A Multidimensional Continuous Response Model for Measuring Unipolar Traits

Abstract

Unipolar constructs are encountered in a variety of noncognitive measurement scenarios that include clinical and forensic assessments, symptoms checklists, addictive behaviours, and irrational beliefs among others. Furthermore, Item Response Theory (IRT) models intended for fitting and scoring measures of unipolar constructs, particularly Log-Logistic models, are fully developed at present, but they are limited to unidimensional structures. This paper proposes a novel multidimensional log-logistic IRT model intended for double-bounded continuous-response items that measure unipolar constructs. The chosen response format is a natural application, and is increasingly used, in the scenarios for which the model is intended. The proposed model is remarkably simple, has interesting properties and, at the structural level can be fitted by using linearizing transformations. Multidimensional item location and discrimination indices are developed, and procedures for fitting the model, scoring the respondents, and assessing conditional and marginal accuracy (including information curves) are proposed. Everything that is proposed has been implemented in fully available R program. The functioning of the model is illustrated by using an empirical example with the data of 371 undergraduate students who answered the Depression and Anxiety subscales of the *Brief Symptom Inventory 18* and also the *Rosenberg Self-Esteem Scale.* The results show the usefulness of the new model to adequately interpret unipolar variables, particularly in terms of the conditional reliability of trait estimates and external validity.

Keywords: unipolar traits, log-logistic unipolar model, continuous response format, multidimensional item response theory.

Unipolar or Positive trait models (Lucke 2014, 2015) are a family of item response theory (IRT) models initially proposed for measuring dimensional variables that (a) could be plausibly scaled to adopt only positive values and (b) have a rightly skewed distribution in the target population. This is the case, for example, of clinical constructs such as depression (Reise et al., 2021), addictions (Lucke, 2015), or suicidal ideation (Morales-Vives et al., 2023) when they are measured in community populations. In these examples, the lower end of the trait continuum merely involves the absence of the pathology (i.e. unipolarity). So, the positive scaling condition is conceptually meaningful (Lucke, 2015, Reise et al., 2021). And, as for the distributional assumption, the majority of individuals in community samples are expected to have absence of the pathology, or very low levels of symptomatology, while the few pathological cases will be expected to spread at the upper tail of the distribution with different levels of severity. In contrast, bipolar normal-range traits such as extraversion-introversion are expected to behave very differently. In this case, both ends of the trait continuum are equally meaningful and have comparable levels of variation; the zero point can be defined at the mean level (i.e. ambiversion), and the distribution in community populations can be plausibly expected to be (approximately) normal (Morales-Vives et al., 2023, Reise et al., 2021).

The three models proposed by Lucke were intended only for binary items and focused on addictive behaviors. Of them, the one that has received the most attention so far is the log-logistic response model (LL-RM), which, since initially proposed, has been expanded in two main directions. On the one hand, LL-RM versions have been proposed for modeling graded (LL-GRM; Reise et al. 2021) and continuous (LL-CRM; Ferrando et al., 2024) responses. On the other, a variety of application domains in which these models could be theoretically appropriate has been explored. These applications

include: clinical and psychiatric measures (Magnus & Liu, 2018, Reise et al. 2018, 2021), suicide ideation (Morales-Vives et al., 2023), superstitious beliefs (Ferrando et al., 2024), maladaptive traits that are still normal-range (Morales-Vives et al., in press), and cognitive constructs such as vocabulary knowledge (Huang et al., in press).

The developments summarized above have been solely concerned with single-trait measures, and we consider that multidimensional extensions of the existing LL-RMs would be of interest for both researchers and professionals. Many psychological instruments, particularly clinical measures, include subscales to assess different pathologies and disorders that can be plausibly modeled as unipolar. This is the case of questionnaires such as the *Brief Symptom Inventory 18* (BSI 18, Derogatis, 2001), explained below, which includes a depression subscale, an anxiety subscale and a somatic subscale. Similarly, there are instruments of this type that include different subscales aimed at assessing a general construct and in which each subscale relates to a different sub-trait or facet. For example, Morales-Vives et al. (in press) have recently provided evidence that the maladaptive personality trait callousness can be plausibly fitted using unipolar models. However, some questionnaires that assess callousness include other subscales which are also likely to behave similarly (e.g. Essau et al., 2006, Morales-Vives et al., 2019).

This article is a first-step towards a multidimensional extension of existing LL-RMs, in which a multidimensional version of the LL-CRM is proposed. The choice of a continuous-response model can be justified for substantive, psychometric and practical reasons. First, at the substantive level, the continuous format is considered to be a natural format for the type of personality and clinical traits to which the proposed model is intended (Bejar, 1977, Liu, 2024, Byrom et al., 2022). Second, from a Psychometric view, models for continuous responses have interesting properties, are generally more

workable and simpler than graded-response models and, for these reasons, a variety of approaches for modeling this type of items have been proposed (e.g. Liu, 2024, Noel & Dauvier, 2007). In the approach considered in this paper, in particular, the model we shall propose can be developed from the transformed item scores and fitted directly by using existing procedures. Furthermore, it can be used as a basis for further developing the graded-response (including binary response) extensions.

From a practical viewpoint, finally, a literature review clearly shows that this old type of format is on the raise, mainly because it is well adapted to computerized administration modalities and psychometric research conducted via Internet. In these settings, its main claimed practical advantages are that is easy to administer, is easily understood by participants, requires little motivation for completion, and provides at least as much discrimination and accuracy as its discrete alternatives (e.g. García-Perez, 2024; McCormack et al., 1988).

Aims

The present article has three main aims. The first is to extend the unidimensional LL-CRM to the multidimensional case. This extension is complete and includes: fitting the structural model and addressing model-data fit, developing multidimensional item response parameters, scoring respondents, and assessing conditional and marginal score accuracy. The emphasis, however, is more on understanding the rationale and functioning of the model than on developing complex estimation procedures.

The second aim of the article is instrumental. Everything that is proposed here has been implemented in a free and user-friendly R program available for the interested readers. The third aim, finally, is to illustrate how the model functions by using a real data example and also provide initial guidelines for its use.

Background and basic framework

The LL-CRM is explained in detail in Ferrando et al. (2024) and, in this section, we shall provide only a brief summary aimed at facilitating the understanding of the multidimensional development. In the next section, we shall provide an updated review of the structural part.

The LL-CRM is a nonlinear IRT model, intended for item scores that can be treated as double bounded-continuous in the (0,1) interval, and designed for measuring a unipolar trait assumed to have a lognormal distribution in the population of interest. While the implied item-trait regressions differ considerably from the typical ogives in standard IRT models, after appropriate transformations of both the item and the person parameters, the structural part of the LL-CRM can be made equivalent to the logistic version of Samejima's (1974) CRM. This equivalence has two main implications. First, the same as the CRM, the LL-CRM can be "linearized" by using further transformations, and fitted (at the structural level) as if it was a Factor-Analytic (FA) model (Bejar, 1977; Ferrando et al., 2024; Mellenbergh, 1994; Samejima, 1974). Second, because either model can be obtained as a transformation of the other, both are expected to attain the same degree of model-data fit when fitted to the same data.

