METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

A Weibull-Link Response Model for Measuring Unipolar-Skewed Constructs With Continuous Responses

Pere J. Ferrando [1], Fabia Morales-Vives [1], José M. Casas [2], David Navarro-González [1]

[1] *Research Center for Behavioral Assessment, Departament de Psicologia, Universitat Rovira i Virgili, Tarragona, Spain.*
[2] *Departament de Psicologia, Universitat Rovira i Virgili, Tarragona, Spain.*

Methodology, 2026, Vol. 22(3), 225–251, https://doi.org/10.5964/meth.20097

Received: 2025-10-03 • **Accepted:** 2026-04-01 • **Published (VoR):** 2026-08-31

Handling Editor: Eduardo Estrada, Autonomous University of Madrid, Madrid, Spain

Corresponding Author: David Navarro-González, Universitat Rovira i Virgili, Research Center for Behavioral Assessment, Departament de Psicologia, Tarragona, Spain. Carretera de Valls s/n, Campus Sescelades, 43007 Tarragona, Spain. E-mail: david.navarro@urv.cat

Abstract

In certain noncognitive applications attempts are made to measure unipolar traits expected to have positively skewed distributions in the target populations, two conditions that clash head-on with the bipolarity and normality assumptions on which most standard IRT models are based. Unipolar-skewed models are an alternative to be used in these conditions, and we propose here a model intended for double-bounded continuous response items in which: (a) the link function or item response function is a cumulative Weibull curve, and (b) the initial latent distribution is lognormal. The model is simple, has desirable psychometric properties, and, conceptually, provides a plausible account of the response functioning. We propose a full development which has been also implemented in a free R program. Moreover, an empirical study is included, and its results show that the model behaves appropriately when used with real data, delivering the anticipated accuracy and external validity outcomes.

Keywords

unipolar traits, nonnormal latent distributions, asymmetric item response functions, Weibull link, continuous response format

The use of Item Response Theory (IRT) models in non-cognitive applications is nowadays a common practice. Not only are they increasingly used but also applied to a larger variety of constructs than those initially considered, i.e., normal-range personality traits and attitudes (e.g., Bejar, 1977; Reise & Waller, 1990; Thissen et al., 1983). Currently, IRT models are used in the measurement of health outcomes (e.g., Smits et al., 2020), quality

of life (e.g., Sijtsma & van der Ark, 2022), psychological and psychiatric measures (e.g., Reise et al., 2018, 2021), etc. To a large extent, however, the choice of the IRT model remains the same as originally: a standard model, (intended for measuring abilities) is fitted, and its appropriateness is assessed via goodness of model-data fit (see e.g., Meijer & Baneke, 2004; Reise et al., 2018, 2021). Furthermore, the most commonly used models are based on two general assumptions. First, that the measured constructs can be conceptualized as dimensional continua. Second, that the relation between the construct and the scores in its indicators is monotonic (e.g., Meijer & Baneke, 2004; Reise et al., 2018, 2021; Tutz & Jordan, 2024). We concur with Reise et al. (2021) in that these assumptions may be approximately met in many noncognitive applications.

Beyond the two starting assumptions above, however, in routine applications three further conditions are, explicitly or implicitly, assumed. First is that the construct is equally meaningful at both ends (i.e., bipolarity). Second is that it is normally distributed in the target population. The third, finally, is that the link function that relates the construct levels with the expected item scores (i.e., the item response function, IRF) is a symmetrical ogive. These assumptions, usually taken by convenience, are reasonable in some cases but highly questionable in others (e.g., Bazán et al., 2006; Goldstein, 1980; Molenaar, 2015; Morales-Vives et al., 2023, 2025; Reise et al., 2021; Shim et al., 2023; Woods & Thissen, 2006).

Background, Aims, and Contributions

The modeling proposed here is intended for measuring a class of non-cognitive constructs in scenarios in which the two initial assumptions of dimensionality and monotonicity are plausible but the three additional assumptions above are not so. More specifically, our model is expected to be appropriate when, (a) the construct can be plausibly considered as unipolar, (b) its latent distribution in the target population is expected to be rightly skewed, and (c) the IRFs behave as asymmetrical power functions. The model is first aimed at offering a substantively consistent account of certain types of data which are found in applications, and, secondly, to be simple and have desirable psychometric properties (e.g., Ramsay, 1989).

We shall now discuss the first two assumptions above. Polarity is considered a conceptual property of the construct (e.g., Tay & Jebb, 2018). So, a bipolar construct is characterized by two opposite poles that are both fully meaningful, for example extraversion-introversion, with no point on the trait continuum representing a natural zero (the absence of the trait). Consequently, when a zero point is used, it typically defines the average point but does not mean the absence of the construct. In contrast, a unipolar construct is fully meaningful at only one pole, because the other pole implies the absence or near-absence of the construct. This would be the case with depression (Reise et al., 2021), suicidal ideation (Morales-Vives et al., 2023) or certain drug addictions (Lucke, 2015). In this type of construct, there is a meaningful zero point which marks the

beginning of the continuum, and so it is reasonable to represent the construct levels with positive numbers (Lucke, 2014, 2015; van der Maas et al., 2011).

That the latent distribution of the construct is rightly skewed is a distributional property of the target population (we shall assume that the model is fitted in a representative sample). Furthermore, the conceptual condition of unipolarity does not imply "per se" that the latent distribution of the construct has to be rightly skewed (e.g., it could be in some populations but not in others). Although both conditions are conceptually independent, however, the scenarios in which both are met are relatively common, mostly when the pole that implies the absence or near-absence of the construct encompasses most of the population, (which is the case, e.g., with suicidal ideation in community populations). The same as in previous developments (Lucke, 2014, 2015) our proposal is solely intended for variables that exhibit both characteristics, and, for completeness we prefer to refer these models as unipolar-skewed.

Our model is intended for measures made of continuous-response items, a format that is considered a natural application in noncognitive domains (Bejar, 1977). At the practical level, it is simple, less time-consuming, easily understood (e.g., McCormack et al., 1988), and well adapted to computerized administration modalities. Furthermore, IRT models intended for this format are parsimonious, and generally far more workable than graded-response models (e.g., Molenaar et al., 2022; Noel & Dauvier, 2007; Tutz & Jordan, 2024). Clearly, the format is on the rise in noncognitive measurement (Huang, 2025).

The model proposed in this article is unidimensional. Therefore, its domain of application is either an instrument that measures a single construct or a multidimensional questionnaire with a structure simple and clear enough to be analyzed on a scale-by-scale basis. A proposal for a multidimensional extension will be also advanced below.

