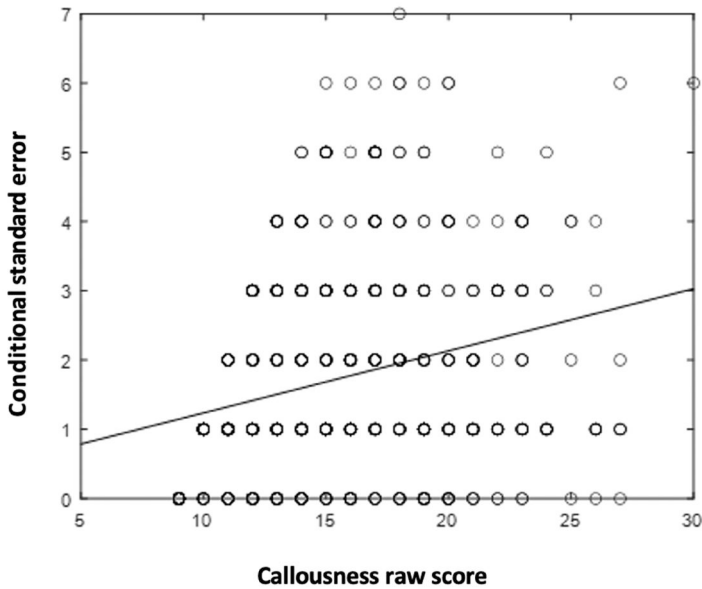


Figure 7. ICU questionnaire: Empirical conditional standard errors of measurement against raw scores.

Note. Darker shades in the circles mean a larger number of observations.

trend. The validity coefficient estimate based on the LL-GRM score estimates was $r = .36$ ($p < .05$), and the validity estimates obtained from the GRM scores and the raw scores were $r = .35$ and $r = .34$ respectively, all very similar as expected (none of the differences was statistically significant). As for the main result, the expected heteroscedasticity relation is evident again in the graphic, and was supported by the results of the formal test ($r = -.18$; $p < .05$). Furthermore, as in the previous study, the heteroscedasticity relation was also obtained with both the GRM-based scores

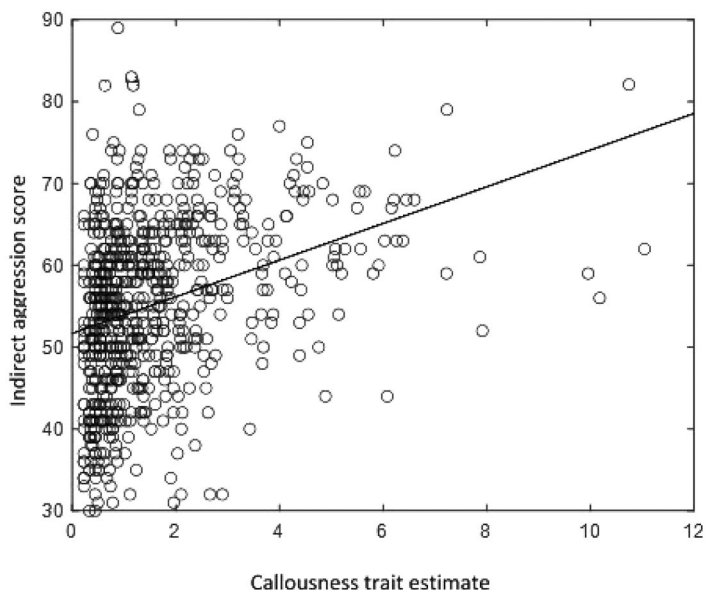


Figure 8. ICU questionnaire: Bivariate scatterplot between callousness and indirect aggression using the LL-GRM model.

($r = -.10$, $p < .05$) and the raw scores ($r = -.10$, $p < .05$). As for the validity coefficient estimates obtained for the cases with θ_U below and above 4 (see Figure 8), they were: $r = .25$ and $r = .33$. The difference (obtained using Bootstrap resampling), however, was nonsignificant in this case.