The fact that the LL-CRM can be linearized lead to the authors that developed it to propose a simple (and robust) two-stage (calibration and scoring) limited-information conditioned fitting procedure (McDonald, 1982) that we shall also propose for the multidimensional extension here. So, in the first, calibration stage, a FA solution is fitted to the transformed item scores and goodness of model-data fit is assessed. Next, the FA item structural estimates are transformed to the LL item estimates. In the second, scoring stage, trait estimates, standard errors of measurement and conditional reliability

estimates are obtained for each respondent on the basis of the first-stage structural estimates

An updated review of the unidimensional case

Consider a test made up of *n* items with a double-bounded continuous format that measure a unipolar trait θ_U. The item scores are scaled to have values in the (0,1) interval and θ_U is assumed to follow a lognormal distribution with parameters $\mu_U = 0$ and $\sigma_U = 1$. The LL-C Item Response curve (IRC) of item *j* for fixed θ_U is:

$$E(X_j \mid \theta_U) = \frac{\alpha_j \theta_U^{\beta_j}}{1 + \alpha_j \theta_U^{\beta_j}}. \quad (1)$$

Were α_j and β_j are both assumed to be positive and are location and form-curvature parameters, respectively. The IRC (1) is a 0-1 scaled power function (e.g. Stevens, 1975): a downward concave curve whose slope tends to increase more strongly for trait values close to zero and flattens as θ_U increases. The α_j parameter has is an "easiness" item parameter (Ferrando et al., 2024, Lucke, 2015, Reise et al.,2021): other things constant, the higher α_j the higher the expected item score becomes. The interpretation of β_j is more complex and is discussed below.

Apart from the basic α_j and β_j parameters, it is of interest to derive additional transformed parameters that (a) have a simple substantive or intuitive interpretation with regards to the item properties and functioning, (b) are in agreement with the parameterization in common use in IRT modeling, and (c) can be meaningfully extended to the multidimensional case. We shall consider two types of transformed item parameters: location indicators and item discriminating power indicators.

A location parameter can be defined as the θ_U trait level at which the expected item score is 0.5 (Ferrando et al. 2024, Lucke, 2015, Reise et al., 2021). So, in the LL-CRM case, the location parameter is the trait level that corresponds to the midpoint of the item response scale.

$$\delta_j = \left(\frac{1}{\alpha_j}\right)^{\frac{1}{\beta_j}}. \tag{2}$$

In clinical applications, δ_j can be appropriately referred as "item severity" (Lucke, 2015, Reise et al., 2021). In others, the generic term "item difficulty" might be more appropriate. Here we shall use the two terms interchangeably.

In terms of δ_j, the IRC (1) can be written as:

$$E(X_j \mid \theta_U) = \frac{\left(\frac{\theta_U}{\delta_j}\right)^{\beta_j}}{1+\left(\frac{\theta_U}{\delta_j}\right)^{\beta_j}}. \tag{3}$$

Expressed in this way, the IRC is a "quotient" curve (e.g. Lord, 1975, Ramsay, 1989) in which the expected item score increases with θ_U at a rate that is inversely proportional to the difficulty.

Reference values for interpretating (2) can be derived from the properties of the lognormal (0,1) distribution (e.g. Aitchison & Brown, 1957). As an approximate reference, δj values below 1 would be interpreted as that the item is "easy" in the sense that it only takes a low trait level to reach the threshold in the item response scale. Values at the 1.65 (the lognormal mean) would already be interpreted as that the item is difficult,

and δj values greater than, say, 3.5 - 4 would be interpreted as that the item is extremely difficult (severe). In practical applications, most of the items are expected to be "difficult" (or expressing high severity).

We turn now to the item discriminating power. The slope of the IRC at each point of θ_U is

$$Slope(\theta_U) = (\frac{\beta_j}{\theta_U})E(X_j|\theta_U)(1 - E(X_j|\theta_U)) \quad (4)$$

which is similar to the typical slope of an ogive-type IRT model but divided by the θ_U value (see Ferrando et al., 2024). This result poses problems when defining a single item discrimination index, because the IRC in (1) has no longer a maximum slope at $\theta_U = \delta_j$. Rather, the slope increases without bound as θ_U approaches zero. To address this issue, we can start by defining β_j as an overall item discriminability index. Second, by using an analogy to the quartile approach (Lucke, 2014) we can report the slopes at the trait levels at which the expected item scores are 0.25, 0.50 (i.e. $\theta_U = \delta_j$) and 0.75 as a summary of the discriminating power the item has at different levels. As for interpretation, based on preliminary results, and, tentatively, we may consider β_j values about between 0.3 and 0.7 to be normal range while values above 1 would already indicate a high level of overall discriminability.

The multidimensional proposal: the Md-LL-CRM

Consider now that the *n*-item test above measures a set of *r* unipolar traits: $\boldsymbol{\theta}_\mathbf{U} = \theta_{U1}$,.. $\theta_{Uk} \ldots \theta_{Ur}$ each of them distributed as lognormal (*0*,*1*). In the model we propose, denoted as multidimensional LL-CRM (Md-LL-CRM), the expected item score for fixed $\boldsymbol{\theta}_\mathbf{U}$, is

$$E\left(X_j\middle|\boldsymbol{\theta}_{\mathbf{U}}\right)=\frac{\alpha_j{\theta_{U1}}^{\beta_{j1}}\ldots.{\theta_{Ur}}^{\beta_{jr}}}{1+\alpha_j{\theta_{U1}}^{\beta_{j1}}\ldots.{\theta_{Ur}}^{\beta_{jr}}}. \quad (5)$$

The item parameters α_j and β_{jk} are all assumed to be positive and determine the location and form of the Item Response Surface (IRS). For the simplest, bidimensional case, Figure S1 shows the IRS obtained with typical item values.

As can be derived from equation (5) and figure (S1), the Md-LL-CRM is a compensatory model, because any change in θ_{U1} can be compensated by a multiplicative change in θ_{U2} resulting in the same product value in the numerator of (5). Furthermore, the IRS is monotonically increasing in the expected score as the elements of $\boldsymbol{\theta}_U$ increase. However, the shape of the surface reflects the power functioning discussed above: strong increases near zero and a tendency to flatten as the values of the θ_{Uk} increase. Finally, the single α_j easiness item parameter retains the same meaning and interpretation that it has in the unidimensional case.

Reckase and co-workers (e.g. Reckase, 1985, Reckase & McKinley, 1991) proposed an IRT-based approach for obtaining multidimensional indices of item difficulty and discrimination that can be easily adapted to the type of models considered here (Ferrando, 2009). For didactic purposes, we shall describe the approach in the bidimensional case of Figure S1 starting by the multidimensional index of difficulty. The approach is two-step. First, the direction on the θ_{U1} θ_{U2} plane (see Figure S1) at which the slope of the IRF is maximal is obtained. Second, the multidimensional item difficulty is defined as the distance from the origin to the point at which the expected item score is 0.5. (i.e. the item response threshold).

In the Md-LL-CRM case, the direction cosines that describe the direction of maximum slope for the general solution in *r* factors are:

$$\cos w_{jk} = \sqrt{\frac{\beta_{jk}}{\Sigma_k \beta_{jk}}}. \tag{6}$$

The Multidimensional item difficulty is found to be

$$Md\_\delta_j = \left(\frac{1}{\alpha_j}\right)^{\frac{1}{\Sigma_k \beta_{jk}}} \left[\frac{\sqrt{\Sigma_k \beta_{jk}}}{\left(\prod_k \beta_{jk}^{\beta_{jk}}\right)^{\left(\frac{1}{2\Sigma_k \beta_{jk}}\right)}}\right]. \tag{7}$$

And, in terms of Md_δ_j, model (5) can be written in quotient form as:

$$E(X_j|\theta_{Uc}) = \frac{\left(\frac{\theta_{Uc}}{Md\_\delta_j}\right)^{\Sigma_k \beta_{jk}}}{1 + \left(\frac{\theta_{Uc}}{Md\_\delta_j}\right)^{\Sigma_k \beta_{jk}}} \tag{8}$$

where θ_{Uc} is the value the linear trait composite in the direction of maximum slope.