As in any IRT model (e.g., Woods & Thissen, 2006) the proposal of a unipolar-skewed model entails two main choices: (a) the latent distribution of the construct, and (b) the link function or IRF. To the best of our knowledge, in all the unipolar formulations so far (regardless of the format), the chosen latent distribution has been the lognormal. For the specific case of continuous responses, only one model has been proposed so far (Ferrando et al., 2024), which combines the lognormal latent distribution with a log-logistic IRF. The present proposal, in comparison, continues considering the lognormal latent distribution but proposes and alternative IRF. The resulting model is essentially an extension of the Wb-UIRM, which Lucke (2014, 2015) proposed for binary responses, to the continuous-response case.

In relation to the existing log-logistic model and related developments, apart from considering an alternative IRF, the present proposal makes four new contributions:

- First is to investigate in depth the model functioning as a quotient model.
- Second is to assess the uncertainty of the parameter estimates and its impact in terms of confidence intervals and confidence bands.
- Third is to develop a multi-faceted approach for assessing model appropriateness.

- Fourth is to propose a procedure for empirically estimating the latent distribution of the construct.

Finally, the proposal is complete, freely implemented in an R program, and illustrated with an empirical example.

The Weibull Continuous Response Model: Wb-CRM

Consider an instrument, made up of *n* items with a doubly bounded continuous response format, scaled in the open interval (0,1), and aimed at measuring a unipolar construct θ_U. In the target population, θ_U is assumed to follow a lognormal distribution with parameters $\mu_U = 0$ and $\sigma_U = 1$. This assumption implies that, if θ_U is log-transformed, the resulting mean and standard deviation are 0 and 1 respectively. The mean and standard deviation of the original variable are 1.65, and 2.16 respectively (e.g., Aitchison & Brown, 1957).

For fixed θ_U, the expected score in item *j* is given by:

$$E(X_j \Big| \theta_U) = \frac{\exp(\alpha_j \theta_U{}^{\beta_j}) - 1}{\exp(\alpha_j \theta_U{}^{\beta_j})} = 1 - \exp(-\alpha_j \theta_U{}^{\beta_j}) \quad (1)$$

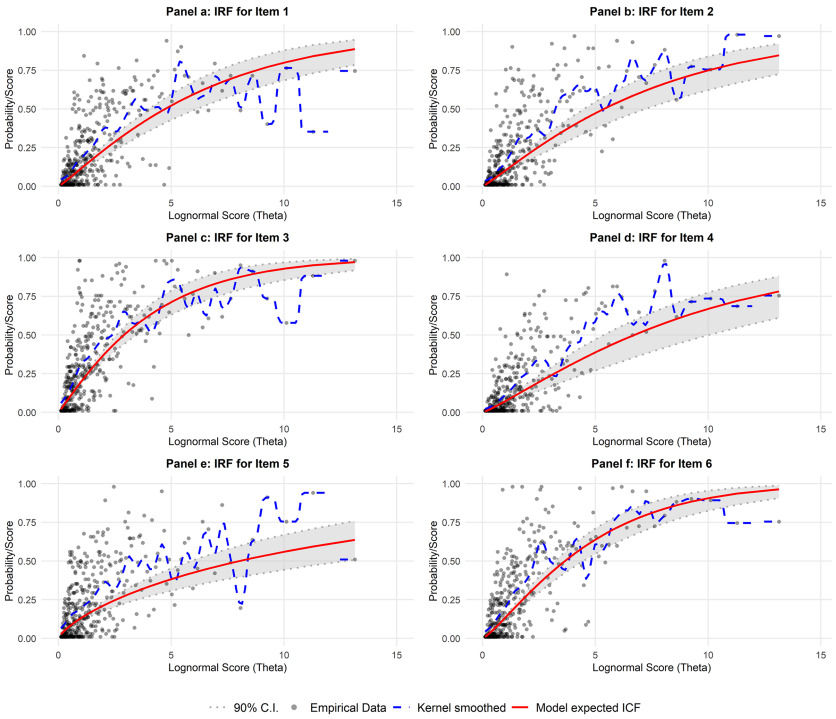
The link function (1) is one of the expressions of the two-parameter Weibull CDF (e.g., McCool, 2012), and, when considered as a function of θ_U, it becomes the IRF of the Wb-CRM. This IRF is, essentially, a 0–1 scaled power function (e.g., Stevens, 1975) and its general features are: (a) it is concave downward, (b) is not symmetrical around any trait level, and (c) its slope tends to increase more strongly for construct values close to zero and flattens as θ_U increases. Typical Wb-CRM IRFs are in Figure 1 below. The α_j and β_j parameters, restricted to have positive values, are original parameters of the Weibull CDF: The α_j parameter is a location/elevation parameter and β_j, is a shape/curvature parameter. Their interpretation as item response parameters is discussed below.

The IRF (1) can be linearized by using the following transformations

$$\ln(\ln(\frac{1}{1 - E(X_j \big| \theta_U)})) = \ln\Big(-\ln\Big(1 - E(X_j \big| \theta_U)\Big)\Big) = \ln(\alpha_j) + \beta_j \ln(\theta_U) = \mu_j + \beta_j \theta_B \quad (2)$$

Figure 1

Item Response Function for each BSI Somatic Items



Expression (2) is the well-known expectation of a unidimensional linear factor analytic (FA) solution (e.g., Mellenbergh, 1994) when applied to transformed variables. Within IRT modeling, the complementary log–log transformation on the left-hand side of (2) has been explicitly proposed by Goldstein (1980), Mellenbergh (1994), and Shim et al. (2023). As for the $ln(\theta_U)$ transformation, θ_B becomes now a variable that ranges from minus to plus infinity with a zero mean and unit variance (we use the *B* subscript for 'bipolar'). If the construct could plausibly be considered as bipolar, then the scale of θ_B would be already a "natural" scale for it (see Lord, 1975), and Model (2) would be equivalent to Goldstein's (1980) when used with continuous responses. Here, however, we regard θ_B as a convenient transformation of a unipolar construct θ_U whose more natural scale is lognormal. So, the linearization of the IRF (1) in (2) is achieved by using a double transformation: complementary log–log (for the item responses) and log (for the trait scale).

From (2), the linearized form of the Wb-CRM we propose here for respondent *i* when answering item *j* is:

$$Y_{ij} = \ln(\ln(\frac{1}{1 - X_{ij}})) = \ln(\alpha_j) + \beta_j \ln(\theta_{Ui}) + \varepsilon_{ij} = \mu_j + \beta_j \theta_{Bi} + \varepsilon_{ij} \tag{3}$$

With a constant residual variance that does not depend on θ_B:

$$\sigma^2_{\varepsilon_j} = Var\left(Y_j\right) - \beta^2{}_j \tag{4}$$

The conditional distribution of X_j for a fixed transformed value of θ_B is given by:

$$f(X_j \mid \theta_U = \exp(t)) =$$
$$= \frac{1}{\sigma_{\varepsilon j}\sqrt{2\pi}}\left(\frac{1}{\ln(\frac{1}{1-X_j})\left(1 - X_j\right)}\right)\exp\left\{-\frac{1}{2}[\frac{ln\left(\ln(\frac{1}{1-X_j}\right)) - \left(\ln(\alpha_j) + \beta_j t\right)}{\sigma_{\varepsilon j}}]^2\right\} \tag{5}$$

When the item response is observed and takes a specific value, then (5) becomes a mathematical function of θ_U: the item likelihood function (e.g., Thissen et al., 1983).