The bivariate scatterplot for non-planning impulsivity with the LL-GRM model, finally, is shown in Figure 9, and the results obtained with this second external variable are consistent with the previous results: (a) a non-linear trend; (b) similar validity estimates among the different scores, with no significant differences: $r = .26$ (LL-GRM); $r = .24$ (GRM); and $r = .24$ (raw scores); and (c) (main result) evidence of a heteroskedastic relation that varies regularly with the estimated trait level. Here too, the formal test supported this evidence ($r = -.17$, $p < .05$), and the relation was also observed with both the GRM-based scores ($r = -.11$, $p < .05$) and the raw scores ($r = -.08$), although in this last case significance was not attained. Finally, the validity coefficient estimates obtained for the cases with θ_U below and above 3 (see Figure 4) were: $r = .11$ and $r = .25$. The difference (obtained using Bootstrap resampling) was statistically significant. Conceptually, the conjecture is again that higher levels of callousness are more relevant for predicting higher non-planning impulsivity, whereas low levels of callousness are less relevant, possibly because at these low levels, impulsive behavior depends more on other variables.

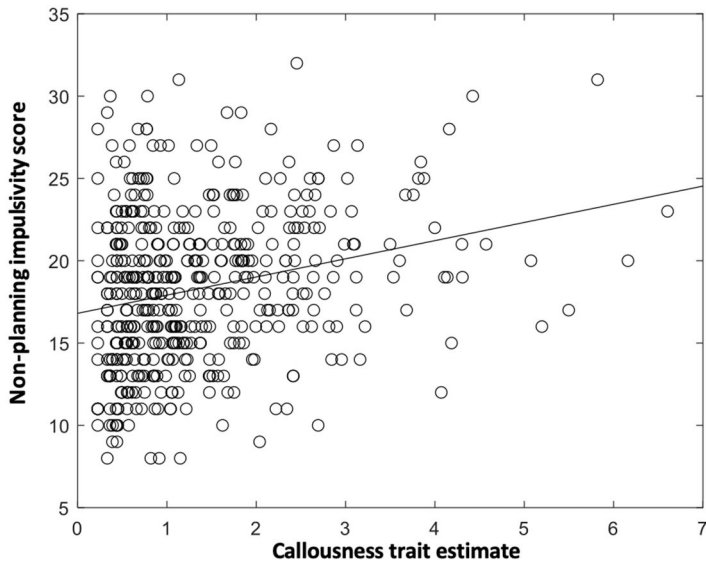


Figure 9. ICU questionnaire: Bivariate scatterplot between callousness and non-planning impulsivity using the LL-GRM model.

Discussion

In recent years, increasing evidence has become available about the unipolar nature of several clinical psychological variables (e.g., Lucke, 2015; Reise et al., 2018, 2021), but the possible unipolar nature of maladaptive or aversive personality traits has not been considered yet (however, see Reise & Rodriguez, 2016). As psychopathy implies a deviation from normality, it seems logical to assume that most people do not have sufficiently high levels of the trait to be considered “psychopaths”. In fact, several studies have reported relatively large skewness for psychopathy scores in community samples (e.g., Anderson & Cheers, 2018; Kajonius et al., 2016). This makes it possible to equate this variable, to some extent, with unipolar clinical variables, since there are also few people who have sufficient symptomatology to be diagnosed as suffering from depression, suicidal ideation, etc. Because CU traits are crucial components of psychopathy (e.g., Cleckley, 1976; Hare, 2003), the current study attempts to determine if callousness behaves as a unipolar variable would be expected to behave. The preliminary results showed that (a) the item scores in both measures were considerably skewed, as expected when a unipolar trait is measured, and (b) The two IRT models that were compared provided equivalent results and fitted equally well at the calibration level. As discussed above, the evidence provided by the first result is not conclusive, and the second result is only to be expected and so is a mere check.

At the scoring level, however, things were different. For both the INCA and ICU measures, opposite patterns of conditional reliability were

obtained for the two models we compared: accuracy was maximum at high levels for the GRM and at low levels for the LL-GRM. This is consistent with previous studies conducted with clinical unipolar variables (e.g., Morales-Vives et al., 2023; Reise et al., 2018, 2021).