Finally, the orthogonal projections of *Md-*δ_j on the θ_{Uk} axes are:

$$\text{Proj}\,(\theta_{Uk}) = Md\_\delta_j \times \cos w_{jk}. \tag{9}$$

With regards to interpretation, first of all, the reference values for *Md_*δ_j are the same as those for the unidimensional δj discussed above. Second, when the structural parameters of a given item approach a simple-structure form (i.e. one substantial weight on one of the traits and the remaining weights approaching zero): (a) the *Md_*δ_j value in (7) approaches the unidimensional δj in (2), as it should be, (b) the direction of maximum slope approaches the axis of the relevant trait, and (c) the projection (9) approaches also δj (the direction cosine for the relevant factor approaches 1 while the remaining cosines approach zero). In the bidimensional, and also in the tridimensional cases, the results so far discussed can be graphically summarized by using a "difficulty" vector plot in which each item is represented by a vector whose coordinates are the projections (9) on the θ_{Uk} axes. So, the length of the vector is the multidimensional difficulty of the item, and the

direction of the vector in the plane reflects the complexity of the item (see Figure 1 below).

As Reckase et al. (1985) noted, if the multidimensional difficulty levels of different items are to be compared, the items must have (at least approximately) the same direction for the comparison to be meaningful. This result can be checked by inspecting the direction cosines (6) and, in the bidimensional or tridimensional case, the vector plot above. Again, this result highlights the advantages of obtaining a solution as "clean" (in the sense of approaching an Independent Cluster) structure as possible, as, in this case, the vectors tend to be aligned around the θ_{Uk} axes.

We turn now to the multidimensional discrimination. The slope of the IRF (5) in the direction of the linear trait composite θ_{Uc} at which this slope is maximal is found to be:

$$Slope(\theta_{Uc}) = \left(\frac{\sum_k \beta_{jk}}{\theta_{Uc}}\right) E\left(X_j \middle| \theta_{Uc}\right)\left(1 - E\left(X_j \middle| \theta_{Uc}\right)\right). \quad (10)$$

So, the multidimensional extensions are quite straightforward here. We can define $\sum_k \beta_{jk}$ as a multidimensional overall item discriminability index, and provide slope values at specified θ_{Uc} levels. Again, we have chosen the values at which the expected item scores are 0.25, 0.50 (i.e. $\theta_{Uc} = M\_\delta_j$) and 0.75 as a summary of the discriminating power this item has at different levels along the dimension of maximal slope. The interpretations of the obtained values are the same as those discussed in the unidimensional case. Also, if different items are to be compared in terms of multidimensional discrimination, they must have (at least approximately) the same direction for the comparison to be meaningful (Reckase & McKinley, 1991).

Fitting the Md-LL-C

Item Calibration

As in the unidimensional case (Ferrando et al., 2024), the linearization of the M-LL-C is achieved by transforming the bounded (0-1) item scores into a real line by using the logit transformation in equation (5). The transformed conditional expectation becomes then:

$$ln(\frac{E(X_j|\mathbf{\theta_U})}{1-E(X_j|\mathbf{\theta_U})}) = ln(\alpha_j) + \beta_{j1}\, ln(\theta_{U1}).. + \beta_{jr}\, ln(\theta_{Ur})..$$
$$= \mu_j + \beta_{j1}\theta_{B1} \ldots \beta_{jr}\theta_{Br}. \quad (11)$$

So, expression (11) is the conditional expectation of a multiple-factor linear FA solution applied to the logit-transformed scores (e.g. Mellenbergh, 1994). Now, if the log-transformed θ_B traits were considered as if they were "natural" bipolar traits, then (11) would be the linearized expression of Samejima's (1974) multidimensional logistic Md-CRM (e.g. Bejar, 1977, Wang & Zeng, 1998). Here, however, the original traits in their "natural" scale are assumed to be the unipolar θ_U traits. So, the linearization (11) is double: First, the unipolar trait levels are log transformed, and second the linearized Md-CRM is proposed as the model for the log transformed θ_B trait levels.

Structural estimation (i.e. item calibration) of the Md-LL-C uses result (11) and consists simply of first fitting a standard multiple FA solution to the logit-transformed item scores and next re-parameterizing the item estimates that need to be transformed. In more detail, the original item scores in the (0-1) interval are first logit transformed (in order to obtain finite transformed scores, the lowest and highest endpoints of the original scores are set at .01 and .99, respectively), and next, the prescribed multiple FA solution is fitted to the transformed item scores. The resulting item structural estimates

are: the intercepts ($\mu_{j)}$, the loadings (β_{jk}), and the residual variances ($\sigma^2_{\varepsilon_j}$; which are only used at the scoring stage). The easiness parameters are next obtained as: $\alpha_j = \exp(\mu_j)$ and the multidimensional indices proposed in equations (6) to (10) can be directly obtained from the α_j and β_{jk} estimates.

If the chosen structural estimation procedure allows so, standard errors and confidence intervals for the β_{jk} estimates will be directly provided in the calibration output. As for the $\alpha_j = \exp(\mu_j)$ estimates, confidence intervals can be readily obtained by the monotonic transformation approach (Raykov & Marcoulides, 2004) and the standard errors can be obtained by the delta method (e.g. Raykov & Marcoulides, 2004). Finally, the multidimensional indices (7) and (10), are complex functions of the basic item parameters and the best approach for obtaining confidence intervals and standard errors is, possibly, to use simulation or resampling approaches.

Individual Scoring, Measures of Score Accuracy, and Empirical Information Curves

The scoring procedure we propose for the LL-CRM is Bayes expected a posteriori (EAP, Bock & Mislevy, 1982). As it is well known, EAP scoring allows finite estimates that do not drift toward implausible values to be obtained for each respondent in each of the traits. Furthermore, EAP estimation has here the advantage that the additional prior information is not arbitrary but based on the model assumptions.

EAP score estimation for Md-LL-CRM solutions is standard (e.g. Bock & Mislevy, 1982) and is a straightforward extension of the unidimensional case described in Ferrando et al. (2024). For each trait, the prior is set to be lognormal ($\mu_U = 0$,$\sigma_U = 1$) and the quadrature approximations use 100 nodes, which ensures accurate point

estimates and posterior standard deviations (PSDs) to be obtained (e.g. Bock & Mislevy, 1982). Simulation checks of the accuracy of the scoring procedure are provided in S2.

We turn now to the score accuracy assessment. For each trait, conditional reliability score estimates at a given trait level are obtained as

$$\rho_{\hat{\theta}_{Uki(EAP)}} = 1 - \frac{PSD(\hat{\theta}_{Uki})^2}{Var(\theta_{Uk})}. \quad (12)$$

where $Var(\theta_{Uk})$ is the variance of the prior lognormal distribution. If the conditional estimates (12) are plotted as a function of the estimated trait levels, an empirical Information Curve (EIC) in reliability metric is obtained for each of the traits that are measured (see Figure 2 below). The relevance of these EICs is discussed below.

An EAP-based marginal reliability estimate for each measured trait can finally be obtained as:

$$\rho_{k_{(EAP)}} = 1 - \frac{E(PSD(\hat{\theta}_{Uki})^2)}{Var(\theta_{Uk})}. \quad (13)$$

Indicator (13) is a useful auxiliar index of overall score accuracy but has to be interpreted together with the corresponding EICs.

Implementation

The proposal so far discussed has been implemented as R script called MULTIPOL, which has been developed in R Version 4.4.1 and runs with R versions more recent than 3.5.0. A description of this program can be found in Supplemental

Material S2. The program can be downloaded from:

https://psico.fcep.urv.cat/utilitats/MULTIPOL-C/index.php

Substantive and practical considerations for using the M-LL-C Model

The Md-LL-CRM is intended to be used in applications in which its assumptions and functioning are theoretically consistent and plausible. This means first that the constructs or traits under study can be conceptually considered as unipolar, i.e. the low end is not a conceptual opposite to the upper end but merely reflects absence of trait manifestations. And second, that the latent traits distributions can be more plausibly assumed to be rightly skewed than normal. Beyond that, the Md-LL-CRM is a demanding model that requires a rigorous test design: all the β_{jk} parameters have to be positive, and the closer the solution approaches an independent-cluster structure, the better will this solution work in terms of stability and interpretability (we accept, however, that a certain degree of item complexity is unavoidable in realistic applications and it is for this reason that we have developed the item multidimensional indices). To sum up, we believe that the use of the model should be, first and above all, derived and supported from the theory, and second, based on a good and solid design.