The density function (5) arises as a transformation system based on a normally distributed variable but does not belong to any standard system (e.g., Ferrando et al., 2024; Molenaar et al., 2022). Its shape and first two moments, however, can be approximated by using quadrature or by using the delta method (e.g., Raykov & Marcoulides, 2004), and, essentially, they agree with that can be expected in the case of doubly bounded response variables (Lord, 1975; Molenaar et al., 2022). So, the shape depends on the fixed construct level, and the distribution is rightly skewed at low construct values and skewed to the left at high values. The conditional mean is approximated by (1) and the conditional variance by,

$$Var(X_j|\theta_U) = \sigma^2_{\varepsilon_j}\left[\frac{\alpha_j \theta_U{}^{\beta_j}}{\exp(\alpha_j \theta_U{}^{\beta_j})}\right]^2 = \sigma^2_{\varepsilon_j}\left(\alpha_j \theta_U{}^{\beta_j}\right)^2\left(1 - E(X_j|\theta_U)\right)^2 \tag{6}$$

which also behaves as expected in that it becomes smaller at both ends of the construct continuum (e.g., Lord, 1980). In this modeling, however, the decrease is generally much more pronounced at the lower end.

We shall now provide an interpretation of the results so far described in terms of the functioning of (1) as an item response model. First, the asymmetric shape of the IRF implies that, as the construct level increases, the item score becomes progressively less sensitive to this increase, which is a most distinctive feature of unipolar IRT models (e.g., Lucke, 2015; Morales-Vives et al., 2023, 2025; Reise et al., 2021). Given that the item scores are doubly bounded while the trait scores are assumed to be only bounded below by zero and unbounded above, this shape is, in principle, plausible: the curve increases evenly from the common lower bound, and then flattens more and more (i.e., a ceiling effect) as

the construct levels increase without bound and the expected item score approaches its unit upper bound (e.g., Bazán et al., 2006; Lord, 1975).

The location parameter α_j has been referred to as the "easiness" index in previous proposals of unipolar models (e.g., Hernández-Dorado et al., 2024; Lucke, 2015; Reise et al., 2021). In principle, this name is not incorrect, because, other things constant, the higher α_j, the higher the expected item score is. However, it does not function in the usual way that difficulty indices are defined in IRT models. This point is clearly seen in the linearized Expression (3): $log(\alpha_j)$ is the expected marginal mean of the transformed item score Y_j (note that the marginal mean of θ_B is zero). It can be further shown that α_j also determines the expected marginal item mean in the original (0,1) response scale. So, essentially, α_j is an unconditional, conventional item location statistic related to the expected mean item score in the total group (Lord, 1980, p. 34). To avoid confusion with the IRT item difficulty index we shall propose below, we prefer to refer to α_j as an elevation index. As for reference values, in applications where the model is considered appropriate, expected α_j values will always be less than 1, generally in the range 0.25–0.50 (Hernández-Dorado et al., 2024).

A conditional location or difficulty item parameter which agrees with an IRT formulation of the model is now defined. The δ_j location parameter, which expresses "extremeness", "severity" or "difficulty", is defined as the θ_U trait level at which the expected item score is 0.5 (i.e., the midpoint of the item response scale; Lucke, 2014, 2015; Reise et al., 2021). For the Wb-CRM this parameter is,

$$\delta_j = \left(\frac{ln2}{\alpha_j}\right)^{\frac{1}{\beta_j}} \tag{7}$$

And, in terms of δ_j, the IRF (1) can be written as:

$$E(X_j|\theta_U) = 1 - \exp\left(-ln2\left(\left(\frac{\theta_U}{\delta_j}\right)^{\beta_j}\right)\right) \tag{8}$$

Examination of (8) shows that δ_j agrees with the definition of a standard IRT difficulty index but has a different functioning than this index has in conventional IRT models. As for the agreements, δ_j is a conditional measure of location (the construct level that is required to attain a 0.5 expected item score) and is in the same scale as the construct θ_U (i.e., lognormal 0,1). So, reference values can be derived by taking into account that the mean, the median and the standard deviation of the lognormal (0,1) distribution are 1.65, 1, and 2.16 respectively. In principle, δ_j values below, say, 1 would be interpreted as that the item is "easy" in the sense that it only takes a low trait level to reach the 0.5 midpoint in the item response scale. Values above 1.65 mean would indicate that the item is difficult, and δ_j values greater than, say, one standard deviation above the mean (3.5–4

or more) would be interpreted as that it is extremely difficult (Hernández-Dorado et al., 2024).

Turning now to the differences. In standard IRT models, the expected item score is a function of the difference between the trait level and the difficulty index, which is the reason why these standard models are known as difference models (see Ramsay, 1989; van der Maas et al., 2011). In (8), however, the expected item score is a function of the quotient between the trait level and the difficulty index. So, the Wb-CRM is a unipolar quotient model (Ramsay, 1989; van der Maas et al., 2011). This distinction becomes clearer if the linearized version of the model (3) is written in terms of δ_j:

$$Y_{ij} = \ln(\ln(\frac{1}{1 - X_{ij}})) = \ln(ln2) + \beta_j(\ln(\theta_{Ui}) - ln(\delta_j)) + \varepsilon_{ij} \quad (9)$$

So, when the original model is linearized using the complementary log–log transformation, the logarithms of both the construct level and the difficulty index behave according a conventional difference mechanism: The expectation of the transformed item score becomes a weighted function of the difference of logarithms (Ramsay, 1989; van der Maas et al., 2011).

The functioning of the IRF (8) then is that the conditional expected score increases with θ_U at a ratio that is inversely proportional to the difficulty. When the item is very "easy" (i.e., δ_j close to zero) the increase is abrupt, and the expectation reaches the unit upper bound already at low trait levels. As δ_j increases, the expected score increases more gradually. The β_j parameter modulates this tendency (accentuating or moderating it) acting in the form of a power exponent.