Two sources of empirical evidence were used for deciding which of the two models better reflects the nature of callousness as a trait and is interesting to note that the predictions derived from these sources go in opposite directions. Thus, in “internal” terms, the prediction is that the trait levels can be more accurately measured at low levels. So, low scores are expected to be more accurate than high scores. However, in “external” terms, the prediction is that trait levels are more relevant predictors when they are high. So, the conditional errors of prediction are expected to be lower for high scores.

Regarding the internal predictions, evidence based on the raw scores (which are not derived from either of the two competing LL-GRM and GRM) consistently suggests that the distribution of the scores is closer to the lognormal than to the normal, and that the empirical measurement errors tend to increase with the scores, as predicted from the LL-GRM. And, as for the “external” predictions, statistically significant evidence of a heteroskedastic relation that varies with the estimated trait level was consistently obtained in the three validity studies. Furthermore, this evidence was also obtained when based on the bipolar GRM scores (in which a homoscedastic relation would be expected) and on the raw scores. Although the effect sizes corresponding to these latter results are small, their consistency suggests that this relation is not an artifact but reflects ‘true’ validity relations.

If it is tentatively accepted that callousness behaves more as a unipolar trait than a bipolar trait, and so, that the LL-GRM is appropriate for calibrating and scoring callousness measures, some important substantive and practical consequences will be derived, starting from item selection and test design. Thus, if the aim is to distinguish between clearly callous individuals and individuals that hardly have manifestations of this trait, then, the present results suggests that measures such as those used here work reasonably well. However, if there is interest in measuring accurately also at high levels, then, more extreme “difficult” items should be included in the test (see Figures 2 and 6). We note that this recommendation is just the opposite of what is currently being done.

Regarding the trait range at which most accuracy is needed, given the relationship between the presence of CU traits and antisocial and delinquent behavior (e.g., Caputo et al., 1999; Frick et al., 2003), it seems particularly relevant to be able to distinguish accurately between adolescents who are characterized by these traits associated to psychopathy, such as

callousness, and adolescents who are not. In fact, most adolescents who come into contact with the juvenile justice system have committed minor offenses and later desist of these behaviors, but a small percentage of adolescents exhibit more problematic behaviors, which, in addition, remain more stable over time (Moffitt, 1993, 2003), and, in this latter case, the CU traits seem to play a key role (Frick, 2006). For this reason, Docherty et al. (2017) have proposed cutoff points for the ICU questionnaire that makes easier to identify those adolescents characterized by these traits. When the aim is precisely to identify the presence of callousness, it should be considered that the LL-GRM (provided that it was the more correct model), applied to measures such as those considered here, would allow this differentiation to be made with greater precision, making easier to predict which teenagers are more likely to engage in serious antisocial behavior and delinquency. This early identification, in turn, may allow educational measures to prevent this kind of behaviors with negative consequences for the young people themselves and for the community to be implemented. So, to implement preventive programs addressed to these teenagers before such behaviors start to appear seems to be particularly useful in community samples. On the other hand, in samples of young offenders, who have already engaged in antisocial behaviors, it may be useful to differentiate between different levels of high callousness, in order to determine the severity of this trait and thus tailor the intervention to the adolescent's characteristics. In this case, as discussed above, and provided again that the LL-GRM was the most correct model, the instruments considered here should be improved, so as to attain reasonable accuracy even at higher trait levels.

To sum up, the evidence provided by the current study suggests that callousness behaves more as a unipolar trait than as a bipolar trait: The score distribution, conditional reliability and external validity results are all consistent with unipolarity, as expected. Admittedly, certain results are not very strong, but all the predictions were fulfilled and most of them attained statistical significance. Also, we believe that the main strength of what is presented here is that similar results have been obtained with two different instruments and in two different samples. So, they do not seem that can be explained by the limitations or characteristics of a particular questionnaire or that are sample specific.