From a more practical point of view, fitting the structural part of the Md-LL-CRM consists on transforming the observed (0,1) bounded item scores to logits, thus converting them to continuous-unbounded, and then fitting the linear FA model to the transformed scores. If the Md-LL-CRM was correct and linear FA model was directly fitted to the untransformed scores, two types of well-known distortions would be expected to appear: (a) spurious evidence of multidimensionality, in the form of additional curvature factors, and (b) differentially attenuated item discrimination estimates (e.g. McDonald, 1982). However, if the item distributions were not too

extreme and the item discriminations not too high, the impact of the distortions at the structural level would be, possibly, relatively low. This means that the fit of the "direct" linear model would be still acceptable and the rank order of the item parameter estimates essentially the same (Ferrando, 2009).

Where substantial differences would be found between the present proposal and the "direct" application of the linear model would be at the scoring stage. The change of the trait scaling from normal to lognormal (see equation 11) would imply that the trait estimates would be compressed at the lower end and expanded at the upper end. So, although the rank ordering of the scores obtained from both approaches would be preserved, their relation would become non-linear. More important: the EICs derived from both approaches would be diametrically opposed. In the approach we propose, maximal conditional reliability would be attained at low trait levels (see equation 10, and Figure 2 below). So, the scores would be particularly appropriate for differentiating individuals who have no or barely any trait manifestations from those who clearly do have them. In contrast, in the standard modeling higher accuracy will be reached at high trait levels. So, if the Md-LL-CRM was the most theoretically appropriate model, the use of the standard model would be expected to lead to score-related misleading conclusions, and this would have implications for such relevant issues as deciding where to set cut-off points to differentiate between subjects with and without a condition or disorder (e.g., Morales-Vives et al., 2023; Morales-Vives et al., in press).

We turn finally to the assessment of model appropriateness. We consider this assessment has to be comprehensive and multifaceted, and go far beyond assessing the degree of model-data fit. This point is particularly relevant here because, as mentioned above, the Md-LL-CRM is indistinguishable from the Md-CRM in pure Goodness-of-fit terms even when both models function very differently. Detailed discussions on this

issue have been provided in Ferrando et al. (2024) and Reise et al. (2018, 2021). So, they will be discussed in the example below providing only a summary here. Beyond the theoretical consistency discussed above: (a) the distributions of the raw item and scale scores should generally be right skewed, with a sizable number of cases piled-up at the lower end; (b) the empirical standard errors of measurement based on the raw scores are expected to increase with the scores; and (c) in terms of external-validity evidence, the relation with the external variables is expected to be stronger at its upper end (because unipolar traits are more meaningful at their upper pole).

Empirical Study

For this empirical study, we used the data of 371 undergraduate students (84.2% women) with ages between 18 and 42 (M = 20.56, SD = 4.14), which were administered the *Brief Symptom Inventory 18* (BSI 18, Derogatis, 2001). For this example, specifically, we used their data in the Anxiety (6 items) and Depression (6 items) subscales. Participants were asked to rate each item on a continuous response format, indicating the extent to which they suffered each symptom (tense, scared, hopelessness, etc.). We also administered the *Rosenberg Self-Esteem Scale* (RSE, Rosenberg, 1979), which consists of 10 items with a five-point Likert response format (1 = Completely disagree, 5 = Completely agree).

Table 1 shows the medians and skewness coefficients for each BSI item (recall that the responses are scaled between 0 and 1). As can be seen, 8 out of 12 items have medians particularly low, with values below .20, and skewness coefficients higher than 1. So, most respondents scored very low in most of the items, giving rise to skewed score distributions (as expected when the Md-LL-CRM is appropriate). The exceptions

are items 3 and 6, because feeling nervous and tense is not so rare in undergraduate students, who are under pressure during their studies because of exams and other assessments they have to pass. Note, however, that the median values for both items are still below .50.

Item Calibration

Based on previous pilot studies, a structural, near IC solution with only two complex items, was fitted to the logit-transformed item scores using robust (mean and variance corrected) ML estimation as implemented in Mplus 8.11. Goodness of model-data fit (GOF) was inspected through several indices that assess different facets of fit: a) absolute fit, with the Goodness of Fit Index (GFI) and the Root Mean Square of Residuals (RMSR); b) comparative fit with respect to the null independence model, with the Comparative Fit Index (CFI), and c) relative fit per degree of freedom, with the Root Mean Square Error of Approximation (RMSEA). The following estimates were obtained: RMSEA = 0.050, 90% CI [0.034, 0.065]; CFI = 0.96; GFI = 0.98; SRMR = 0.38. Overall, the fit can be considered as quite acceptable. Because the Md-LL-CRM and the Md-CRM are indistinguishable in pure GOF terms, however, these results only tell us that both models are tenable at the structural level. The item parameter point estimates are in Table 1 (standard errors for columns 5-7 are available from the authors) and the corresponding vector difficulty plot is in Figure 1. The estimated correlation between both unipolar lognormal traits (e.g. Zerovnik et al., 2013) was $\varphi = .60$.

INSERT FIGURE 1 AND TABLE 1ABOUT HERE

Starting with the overall features in Table 1, it seems clear that most of the items are both extremely difficult (in terms of α and Md_δ) and rather discriminating (overall discriminations near or above 1). Note that, for the items that follow this trend, the slope

is only nonnegligible at the lowest evaluation point (0.25) and then flattens almost completely. This is, indeed, the expected behavior of the IRS in items that are in accordance to the unipolar modeling

Certain items in Table 1, however, depart somewhat from this general trend. Items 9, 12, and 17 are extremely difficult but their slope is rather flat from the beginning. So, in graphical terms, the IRSs of these items have a low initial slope which becomes practically flat before reaching the 0.5 threshold. Their extreme *Md_δ* values, which have also been obtained in previous LL-M applications (e.g. Luke, 2014, Table 13.2), reflect these features and suggests that the symptoms or behaviors the items refer to are extremely rare in the target population. This was already expected, because item 17 refers to suicidal ideation, item 12 refers to outbreaks of terror or panic, and item 9 refers to sudden and unreasonable fears, which are not very common symptoms in the population, in comparison with others less extreme such as feeling tense (item 6), nervousness (item 3) or feeling blue (item 8).

The vector plot in Figure 1 reflects the remarkable simplicity of the solution, and most items are aligned along their corresponding axes with different lengths. For scaling reasons, the most difficult items (9, 12, and 17) exceed the graph rule (their lines do not have the final arrow head). As for the two complex items, item 8 is located at 29º from the first axis and 61º from the second. So, it would be a more saturated measure of Depression than of Anxiety. It is also rather easy. In contrast, item 18 is considerably more difficult, and is located at 59º of the Depression axe and at 31º of the Anxiety axe. So, it is a more saturated measure of anxiety.

We turn now to the scoring-related results. For both Depression (dashed line) and Anxiety (solid line) score estimates, Figure 2 shows the conditional reliabilities

obtained with equation (12) against the EAP point estimates. The two curves are very similar, with the trait estimates being highly reliable between 0 and 2 units for Anxiety and between 0 and 2.5 units for Depression, and with reliability levels dropping significantly above these values, especially for Anxiety. Although there is a sharp decline in reliability in both cases, it appears that reliability levels remain somewhat higher for Depression than for Anxiety. However, in both cases reliability levels are quite low at medium and high trait levels. The marginal reliabilities are consistent with these findings, with a higher estimated value for Depression (.87) than for Anxiety (.82). The conditional reliability curves suggest that the score estimates are useful for distinguishing between people who do not suffer from anxiety or depression and those who do. However, they are not as useful for identifying different levels of severity among those who do suffer from these disorders.