We turn now to item discriminating power. To address this property from an IRT framework, we shall first derive the slope of the curve at each point of θ_U, in terms of the quotient formulation (8). It is given by:

$$Slope(\theta_U) = \frac{dE(X_j|\theta_U)}{d\theta_U} = \left(1 - E(X_j|\theta_U)\right)\frac{ln2\beta_j}{\theta_U}\left(\frac{\theta_U}{\delta_j}\right)^{\beta_j} \quad (10)$$

So, the slope increases generally when, (a) the construct level decreases (and so the expected item score decreases), and (b) when β_j increases. However, as advanced above, the impact of β_j is not constant (as it occurs in a standard difference IRT model). Rather, in the Wb-CRM, this impact depends on the magnitude of the quotient θ_U/δ_j . Thus, in a very "easy" item, the slope would be very high at low construct levels. If, in addition, β_j was also high, the curve will be near vertical at low levels and then would flatten completely. At the other end, in a highly difficult item, the curve will increase gradually, with a gentler, more constant slope. If, in addition, β_j was not large, not only would the increase be even more gradual, but the expected score would not seem to reach the unit upper bound even for very high trait values.

Further insight on the role of β_j as a determinant of the slope can be obtained by noting that, when $\beta_j \leq 1$, the slope of the IRF increases without bound as θ_U approaches zero. When $\beta_j > 1$, the maximal slope is attained at:

$$\theta_U = \delta_j \left(\frac{\beta_j - 1}{ln2\beta_j} \right)^{1/\beta_j} \tag{11}$$

The results so far lead us to propose β_j as a general or "basic" index of item discrimination. However, the complete study of the discriminating power of a Wb-CRM item with a given difficulty parameter at different construct levels requires studying the behavior of the slope (10) and an inspection of the IRFs (see Figure 1 and corresponding comments). As for reference values for β_j finally, they are similar to those proposed for the discrimination parameter in standard IRT models (Bejar, 1977; Ferrando & Morales-Vives, 2023). Tentatively, and based on our experience, we may consider β_j values about between 0.3 and 0.7 as be moderately discriminating while values above 1 would already indicate a high level of "basic" item discriminating power.

Model Estimation

The results (2)–(4) allow a simple, limited-information and robust two-stage procedure for calibrating the items and scoring respondents to be proposed. In the first, calibration stage, the FA solution (3) is fitted to the transformed item scores and, if considered appropriate, the FA item structural estimates are next transformed to the Wb-CRM item estimates and taken as if they were fixed and known. In the second, scoring stage, individual trait estimates, with the corresponding standard errors of measurement and conditional reliability estimates are obtained on the basis of the first-stage fixed structural estimates.

Structural Estimation: Item Calibration

In this stage, a standard unidimensional FA solution is fitted to the complementary log–log transformed item scores (3). In order to obtain finite estimates, the lowest and highest endpoints of the original 0–1 scores are set at .01 and .99 (e.g., Shuford et al., 1966). The solution is identified by fixing the mean and standard deviation of θ_B to *0* and *1* respectively, which follows from the prior lognormal assumption. The item structural estimates directly obtained from the FA output are: the intercepts (μ_j), the loadings (β_j), and the residual variances ($\sigma^2_{\varepsilon_j}$). Next, the elevation estimates are obtained as: $\alpha_j = \exp(\mu_j)$, and the location/difficulty δ_j parameters are obtained from Equation (7). Fitting procedures and structural goodness-of-fit assessment are conventional (e.g., Raykov & Marcoulides, 2012).

If the chosen estimation procedure allows standard errors for the structural estimates to be obtained, then, the standard errors and confidence intervals for the β_j estimates are

directly provided in the calibration output. As for the $\alpha_j = \exp(\mu_j)$, and δ_j estimates, they can be obtained by the delta method (e.g., Raykov & Marcoulides, 2004).

An advantage of the simple calibration procedure so far described is its expected robustness against misspecification of the latent trait distribution. If the normality assumption for θ_B (or, equivalently, the log-normality assumption for θ_U) was incorrect, the quality of the β_j and α_j estimates would not be expected to be appreciably affected (Asparouhov & Muthén, 2010), particularly, if a robust estimation procedure was used.

Scoring

The scoring schema we propose is Bayes expected a posteriori (EAP; Bock & Mislevy, 1982). It is standard and the only specific features are that, (a) the prior is lognormal (0,1), and (b) the likelihood function is obtained from the conditional distribution (5). For each individual, the output consists of the trait θ_{Ui} point estimate and the corresponding posterior standard deviation (PSD) (e.g., Bock & Mislevy, 1982).

EAP-based conditional and marginal reliability estimates (e.g., Mellenbergh, 1994) are as follows. The conditional reliability estimate $\widehat{\theta}_{Ui(EAP)}$ is obtained as

$$\rho_{\widehat{\theta}_{Ui}(EAP)} = 1 - \frac{PSD(\widehat{\theta}_{Ui})^2}{Var(\theta_U)}. \tag{12}$$

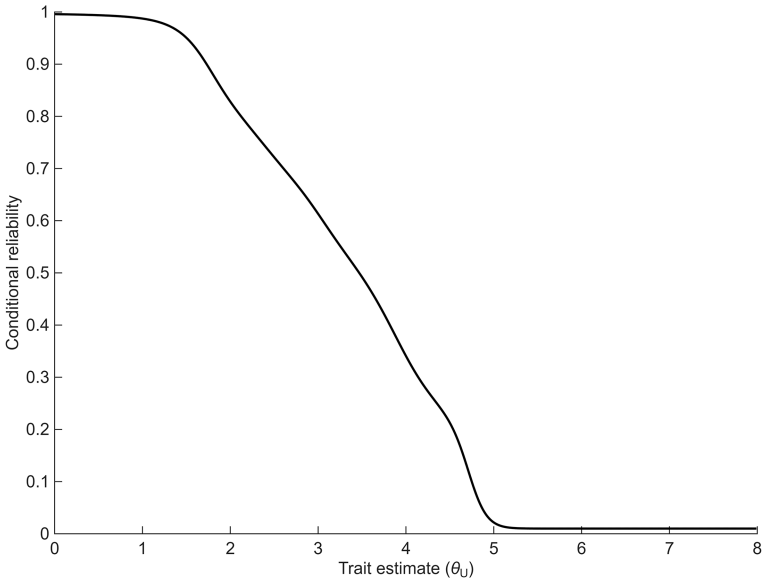
Where $Var(\theta_U)$ is the variance of the prior lognormal distribution. The marginal reliability estimate is:

$$\rho_{(EAP)} = 1 - \frac{E(PSD(\widehat{\theta}_{Ui})^2)}{Var(\theta_U)}. \tag{13}$$

EAP-based information curves can be obtained by plotting the PSDs or the conditional reliabilities against the point-estimates. Figure 2 in the example below is a typical example of that can be expected: in accordance with the functioning of the model, the curves are decreasing functions in which maximal information/accuracy is obtained at low trait levels.