At the methodological level, the modeling we propose here can be considered as a novelty in the field. And, so far, there is still no available software for fitting the LL-GRM in applications. For this reason, the LL-GRM has recently been implemented in R under the name UNIPOL-GC and is available, together with a complete manual and demonstrative examples at the following link: https://www.psicologia.urv.cat/media/upload/domain_2082/arxius/Utilitats/UNIPOL-GC/description.html

Like any novel study, this one has its share of limitations. To start with, it focuses on a specific trait: callousness, which is a component of psychopathy, but further studies should be done with other instruments that assess the construct of psychopathy more broadly, to gain further evidence on the possible unipolar nature of this construct. Furthermore, other possible traits in the maladaptive but still-normal-range should also be assessed. Secondly, although the present results are consistent, most of the obtained effect sizes are only modest. So, far more evidence is needed to determine the strength and generalizability of the present results, which are only a first step. Regarding this last point, it would be interesting to undertake an extended study in which the results obtained in the modeling of callousness were compared with those obtained with a variable that can clearly be considered to be unipolar. For this comparison to be meaningful, this other variable should be more extreme and maladaptive than callousness, so that it could serve as a benchmark to determine whether the performance of callousness scores is close enough to that obtained with the univocally unipolar variable, despite not being as extreme.

A different type of limitation, although possibly of less practical relevance, concerns the choice of the theoretical latent trait distribution. As discussed above, the lognormal distribution agrees with that expected in a unipolar model and has important psychometric advantages. Apart from that, however, there are no further substantive reasons for choosing this particular distribution among other theoretical distributions that might also work well. So, in future studies it would be interesting to consider alternative latent distributions, such as Weibull's. Another possibility in the future is to consider a unipolar model as appropriate at the structural level, start by using a mild prior, and next estimate the latent density empirically.

Disclosure statement

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

Funding

This work was supported by the Spanish Ministry of Science and Innovation under grant PID2023- 148374NB-I00, and by the Catalan Ministry of Universities, Research and the Information Society under grant 2021 SGR 00036. The funding source was not involved in any step of the research process, neither in the writing and publication process.

ORCID

Fabia Morales-Vives  <http://orcid.org/0000-0002-2095-0244>

Pere J. Ferrando  <http://orcid.org/0000-0002-3133-5466>

Ana Hernández-Dorado  <http://orcid.org/0000-0001-9502-9735>

References

- Albert, R. S., Brigante, T. R., & Chase, M. (1959). The psychopathic personality: A content analysis of the concept. *The Journal of General Psychology*, *60*(1), 17–28. doi:10.1080/00221309.1959.9710204
- Anderson, J., & Cheers, C. (2018). Does the dark triad predict prejudice?: The role of machiavellianism, psychopathy, and narcissism in explaining negativity toward asylum seekers. *Australian Psychologist*, *53*(3), 271–281. doi:10.1111/ap.12283
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. doi:10.1177/014662168200600405
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons. doi:10.1002/9781118619179
- Book, A., & Power, J. (2016). Callousness. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–3). Springer International Publishing. doi:10.1007/978-3-319-28099-8_1051-1
- Bryant-Lees, K. B., & LaHuis, D. M. (2022). Clarifying personality measurement in industrial-organizational psychology: The utility of item response tree models. *Personality and Individual Differences*, *187*, 111410. doi:10.1016/j.paid.2021.111410
- Caputo, A. A., Frick, P. J., & Brodsky, S. L. (1999). Family violence and juvenile sex offending. *Criminal Justice and Behavior*, *26*(3), 338–356. doi:10.1177/0093854899026003004
- Cardinale, E. M., Ryan, R. M., & Marsh, A. A. (2021). Maladaptive fearlessness: An examination of the association between subjective fear experience and antisocial behaviors linked with callous unemotional traits. *Journal of Personality Disorders*, *35*(Suppl A), 39–56. doi:10.1521/pedi_2020_34_486
- Chahin, N., Cosi, S., Lorenzo-Seva, U., & Vigil-Colet, A. (2010). Stability of the factor structure of Barrat's Impulsivity Scales for children across cultures: A comparison of Spain and Colombia. *Psicothema*, *22*(4), 983–989. <https://www.psicothema.com/pdf/3830.pdf>
- 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. doi:10.1007/s10862-013-9384-z
- Cleckley, H. (1976). *The mask of sanity*. Mosby.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. LEA.
- Crego, C., & Widiger, T. A. (2015). Psychopathy and the DSM. *Journal of Personality*, *83*(6), 665–677. doi:10.1111/jopy.12111
- Docherty, M., Boxer, P., Huesmann, L. R., O'Brien, M., & Bushman, B. (2017). Assessing callous-unemotional traits in adolescents: Determining cutoff scores for the inventory of callous and unemotional traits. *Journal of Clinical Psychology*, *73*(3), 257–278. doi:10.1002/jclp.22313
- Eysenck, H. J., & Eysenck, S. B. G. (1976). *Psychoticism as a dimension of personality*. Crane, Russak & Company, Inc. doi:10.1007/s10862-008-9111-3
- Ferrando, P. J. (1994). Fitting item response models to the epi-A impulsivity subscale. *Educational and Psychological Measurement*, *54*(1), 118–127. doi:10.1177/0013164494054001016