INSERT FIGURE 2 ABOUT HERE

Finally, in this example we shall consider "external" evidence of the appropriateness of the unipolar solution in terms of the validity relations with a relevant external variable, more specifically Self-esteem, which was assessed with the RSE questionnaire. As discussed above, we expect the unipolar scores to become increasingly predictive as they increase, which implies a heteroscedastic relation in which the scatter of points around the regression line is more dispersed at the lower end of the trait continuum than at the upper end. Figure 3 shows the bivariate regressions of the Self-esteem scores against the estimated Depression scores (panel a) and Anxiety scores (panel b). The regression lines were fitted using kernel-smoothed nonparametric regression; specifically, the Nadaraya-Watson kernel estimator (e.g. Härdle, 1990).

The results in Figure 3 (panel a) suggests that the regression of the Self-esteem scores on the Depression scores is nonlinear. The correlation between the (kernel) predicted and observed criterion scores was $r = .66$, which is substantial. More important here, however, is that the expected heteroscedasticity relation seems evident in the graphic. The formal test of this hypothesis (e.g. Cohen, 1988) gave a significant and negative correlation between the absolute regression residuals and the estimated trait levels: $r = -.22$ ($p < .05$). Although the graph shows that most people without depression tend to have higher self-esteem than those with depression, this does not necessarily involve that all people without depression have a very high self-esteem. For this reason, a wide spread of self-esteem scores at very low levels of depressive symptomatology would be expected. In contrast, people with high levels of depressive symptomatology tend to have low levels of self-esteem, as also expected. These findings seem to be consistent with the vulnerability theory (Beck, 1967), which suggests that low self-esteem contributes to depression, although this relationship also depends on other variables, as for example rumination (e.g., Orth & Robins, 2013). Therefore, low self-esteem alone is not sufficient to develop a depressive state, and for this reason there is a greater dispersion observed at low levels of depressive symptomatology.

With regards to the Anxiety results in Figure 3 (panel b), the trends just discussed are observed again but in a more moderate way. The relation is more linear here, the predicted-observed correlation is $r = .46$, and the heteroscedasticity test is again significant: $r = -.20$ ($p < .05$). Previous studies have also shown a relationship between anxiety and low self-esteem (e.g., Sowislo & Orth, 2013). However, the results of the current study suggest that low levels of anxiety are not necessarily associated with very high levels of self-esteem, which leads to a greater dispersion in self-esteem

scores for low levels of anxiety. In contrast, people with high levels of anxiety tend to have low levels of self-esteem. These differential levels of dispersion in self-esteem at low and high levels of anxiety follow a similar trend to that observed for depressive symptomatology.

INSERT FIGURE 3 ABOUT HERE

Discussion

While the distinction between bipolar and unipolar constructs is well established in non-cognitive measurement (e.g. Ferrando, et al., in press; Tay & Jebb, 2018), it does not appear to have substantially affected the way in which instruments that measure one or another of these two types of constructs are psychometrically modelled. To a certain extent, this state of affairs is understandable, because for most normal-range traits, the standard models, which are based on bipolarity assumptions, are expected to work reasonably well for both types. In other cases, however, routine application of these models is more questionable. In particular, most clinical variables are only meaningful and well defined at their upper pole. Furthermore, when measures of these variables are administered in community samples (probably the most common application), the "default" normality assumption regarding the distribution of the construct becomes untenable. In these cases, the use of models like that proposed here is, possibly, a better option. In addition, recent research suggests that there is a non-negligible array of variables beyond clinical scales that could be more appropriately modelled with unipolar models (e.g., Ferrando et al., 2024; Huang & Bolt, 2023; Huang et al., in press; Morales-Vives et al., in press).

So far, the development of unipolar IRT models, particularly Log-logistic models, has been limited to unidimensional structures, and the need to extend them to

the multidimensional space seems to be clear. This paper is a first step in this direction and proposes a multidimensional log-logistic unipolar model for double-bounded continuous item responses. The model is relatively simple and enjoys interesting psychometric properties. However, as discussed above, the choice of a continuous format is not solely justified for these reasons.

The Md-LL-CRM proposed here is a natural and plausible extension of the existing unidimensional LL-CRM. To start with, at the conceptual level, the proposed multidimensional item indices are meaningful extensions of their unidimensional counterparts, and reduce to them when there is only a single dimension (as it should be). In terms of model fitting, the Md-LL-CRM retains the logit-linearization property of the LL-CRM, which means that, at the structural level, it can be fitted as if it was a transformed multiple FA model. So, available well-known standard structural modeling procedures can be used for fitting the extended (mean and covariance) structure and for assessing model data fit. The scoring stage finally is also a straightforward extension of the unidimensional case. Overall, we would like to stress that the modeling proposal as well as the calibration and scoring procedures have been purposely kept as simple and robust as possible.

The empirical study carried out illustrates how the Md-LL-CRM behaves when used on real data, with two plausibly unipolar variables: Depression and anxiety. The results are equivalent to those obtained in previous studies with the unidimensional models (e.g. Ferrando et al., 2024; Morales et al., in press; Morales-Vives et al., 2023). While there are no differences between the unipolar and the standard bipolar models at the goodness of structural fit, the behavior of the items is entirely consistent with what is expected in the proposed model, with many items having significantly skewed score distributions and extreme values of difficulty and discrimination. Similarly, the results

at the conditional reliability level are similar to those previously obtained. Specifically, maximum values of conditional reliability are attained at low levels of trait scores, with conditional reliability decreasing abruptly at medium and high levels for both in depression and anxiety. Therefore, greater precision is achieved in the estimation of scores at low trait levels, which makes it possible to discriminate adequately between subjects who do not suffer from these conditions and those who do. In return, scores are not so accurate for differentiating among different high trait levels, being this tendency even more pronounced for anxiety than for depression.

The results regarding external validity are also equivalent to those obtained in previous unidimensional studies. In this case, there is a greater dispersion of self-esteem scores for subjects with low levels of anxiety and depression than for subjects who do present these emotional problems. Similar results were obtained regarding the relationship between the unipolar variable suicidal ideation and life satisfaction (Morales-Vives et al., 2023), or between the unipolar variable callousness and the variables indirect aggressiveness and non-planning impulsiveness (Morales-Vives et al., in press). To sum up, the empirical example of the present study provides evidence on the performance of the Md-LL-CRM model with real data, revealing results that were theoretically expected and in agreement with those obtained in previous studies with variables of a similar nature.

As any novel, wide-scope, proposal, this one has its share of limitations, issues that can be made more complete, and further possibilities of extension. Thus, starting from the fitting structural procedures, the simple limited-information two-stage approach we propose should be extended to incorporate standard errors and confidence intervals for the multidimensional indices developed here. More generally, alternative full-information estimation procedures could be considered. Having said that, however,

it is not clear to us that these far more complex procedures will be clearly superior in practice to that we propose here (e.g. McDonald, 1982). In any case, their theoretical advantages make them worth trying.

At the scoring level, the extensions we consider of more interest are two. First, is to develop procedures for estimating the unipolar latent trait distributions. At present, the latent priors are set to be lognormal (0,1). However, if the latent densities could be also estimated from the data, this feature would provide more flexibility to the modeling, as the priors could be modified to have different degrees of skewness. The second extension would consist of developing person-fit indices tailored for this type of model (in which most of the scores undifferentiated at the lower end and a few scores that can attain very high values).

Any new proposal like this only has a real chance of being used in practice if it is implemented in a simple, user-friendly and preferably free program. Our program is a first step in this direction. However, we plan to further develop it in the future and end up with a complete R application able to fit the model directly from the raw data.

To sum up, the Md-LL-CRM model is an original proposal that fills a gap in the study of unipolar constructs. Despite its limitations, it allows, for the first time, to adequately analyze potentially unipolar variables assessed through multidimensional instruments, which is the most common situation in the field of social sciences. Therefore, this article opens up a new range of possibilities and may be the beginning of new research in this field. It may also lead to the development of new models that respond to the needs of both practitioners and researchers, as, for example, new models for binary and graded-response items.