Figure 2

Empirical Information Curves for the BSI Somatic Score Estimates



Assessing Model Appropriateness: A Multi-Faceted Proposal

While model appropriateness in noncognitive IRT applications has been mostly focused on goodness of model-data fit (GOF) assessment, we consider this assessment as largely insufficient (cf. Pitt & Myung, 2002), and propose instead a multi-faceted approach, in which different sources of evidence are collected, for the Wb-CRM.

First source is obtained via content and descriptive analysis. The functioning of the model must be plausible and theoretically consistent, which means that the construct should fit in the unipolar conception (e.g., Tay & Jebb, 2018), and its distribution should be expected to be rightly skewed in the target population (e.g., Woods & Thissen, 2006). The second assumption implies that 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. This, however, is only a necessary condition (see Molenaar, 2015).

The second source of evidence is conventional GOF. Because the Wb-CRM is fitted at the calibration level as if it was a Structural Equation Model (SEM), the powerful machinery for assessing GOF within this framework can be used directly here (e.g., Raykov & Marcoulides, 2012).

Still at the calibration level, the third proposed source of evidence is a graphical residual check on an item-by-item basis (e.g., Wells & Hambleton, 2016). This check is useful for comparing the functioning of the empirical Wb-CRM IRFs against their

model-expected counterparts, or against other possible IRFs. It is also useful for detecting unexpected item features (Meijer & Baneke, 2004; Reise et al., 2021). The specific proposed procedure is as follows. First, empirical percentiles are obtained from the sum scores and transformed to the corresponding value from the lognormal distribution. Second, for each item, the item scores are plotted against the transformed scores. Finally, both the model-expected IRF, obtained using Equation (1) and the kernel-smoothed IRF (see Ramsay, 1991) are fitted to the regression scatterplot. If standard errors for the structural item parameters are available from the calibration results, confidence bands for the expected IRFs can also be obtained by specifying a grid of values across the transformed scores, obtaining conditional expected values at each grid point based on generated parameter-estimated values that contain error, and obtaining confidence intervals from the array of resulting conditional expected values (e.g., Bollen & Stine, 1992).

The fourth and fifth sources focus more on the properties of the scores. The fourth source compares the specified prior distribution to the average of the latent posterior distribution estimated in the sample (Monroe, 2021). More specifically, the estimated posterior distribution for a given sample response vector $\mathbf{X_i}$ is:

$$f(\quad \theta_U \quad |\mathbf{X_i}) = \frac{f(\mathbf{X_i} \quad | \quad \theta_U) g(\quad \theta_U)}{\pi(\mathbf{X_i})} \tag{14}$$

Where $f(\mathbf{X_i}|\theta_U)$ is obtained from Equation (5). The sample average of the (14) estimates provides the posterior estimate corresponding to a given θ_U value.

Strong discrepancies between the prior and the posterior distribution would indicate a misspecification of the Wb-CRM either in the link functions, the latent prior, or both (e.g., Monroe, 2021; Sinharay et al., 2006). However, if the residual item analyses suggest that the IRFs are appropriate, the problem must be sought further in the prior latent distribution.

As in the third point above, the posterior analysis here is mostly based on graphical procedures, in which the prior and posterior smoothed distributions based on the quadrature histograms (see next section) are jointly displayed. As a descriptive discrepancy measure, the root mean squared difference between the prior and posterior probabilities is also obtained.

Finally, some studies have used results concerning external validity as further evidence of the plausibility of the unipolar-skewed modeling (e.g., Ferrando et al., 2024; Morales-Vives et al., 2023, 2025). If the variable is truly unipolar, it can be expected that the relationship with most relevant external variables will be more pronounced for subjects with high scores because this part of the trait continuum is more meaningful than the lower pole.

In closing, we believe it is crucial to collect sufficient evidence to determine model appropriateness thereby justifying its application. Using models that match both the unipolar nature of the variables and the latent distributions makes it possible to better

establish at which trait levels the accuracy of estimations is highest (e.g., Ferrando et al., 2024; Morales-Vives et al., 2023). This choice has important implications, for example, in terms of establishing adequate cut-off points for identifying individuals with certain conditions or disorders (e.g., Morales-Vives et al., 2023; Morales-Vives et al., 2025).

Model Modifications: Estimating the Latent Distribution of θ_U

When the structural GOF results are considered acceptable, but the posterior analysis reveals substantial discrepancies between the prior and the empirical posterior distributions, it is possible to estimate the θ_U distribution. In our proposal, the item parameter estimates obtained at the calibration stage are taken as fixed and known, and the latent θ_U distribution is estimated by using an adaptation of Mislevy's (1984) empirical histogram approach (see also Woods, 2007). Wang and Zeng (1998) proposed an extension of this approach for the case continuous response items that behaved according to Samejima's (1974) CRM. The procedure in our case is the same as they summarize in their Equation (21), with the Wb-CRM specific features that, (a) the prior is lognormal ($\mu_U = 0$, $\sigma_U = 1$), and (b) the likelihood function is obtained from the conditional distribution (5).

If the empirical histogram solution is consistent and departs substantially from the prior latent distribution, the EAP scores, PSDs, and reliability measures can be re-estimated by using the final histogram as the "input" quadrature instead of the prior quadrature. The behavior and advantages of these refined scores (particularly their stability), however, is an issue that deserves intensive research (see Woods, 2007).

A Template for a Multidimensional Extension

A direct extension of (1) to the case of a test that measures a set of *r* unipolar traits: $\boldsymbol{\theta_U} = \theta_{U1}, .. \theta_{Uk} ... \theta_{Ur}$, each of them distributed as lognormal (*0,1*), and that we propose as the multidimensional extension of the Wb-CRM, is given by:

$$E(X_j|\boldsymbol{\theta_U}) = \frac{\exp(\alpha_j \theta_{U1}^{\beta_{j1}} .... \theta_{Ur}^{\beta_{jr}}) - 1}{\exp(\alpha_j \theta_{U1}^{\beta_{j1}} .... \theta_{Ur}^{\beta_{jr}})} = 1 - \exp(-\alpha_j \theta_{U1}^{\beta_{j1}} .... \theta_{Ur}^{\beta_{jr}}) \quad (15)$$

By using the double (log–log and log) transformation proposed for the unidimensional model, the expectation in (15) is transformed to the expectation of a linear multiple factor-analytic solution. So, the two-stage estimation we proposed for the Wb-CRM is also feasible here. We consider that a full development of (15) in the same line as has been done with related unipolar models (Ferrando et al., 2026) is of interest and we plan to developed it in the future.