- Ferrando, P. J., Morales-Vives, F., & Hernández-Dorado, A. (2024). Measuring unipolar traits with continuous-response items: Some methodological and substantive developments. *Educational and Psychological Measurement*, 84(3), 425–449. doi:10.1177/00131644231181889
- Frick, P. J. (2004). *The inventory of callous-unemotional traits. Unpublished rating scale.* University of New Orleans.
- Frick, P. J. (2006). Developmental pathways to conduct disorder. *Child and Adolescent Psychiatric Clinics of North America*, 15(2), 311–331, vii. doi:10.1016/j.chc.2005.11.003
- Frick, P. J., Cornell, A. H., Barry, C. T., Bodin, S. D., & Dane, H. E. (2003). Callous-unemotional traits and conduct problems in the prediction of conduct problem severity, aggression, and self-report of delinquency. *Journal of Abnormal Child Psychology*, 31(4), 457–470. doi:10.1023/A:1023899703866
- Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, 40(6), 374–377. doi:10.1037/h0048071
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluw.
- Hare, R. D. (2003). *Manual for the revised psychopathy checklist* (2nd ed.). Multi-Health Systems.
- Huang, Q., & Bolt, D. M. (2023). Unipolar IRT and the author recognition test (ART). *Behavior Research Methods*, 56(6), 5406–5423. doi:10.3758/s13428-023-02275-2
- Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment*, 21(1), 28–41. doi:10.1177/1073191113514105
- Kajonius, P. J., Persson, B. N., Rosenberg, P., & Garcia, D. (2016). The (mis) measurement of the Dark Triad Dirty Dozen: Exploitation at the core of the scale. *PeerJ*, 4, e1748. doi:10.7717/peerj.1748
- Leenaars, L. S., & Rinaldi, C. M. (2015). Indirect victimization. In W. G. Jennings (Ed.), *The Encyclopedia of Crime and Punishment* (pp. 734–736). Wiley Blackwell. doi:10.1002/9781118519639.wbecpx180
- 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. doi:10.1007/s10862-014-9469-3
- Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, 38(1), 88–91. doi:10.3758/BF03192753
- Lucke, J. F. (2013). Positive trait item response models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, and C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 199–213). Springer.
- Lucke, J. F. (2015). Unipolar item response models. In S. P. Reise and D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 272–284). Routledge/Taylor & Francis Group. doi:10.4324/9781315736013
- Magnus, B. E., & Liu, Y. (2018). A zero-inflated Box-Cox normal unipolar item response model for measuring constructs of psychopathology. *Applied Psychological Measurement*, 42(7), 571–589. doi:10.1177/0146621618758291
- McDonald, R. P. (1982). Linear versus models in item response theory. *Applied Psychological Measurement*, 6(4), 379–396. doi:10.1177/014662168200600402
- Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, 100(4), 674–701. doi:10.1037/0033-295X.100.4.674