References

Aitchison, J., & Brown, J.A.C. (1957). *The Lognormal Distribution.* Cambridge University Press.

Beck A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects.* Harper & Row.

Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement*, *1*(4), 509-521. https://doi.org/10.1177/014662167700100407

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. https://doi.org/10.1177/014662168200600405

Byrom, B., Elash, C. A., Eremenco, S., Bodart, S., Muehlhausen, W., Platko, J. V., Watson, C., Howry, C. (2022). Measurement comparability of electronic and paper administration of visual analogue scales: A review of published studies. *Therapeutic Innovation & Regulatory Science, 56*(3), 394-404. https://doi.org/10.1007/s43441-022-00376-2

Cohen, J. (1988). Statistical power analysis for the behavioral sciences.

Derogatis, L. R. (2001). *BSI 18, Brief Symptom Inventory 18: Administration, scoring and procedures manual*. NCSPearson, Inc.

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

Ferrando, P.J. (2009). Difficulty, discrimination, and information indices in the linear factor analysis model for continuous item responses. *Applied Psychological Measurement*, *33*(1), 9-24. https://doi.org/10.1177/0146621608314608

Ferrando, P. J., Morales-Vives, F., Casas, J. M., & Muñiz, J. (in press). Designing, constructing, and using Likert scales: A practical guide. *Psicothema*

Ferrando, P. J., Morales-Vives, F., Hernandez-Dorado, A. (2024). Measuring unipolar traits with continuous-response items: Some methodological and substantive developments. *Educational and Psychological Measurement*, *84*(3), 425-449. https://doi.org/10.1177/00131644231181889

García-Pérez, M. A. (2024). Are the steps on Likert scales equidistant? responses on visual analog scales allow estimating their distances. *Educational and Psychological Measurement*, *84*(1), 91-122. https://doi.org/10.1177/00131644231164316

Härdle, W. (1990). Applied nonparametric regression. London: Chapman & Hall.

Huang, Q., & Bolt, D. M. (2023). Unipolar IRT and the Author Recognition Test (ART). *Behavior Research Methods*, 1-18. https://doi.org/10.3758/s13428-023-02275-2

Huang, Q., Bolt, D. M., & Liao, X. (in press). Theory-Driven IRT Modeling of Vocabulary Development: Matthew Effects and the Case for Unipolar IRT. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12433

Liu, C. W. (2024). Multidimensional item response theory models for testlet-based doubly bounded data. *Behavior Research Methods*, *56*(6), 5309-5353. https://doi.org/10.3758/s13428-023-02272-5

Lord, F. M. (1975). The 'ability'scale in item characteristic curve theory. *Psychometrika*, *40*(2), 205-217. https://doi.org/10.1007/BF02291567

Lucke, J.F. (2014). 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. https://doi.org/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. https://doi.org/10.1177/0146621618758291

McCormack, H. M., David, J. D. L., & Sheather, S. (1988). Clinical applications of visual analogue scales: a critical review. *Psychological medicine*, *18*(4), 1007-1019. https://doi.org/10.1017/S0033291700009934

McDonald, R. P. (1982). Linear versus models in item response theory. *Applied Psychological Measurement*, *6*(4), 379-396. https://doi.org/10.1177/014662168200600402

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*(2), 300–307. https://doi.org/10.1037/0033-2909.115.2.300

Morales-Vives, F., Cosi, S., Lorenzo-Seva, U., & Vigil-Colet, A. (2019). The INventory of Callous-unemotional traits and Antisocial behavior (INCA) for young people: Development and validation in a community sample. *Frontiers in Psychology*, *10*, 1-12. https://doi.org/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*, 1-15. https://doi.org/10.1007/s12144-022-03224-6

Morales-Vives, F., Ferrando, P. J., Hernandez-Dorado, A. (in press). Modelling maladaptive personality traits with unipolar item response theory: The case of Callousness. *The Journal of General Psychology*. http://dx.doi.org/10.1080/00221309.2024.2404398

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*(1), 47-73. https://doi.org/10.1177/0146621605287691

Orth, U., & Robins, R. W. (2013). Understanding the link between low self-esteem and depression. *Current Directions in Psychological Science*, *22*(6), 455-460. https://doi.org/10.1177/0963721413492763

Ramsay, J.O. (1989). A comparison of three simple test theory models. *Psychometrika*, *54*(3), 487-499. https://doi.org/10.1007/BF02294631

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, *11*(4), 621-637. https://doi.org/10.1207/s15328007sem1104_7

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied psychological measurement*, *9*(4), 401-412. https://doi.org/10.1177/014662168500900409

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied psychological measurement*, *15*(4), 361-373. https://doi.org/10.1177/014662169101500407

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. https://doi.org/10.1007/s11336-021-09802-0

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*, 363–374. https://doi.org/10.1080/00223891.2017.1381969

Rosenberg, M. (1979). *Conceiving the Self*. Basic Books.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39, 111-121. https://doi.org/10.1007/BF02291580

Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological Bulletin*, *139*(1), 213-240. https://doi.org/10.1037/a0028931

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Transaction Publishers.

Tay, L., & Jebb, A. T. (2018). Establishing construct continua in construct validation: The process of continuum specification. *Advances in Methods and Practices in Psychological Science*, *1*(3), 375-388. https://doi.org/10.1177/2515245918775707

Wang, T., & Zeng, L. (1998). Item parameter estimation for a continuous response model using an EM algorithm. *Applied Psychological Measurement*, *22*(4), 333-344. https://doi.org/10.1177/014662169802200402

Žerovnik, G., Trkov, A., Smith, D. L., & Capote, R. (2013). Transformation of correlation coefficients between normal and lognormal distribution and implications for nuclear applications. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *727*, 33-39. https://doi.org/10.1016/j.nima.2013.06.025

Table 1

Median, Skewness and item parameter estimates for the 12 BSI items

	Item number	Median	Skewness	α	$\beta_1(Dep)$	$\beta_2(Anx)$	Overall Disc	Md_δ	Slope (0.25)	Slope (0.50)	Slope (0.75)
Depression	2. No interest	.29	0.65	0.35	0.87	0	0.87	3.40	0.17	0.06	0.01
	5. Lonely	.17	1.01	0.18	1.37	0	1.37	3.44	0.17	0.10	0.03
	8. Blue	.36	0.32	0.48	1.04	0.25	1.29	2.27	0.25	0.14	0.04
	11. Worthlessness	.11	1.50	0.10	1.66	0	1.66	4.04	0.15	0.10	0.02
	14. Hopelessness	.10	1.38	0.11	1.39	0	1.39	4.84	0.12	0.07	0.01
	17. Suicide	.01	5.20	0.02	0.63	0	0.63	90.00	0.01	0.00	0.00
Anxiety	3. Nervousness	.41	0.24	0.57	0	1.22	1.22	1.57	0.36	0.20	0.06
	6. Tense	.46	0.09	0.71	0	1.17	1.17	1.34	0.42	0.22	0.06
	9. Scared	.06	1.95	0.07	0.91	0	0.91	20.86	0.03	0.01	0.00
	12. Panic	.02	2.34	0.04	0	1.17	1.17	15.68	0.04	0.02	0.00
	15. Restlessness	.12	1.12	0.14	0	1.00	1.00	7.35	0.08	0.03	0.00
	18. Fearful	.13	1.29	0.14	0.37	0.93	1.30	6.30	0.10	0.05	0.01

Note. The text of the items is presented in summary form.