Implementation and Simulation Checks

The proposal so far discussed has been implemented as R script called UNI_ASK_U, which has been developed in R Version 4.5.1 and runs with R versions more recent than 3.5.0. A description of this program can be found in https://psico.fcep.urv.cat/utilitats/UNI_ASK_U/index.php.

The program is released in a compressed folder, which contains two main R scripts: UNI_ASK_U.R and UNI_ASK_U_GUI.R. The first script is the code function, where the input values have to be provided through console, and the output is also printed in the console. When using this script, the user must source the function before its utilization:

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

The second script, UNI_ASK_U_GUI, is a shiny app (Chang et al., 2025), which provides a Graphical User Interface version of the program. This can be initialized by:

```
> library(shiny); runApp('UNI_ASK_U_GUI.R')
```

More information about the usage of the script can be found at the aforementioned website.

Empirical Study

For this study, we used the somatization subscale of the *Brief Symptom Inventory 18* (BSI 18, Derogatis, 2001), which consists of 6 items referring to different somatic manifestations (faintness, nausea, numbness, etc.). Over the years, L. R. Derogatis has developed various questionnaires, such as the BSI-18 or the SCL-90-R, that aim to assess a broad, general construct of psychological distress across different dimensions, being somatization one of them. In his studies, the somatization subscale is treated in the same way as the other subscales, such as the depression or anxiety subscales, considering that the items that compose the subscale function as indicators, or are more specific manifestations, of the general construct of psychological distress. In fact, the BSI-18 allow overall scores to be calculated as a general measure of psychological distress (Global Severity Index), which includes the somatization items along with the other items. Based on this conceptualization, the present study also adopts this premise, and considers that the scores on this subscale reflect an underlying sub-dimension of distress specifically related to physiological experiences, whereas the other subscales, reflect other specific kinds of distress. According to this premise, this subscale is suitable to be used as an effect-indicators of a construct in this empirical example. Furthermore, it should be considered that the DSM-5 includes a diagnosis of *Somatic Symptom and Related Disorder*, and also considers that its indicators are not formative symptoms (they cannot be fully explained by a known general medical condition), but effects of a psychological distress dimension, which is consistent with this conceptualization.

The subscale was administered to 367 undergraduate students (84.7% women) with ages between 18 and 42 years old (M = 20.56, SD = 4.15). Participants were asked to rate to which extent they suffered each of these symptoms using a continuous response format that was scaled to 0-1 as described above. Participants were also asked to complete the *Rosenberg Self-Esteem Scale* (RSE, Rosenberg, 1979) to be used as external variable.

The medians and skewness coefficients for each somatic symptom are shown in Table 1. Considering the 0–1 scaling, the medians are quite low, with values lower than .28, except for the "Nausea" item (whose median is also not high: .34). The median in the "Short breath" item is especially low (.18). The skewness coefficients are close to 1, or higher than 1, in all cases except the "Nausea" item. Therefore, as expected in unipolar-skewed scenarios, participants tended to score very low, giving rise to positive skewed score distributions.

Table 1

Medians and Skewness Coefficients of BSI Somatic Items

Subscale	Item	Median	Skewness
Somatic subscale	1. Faintness	.24	0.94
	2. Chest pains	.23	1.17
	3. Nausea	.34	0.65
	4. Short breath	.18	1.46
	5. Numbness	.25	0.91
	6. Weakness	.27	0.92

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

Item Calibration

Using the UNI-ASK-U program, the unidimensional FA solution (3) was fitted to the transformed item scores by specifying a robust (mean and variance corrected) maximum-likelihood (ML) estimation criterion. GOF results at the structural level were quite good: RMSEA = .036, 90% CI [0.001, 0.077]; CFI = .99; GFI = .99; SRMR = .023. The item parameter point estimates together with the confidence intervals are in Table 2. Regarding the point estimates, all the items but the "Nausea" item have δj values greater than 3.5–4, so they can be considered as extremely difficult (particularly Items 4 and 5): a very high trait level is needed to reach the 0.5 midpoint in the item response scale. The α_j elevation estimates are consistent with these results, as all the items have considerably low values, close to zero, with the exception of the "Nausea" item, which is slightly higher. Regarding the β parameter, 5 out of 6 items can be considered as highly discriminating, as their values are higher than 1, while item 5 is moderately discriminant. To sum up, all the items can be considered as moderately/highly or highly discriminating and difficult.

Table 2

Item Parameter Estimates for the BSI Somatic Items with Confidence Intervals

Item	α (90% CI)	β (90% CI)	δ (90% CI)
1. Faintness	0.12 (0.11; 0.14)	1.13 (1.00; 1.26)	4.72 (3.73; 5.71)
2. Chest pains	0.11 (0.09; 0.12)	1.12 (0.98; 1.26)	5.17 (3.86; 6.48)
3. Nausea	0.22 (0.20; 0.23)	1.06 (0.92; 1.20)	2.95 (2.43; 3.47)
4. Short breath	0.07 (0.06; 0.10)	1.18 (1.05; 1.31)	6.98 (4.87; 9.09)
5. Numbness	0.14 (0.12; 0.16)	0.78 (0.64; 0.92)	7.77 (4.15; 11.39)
6. Weakness	0.14 (0.13; 0.16)	1.23 (1.10; 1.36)	3.67 (3.01; 4.33)

With regards to the 90% confidence intervals, they are all reasonably narrow, particularly taking into account the modest sample size. As expected, the elevation α_j estimates are the most accurate (they are intercepts), and the δ_j estimates are the ones with the greatest sampling error (they are a complex function of intercepts and slopes) although, on the other hand, they are the most interpretable. Note also that the width of the confidence intervals (and so, the magnitude of the standard errors) for δ_j increases with δ_j.

Figure 1 displays the IRFs of the six items. In all cases the model-expected IRF curve follows a general pattern similar to that of the empirical (kernel-smoothed) curve and the scatterplot, suggesting that the model is appropriate for this kind of items. As for the confidence bands around the expected IRF they are reasonable in all cases, in agreement with the results in Table (2), and Equation (6), and their width generally increases with the magnitude of the transformed scores, which is consistent with the greater dispersion of points at the upper side of the trait continuum.

It is also noted that the curves of the two most difficult Items (4 and 5) are slightly different from those for the rest: their increase is very gradual and seem not to attain the upper ceiling of 1. In other words, the items are so difficult that, even at moderate or high trait levels, there is little chance of scoring very high on them. In contrast, the increase of the curve is more pronounced in easiest items, especially in the "Nausea" item, reaching sooner the upper limit of one. All these results are in agreement with the functioning of the model discussed above.