- Moffitt, T. E. (2003). Life-course-persistent and adolescence-limited antisocial behavior: A 10-year research review and a research agenda. In B. B. Lahey, T. E. Moffitt & A. Caspi (Eds.), *Causes of conduct disorder and juvenile delinquency* (pp. 49–75). Guilford Press.
- Morales-Vives, F., Cosi, S., Lorenzo-Seva, U., & Vigil-Colet, U. (2019). The Inventory of Callous-unemotional traits and Antisocial behaviour (INCA) for young people: Development and validation in a community sample. *Frontiers in Psychology, 10*, 713. doi:10.3389/fpsyg.2019.00713
- Morales-Vives, F., Ferrando, P. J., & Dueñas, J.-M. (2023). Should suicidal ideation be regarded as a dimension, a unipolar trait or a mixture? A model-based analysis at the score level. *Current Psychology, 42*(25), 21397–21411. doi:10.1007/s12144-022-03224-6
- Morales-Vives, F., Gómez-Herrera, M., & Vigil-Colet, A. (2020). INCA-M: Mexican adaptation of the inventory of callous-unemotional traits and antisocial behavior. *Frontiers in Psychology, 11*(753), 753. doi:10.3389/fpsyg.2020.00753
- Preston, O. C., Gillen, C. T., Anestis, J. C., Charles, N. E., & Barry, C. T. (2020). The validity of the personality assessment inventory—Adolescent in assessing callous-unemotional traits in at-risk adolescents. *Journal of Personality Assessment, 103*(1), 48–56. doi:10.1080/00223891.2019.1705462
- Reise, S. P., & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: Some empirical and conceptual issues and challenges. *Psychological Medicine, 46*(10), 2025–2039. doi:10.1017/S0033291716000520
- Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*(1), 45–58. doi:10.1177/014662169001400105
- Reise, S. P., Du, H., Wong, E. F., Hubbard, A. S., & Haviland, M. G. (2021). Matching IRT models to patient-reported outcomes constructs: The graded response and log-logistic models for scaling depression. *Psychometrika, 86*(3), 800–824. doi:10.1007/s11336-021-09802-0
- Reise, S. P., & Moore, T. M. (2012). An introduction to item response theory models and their application in the assessment of noncognitive traits. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 699–721). American Psychological Association. doi:10.1037/13619-037
- Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to addressing non-normal distributions in the application of IRT models to personality measures. *Journal of Personality Assessment, 100*(4), 363–374. doi:10.1080/00223891.2017.1381969
- 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. doi:10.1177/1073191109344153
- Ruiz-Pamies, M., Lorenzo-Seva, U., Morales-Vives, F., Cosi, S., & Vigil-Colet, A. (2014). I-DAQ: A new test to assess direct and indirect aggression free of response bias. *The Spanish Journal of Psychology, 17*(e41), E41. doi:10.1017/sjp.2014.43
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Psychometric Society. <http://www.psychometrika.org/journal/online/MN17.pdf>. doi:10.1007/BF03372160
- Sanz-García, A., Gesteira, C., Sanz, J., & García-Vera, M. P. (2021). Prevalence of psychopathy in the general adult population: A systematic review and meta-analysis. *Frontiers in Psychology, 12*, 661044. doi:10.3389/fpsyg.2021.661044
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.

- Thorndike, R. L. (1971). *Educational Measurement*. America Council on Education.
- Toldos, M. P. (2005). Sex and age differences in self-estimated physical, verbal and indirect aggression in Spanish adolescents. *Aggressive Behavior*, 31(1), 13–23. doi:10.1002/ab.20034
- van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, 118(2), 339–356. doi:10.1037/a0022749
- Warren, G. C., & Clabour, J. (2009). Relationship between psychopathy and indirect aggression use in a noncriminal population. *Aggressive Behavior*, 35(5), 408–421. doi:10.1002/ab.20317
- Webster, G. D., & Jonason, P. K. (2013). Putting the “IRT” in “Dirty”: Item response theory analyses of the Dark Triad Dirty Dozen—An efficient measure of narcissism, psychopathy, and Machiavellianism. *Personality and Individual Differences*, 54(2), 302–306. doi:10.1016/j.paid.2012.08.027
- Zuckerman, M. (2012). *Psychobiology of personality* (2nd ed.). Cambridge University Press. doi:10.1017/CBO9780511813733.006