Figure 1

Difficulty vector plot for the 12 BSI items

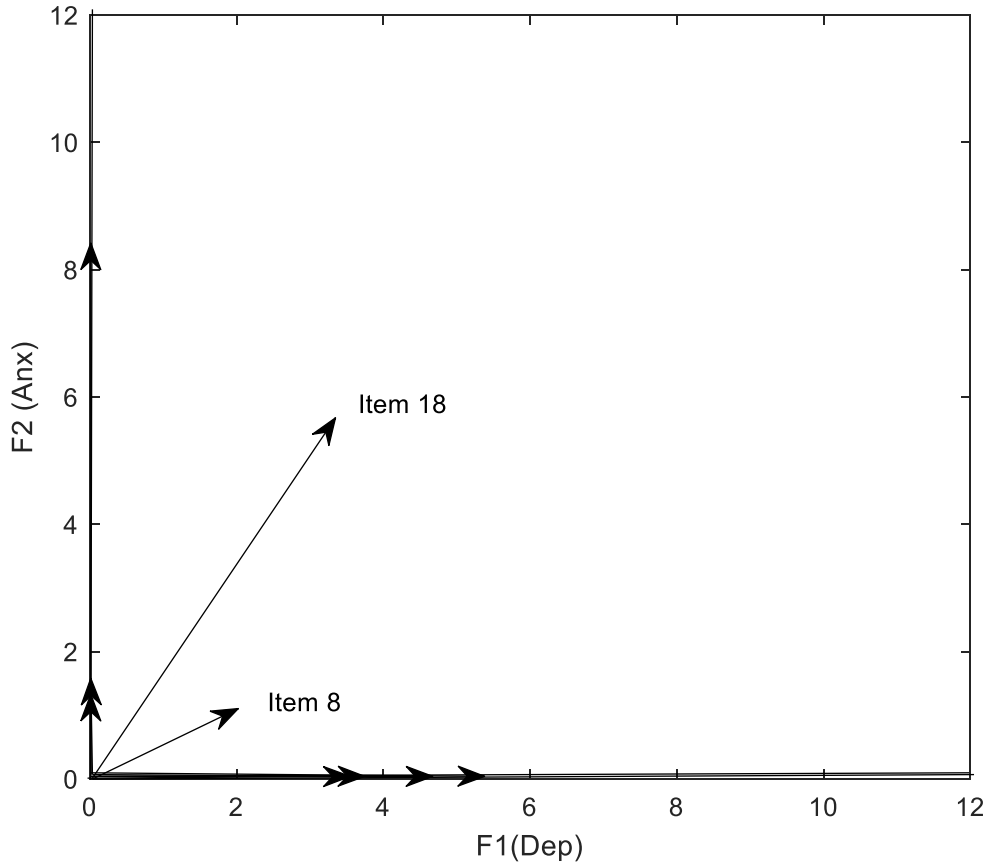


Figure 2

Empirical Information Curves for the Depression and Anxiety BSI score estimates

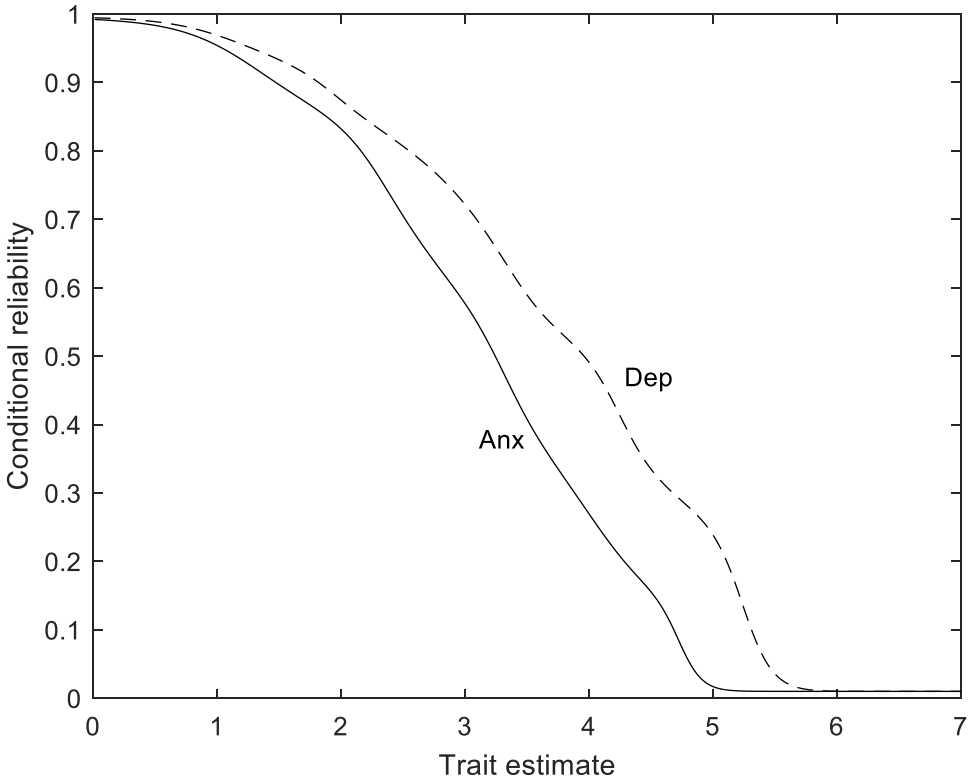
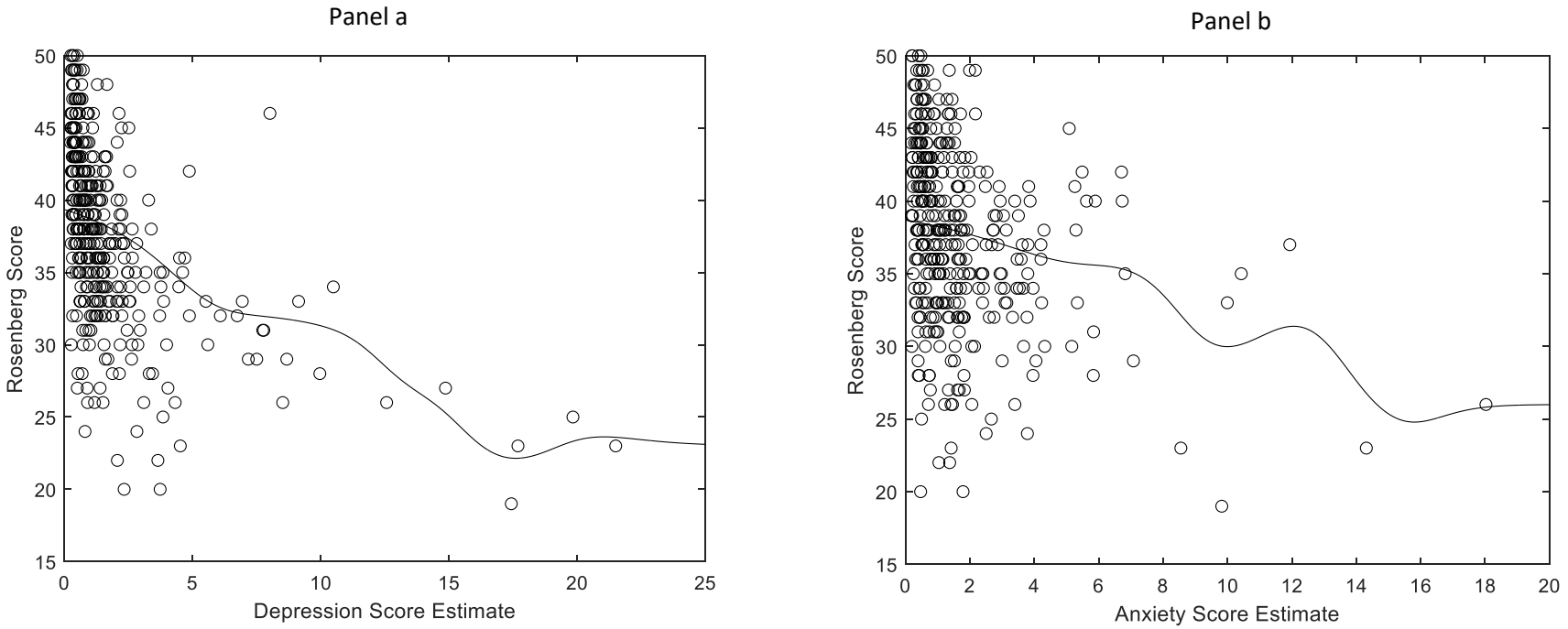