Turning now to the scoring-related results. Figure 2 shows the conditional reliabilities against the EAP point estimates. As can be seen, maximal reliability is achieved at a range of construct levels between, approximately, 0 and 1. And, although this 0–1 range seems narrow, it contains the lower 50% of the distribution (the median of the latent distribution is 1). In other words, the most accurate results are obtained for the half of the sample that has no or very few symptoms. This result suggests, as expected,

that score estimates can effectively distinguish between individuals with very low somatization levels at the cost of not allowing for such accurate differentiation between different high levels of severity. Even so, however, up to a construct level of about 2.5, the conditional reliability still remains acceptable, and this level is the 82nd percentile. Finally, for estimates that are one standard deviation above the mean (about 3.8) the accuracy is, effectively, quite low (around .50), contrasting with the accuracy at low levels. However, we are talking here about individuals who are above the 90th percentile. So, in our view, the criticism that the information/conditional reliability function derived from unipolar modeling is somewhat implausible (too exaggerated; Reise et al., 2018) should be qualified, at least for the present model and response format.

A practical issue derived from the scoring results so far (e.g., Reise & Waller, 2009) refers to the consequences of using simple sum scores instead of the IRT scores proposed here if the Wb-CRM was the correct model. To address this issue, Figure 3 displays, for each of the 367 participants, the sum score against the estimated trait level (the EAP score). Included in Figure 3 is also the expected sum score, given the structural item estimates and the estimated trait level. Results are clear. Note first, that the curve of the expected sum score is, in IRT terms, the test response function (TRF) and its shape is an average of the IRFs in Figure 1 (e.g., Lord, 1980). Note also the closer agreement between the TRF and the scatter of points, which provides additional evidence supporting model appropriateness. The relation is nonlinear (a power function), and, again, at the lower end, small increases in the IRT score would translate into large differences in terms of sum score and the opposite trend would be observed at the upper end. So, for many purposes (individual comparisons, clinical decisions, analysis of change) the two types of scores would not be interchangeable (see Reise & Waller, 2009 for a discussion). However, even when the relation is clearly nonlinear, the product-moment correlation between both sets of scores is still $r = 0.93$. So, if ranking individuals was the only applied purpose, then the scores would be essentially interchangeable.

We turn now to evidence of appropriateness related to scoring results. Figure 4 displays the prior and posterior smoothed trait distributions. The root mean squared discrepancy between both distributions was RMSD = 0.0019. So, the results again suggest that, (a) the Wb-CRM is quite appropriate in this case, and (b) there is no need to further estimate the latent distribution.

Figure 3

Plot of Sum Scores Against Trait Estimates

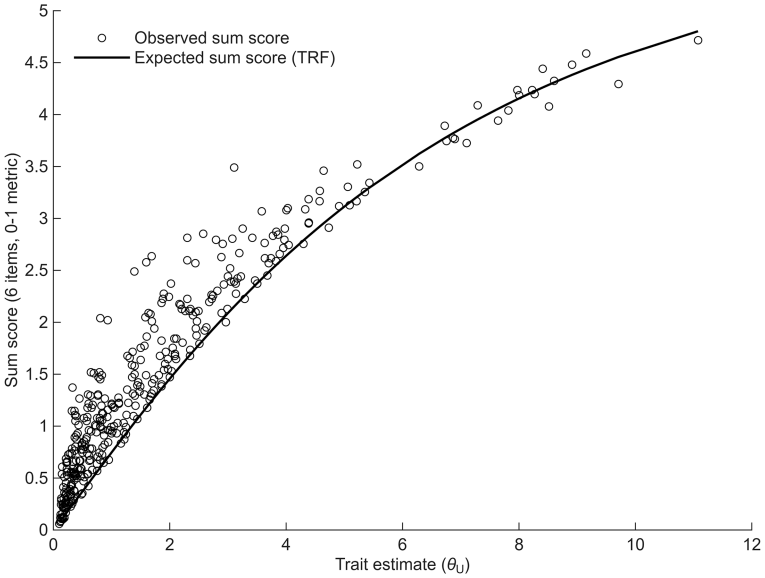
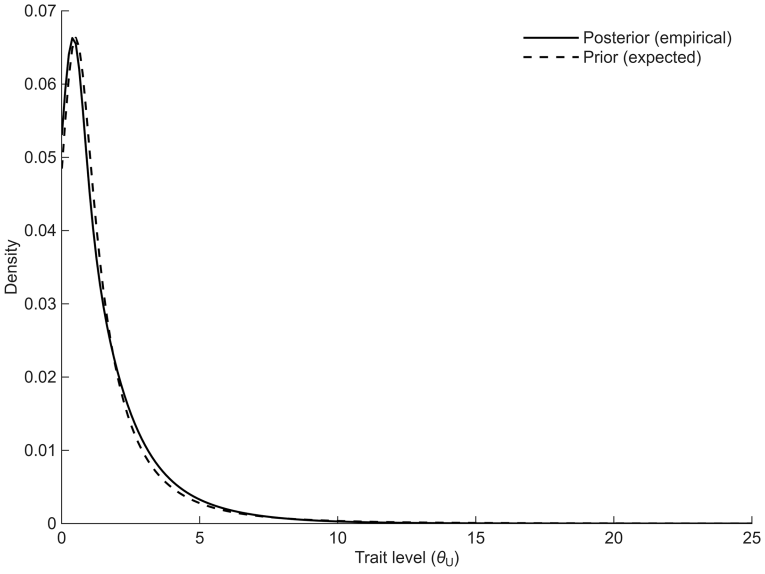