Figure 3

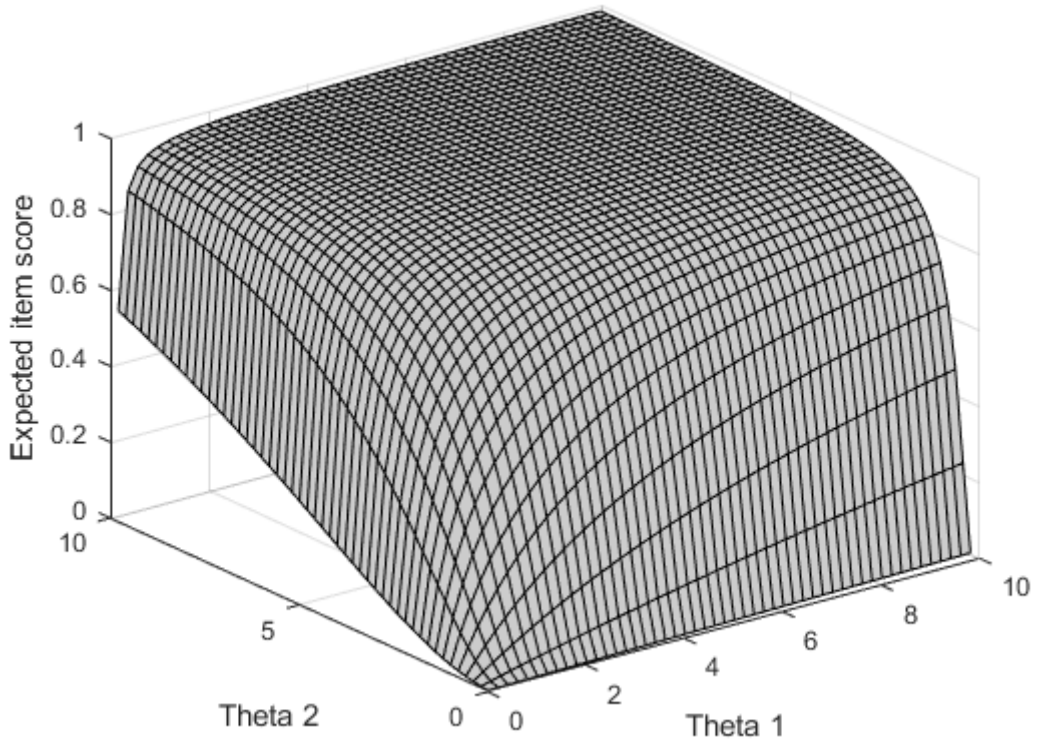
Validity evidence: Rosenberg self-esteem scores against BSI Depression Unipolar score estimates (panel a) and against Anxiety Unipolar score estimates (panel b)



Supplemental Material S1

Figure S1

Item Response Surface of the Md-LL-CRM



Supplemental Material S2

The MULTIPOL R script uses as input the untransformed data in 0-1 format, as well as the structural results based on the FA linearization. Most precisely, it requires the covariance-based factor loadings obtained from fitting the linearized covariance structure in equation (11), as well as the inter-factor correlation matrix.

The program is released in a compressed folder, which includes two main R scripts: MULTIPOL and MULTIPOL_GUI. The first script, MULTIPOL, was designed for advanced R users, where the input values have to be provided through code and the output is also printed in the R console or returned silently. When using this function, the user must source the function before its utilization:

```
>source("MULTIPOL.R")
```

The second script, MULTIPOL_GUI, is a shiny app (Chang et al. 2024), which provides a Graphical User Interface version of the script, more suitable for users less familiar with R language. It requires shiny package to be installed.

Finally, a data folder is also provided, containing an example dataset as well as all the required input data for using MULTIPOL.

The program can be downloaded from:
https://psico.fcep.urv.cat/utilitats/MULTIPOL-C/index.php

The structural input information required as input in MULTIPOL can be obtained with any of the existing commercial and non-commercial structural modeling programs that allow a restricted or an unrestricted solution to be estimated. So, it does

not make much sense to perform simulation studies on the structural or calibration part of the model data fitting process.

EAP score estimation is also standard, but, in this case, it has been fully implemented in our program. So, the functioning of this part needs to be checked. To do so, we generated data, for which the Md-LL-CRM held, in a variety of settings in which both the structural and the individual (i.e. true unipolar trait levels) parameters were known. In all cases, the relevant estimates: individual EAP estimates, conditional reliabilities, and marginal reliabilities, appropriately recovered their corresponding parameters within the limits of sampling error. More specifically, we provide below the results of one of the simulations checks we undertook.

Pseudo-Samples of N = 500 were drawn from true population solutions under three conditions: C1; two unipolar traits and 6 indicators per trait; C2; three unipolar traits and 7 indicators per trait, and C3 four unipolar traits and 8 indicators per trait. In all cases, each indicator had a salient (discrimination) β value of 1.0 on the trait it indicated, and a secondary β value of 0.4 on the remaining traits. In all cases, the traits were orthogonal. As for the α_j values, they were set to 1.20 for all items. We shall note that the specified values are realistic and agree with those expected in a typical application of the model.

From each condition, 25 replications were used, and the EAP estimates were obtained using 100 quadrature points, as discussed in the article. For each trait, he outcomes of interest in this study were: (1) the mean of the EAP estimates, (2) the standard deviation of the EAP estimates, and (3) the product-moment correlation between the EAP estimates and the corresponding 'true' trait values. The average results across replications are summarized in Table S2. Furthermore, for each reported

outcome, the corresponding expected result is shown next to it within brackets. These expected results were obtained using Kelley's regression formulas (see Bock & Mislevy, 1982), and so, are marginal approximations. They proved, however, be quite accurate.

Table S2

EAP scoring. Results of the reported simulation study

		C1	C2	C3
		2 traits with 6 indicators each	3 traits with 7 indicators each	4 traits with 8 indicators each
T1	$mean\ \hat{\theta}\ (expected)$	1.62 (1.64)	1.64 (1.64)	1.65 (1.64)
	$std\ \hat{\theta}\ (expected)$	1.67 (1.74)	1.70 (1.73)	1.72 (1.72)
	$r\ (\hat{\theta}, \theta)\ (expected)$	0.90 (091)	0.91 (0.92)	0.92 (0.93)
T2	$mean\ (\hat{\theta})(expected)$	1.65 (1.64)	1.65 (1.65)	1.65 (1.64)
	$std\ (\hat{\theta})\ (expected)$	1.68 (1.74)	1.75 (1.73)	1.78 (1.72)
	$r\ (\hat{\theta}, \theta)\ (expected)$	0.90 (0.91)	0.91 (0.92)	0.92 (0.93)
T3	$mean\ (\hat{\theta})(expected)$		1.65 (1.64)	1.64 (1.64)
	$std\ \hat{\theta}\ (expected)$	-	1.74 (1.73)	1.74 (1.72)
	$r\ (\hat{\theta}, \theta)\ (expected)$		0.91 (0.92)	0.92 (0.93)
T4	$mean\ (\hat{\theta})(expected)$			1.64 (1.64)
	$std\ \hat{\theta}\ (expected)$	-	-	1.74 (1.73)
	$r\ (\hat{\theta}, \theta)\ (expected)$			0.92 (0.93)

Results in Table S2 are clear. In all cases the obtained results are quite close to their expectations without systematic trends in the discrepancies that could indicate malfunction or bias.

References in S2

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2025). *shiny: Web Application Framework for R. R package version 1.10.0.9000*. https://github.com/rstudio/shiny, https://shiny.posit.co/