Figure 4

Prior and Posterior Smoothed Trait Distributions



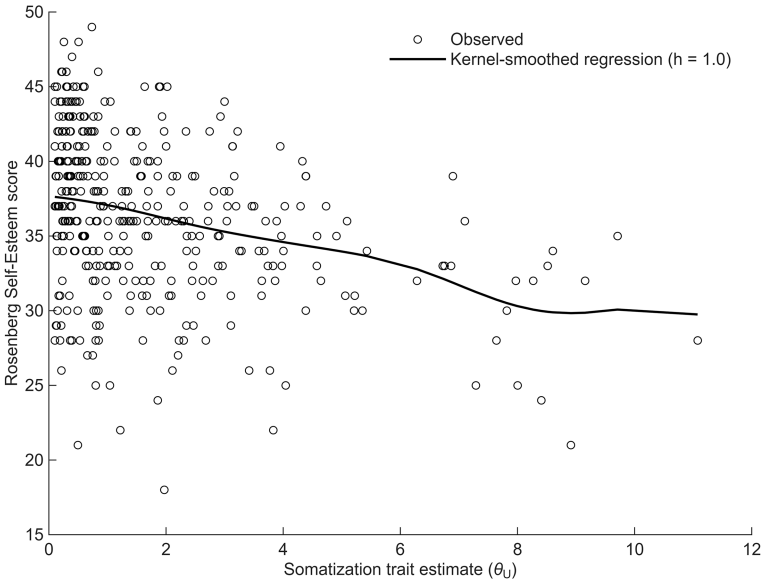
We turn finally to external validity evidence using the self-esteem RSE scores as external variable. As discussed above, we expected the somatic scores to become increasingly predictive as they increase, which implies a heteroscedastic relation in which the scatter of points around the regression line is more dispersed at the lower end of the trait continuum. Figure 5 shows the bivariate scatterplot of the self-esteem scores against the estimated somatic scores with the kernel smoothed regression line. Results suggest an essentially linear but heteroscedastic relation in the expected direction. To formally test this trend, we computed Cohen's (Cohen & Cohen, 1983) test of heteroscedasticity, which gave a significant and negative correlation between the absolute regression residuals and the estimated trait levels: r = -.12 (-0.20; -0.04), although the differential effects are not huge. However, this is understandable given that the validity coefficient is not very large either. Additionally, the correlation between the self-esteem scores on the somatic scores (validity coefficient) in the full sample was r = -.36, which involves a moderate effect size. The corresponding correlation in the majority group with low to moderate somatic scores (below 4) was r = -0.25, with a 90% confidence interval of -0.34 and -0.17. The corresponding correlation in the minority group with high scores was r = -0.48, a value that falls outside the limits of the reported confidence interval. Admittedly this second evidence based on group splitting has a component of arbitrariness and it is best to take it as supplementary evidence. Overall, results suggest that those people that suffer somatic symptoms tend to have lower levels of self-esteem, while the relationship between these variables is not so obvious for those people that do not suffer, or barely suffer, them. In other words, somatic symptoms may affect self-esteem, but not suffering these symptoms is not enough for manifesting a high self-esteem.

Discussion

In our view, for many noncognitive constructs and populations, standard IRT models are plausible choices. In other scenarios, the choice of these models might not be very well founded, but they can still function reasonably well (Reise et al., 2021). There are, however, certain scenarios in which both the scaling of the construct and its expected latent distribution clash head-on with the assumptions of a standard model. In these cases, many inferences based on standard results would likely to be wrong (Morales-Vives et al., 2023, 2025; Ferrando et al., 2024). The model proposed here is an alternative to be used in the third type of scenarios, and involves constructs that can be considered unipolar with an expected right-skewed latent distribution in the target population, and measured with doubly-bounded continuous response items.

Models related to the present proposal can be developed by playing with two determinants: the latent distribution and the link function or IRF. And, so far, the most convenient choice for the first, appears to be the lognormal (0,1) distribution, which is also the basic choice we made here, although we have also considered the possibility

Figure 5



Validity Evidence: Rosenberg Self-Esteem Scores Against BSI Somatic Score Estimates

of using it as a starting point. As for the link function the common choice so far has been the log-logistic (Morales-Vives et al., 2023, 2025; Ferrando et al., 2024), and we have proposed here a different link that is related to previous psychometric and biometric proposals: the Weibull link, that, when linearized, becomes the complementary log–log link. This is the main novelty of our proposal.

Beyond the new basis link, our proposal has attempted to include new developments that, to the best of our knowledge, had not been proposed until now in the field of this family of models. They include: (a) an intensive study of the model functioning as a quotient model, (b) taking into account estimation uncertainty, (c) improved procedures for assessing model appropriateness, and (d) the possibility of empirically estimating the latent distribution. Furthermore, the proposal has been fully operationalized using by simple and robust estimation procedures, implemented in a free user-friendly program, and illustrated with an empirical example. Even with all these contributions, however, we recognize that what we propose here is only a first step.

Limitations and Future Directions

The Wb-CRM has three main advantages. First, it is plausible and substantively consistent in the scenarios for which it has been designed, Second, its functioning is aligned

with the most common intended usages in clinical measurement (distinguish between people which are really low on the construct from those who are not). Third, it is relatively simple and has interesting psychometric properties. Even so, however, we still do not have conclusive evidence that a unipolar model is "genuinely correct", and, if so, that the Wb-CRM is the most appropriate choice among the potential alternatives. As for the first point, the new contributions of this article regarding appropriateness assessment (mainly graphical IRF residual analysis and posterior distribution analysis) are improvements with respect to previous developments, but they are still a first step.

In our view, gathering further evidence about model correctness mostly involves, first, linking the developments to theoretical foundations, and second, "getting out of the loop", making tangible outside predictions from the model and verifying them (Lord, 1980; Wells & Hambleton, 2016). The differential validity study in the example goes in this direction. As a further example, if the item set was large enough (unfortunately not the case in our example), the split-half method could be used, and the prediction that the standard errors should increase with trait level could be empirically checked. Ultimately, if it could be demonstrated that the use of the model leads consistently to improved cut-offs and better clinical diagnostic decisions, then, support for its correctness would be greatly enhanced.

With regards to model comparisons, we believe that they should also be based on the same rationales as outlined above and go beyond pure GOF comparisons. Again, residual and posterior analyses are a first step, but theoretical foundations, validity, and clinical functioning should be the ultimate criteria. This evidence should be based on intensive research and a variety of datasets obtained from different samples and measuring instruments. And, in this sense, the comprehensive implementation of different unipolar models in a general program we are conducting would allow applied researchers to carry out these investigations independently and gain knowledge about their functioning.

As for future directions that go beyond the general lines mentioned above, we believe that the study of the topic is in its early stages and has considerable expansive potential in both psychometric and applied research. In this sense, the reviewers of the first version provided valuable directions. Below is a summary of points that deserve further study.

- Improving and extending the procedures for assessing model appropriateness and for comparing alternative models. In particular, the residual IRF and the posterior analysis checks are very basic and can be clearly improved (Douglas & Cohen, 2001; Monroe, 2021; Sinharay et al., 2006; Wells & Hambleton, 2016). Also, procedures for comparing the functioning of competing models that go beyond GOF and that are being used in conventional IRT models, should be incorporated into these new models (e.g., Bazán et al., 2006; Wells & Hambleton, 2016).
- Improving the procedures for estimating the latent distribution. The empirical histogram approach is very basic, can provide little information in some cases

(particularly short tests) and can be improved (e.g., Woods, 2007). Alternative approaches should also be considered (e.g., Woods & Thissen, 2006).
- A more in-depth study of the uncertainty of the structural estimates and its impact on the score estimates. While we consider the simple two-stage approach to be quite defensible, the impact of considering the structural estimates as if they were fixed and known at the scoring stage should be further studied.
- Considering alternative score estimates. In particular, in the Wb-CRM maximum likelihood score estimates can be obtained in closed form and are not impacted by the choice of the prior distribution. On the other hand, they are unstable at the extreme levels and do not generally work well in short tests. Its usefulness should be assessed.
- Fully extending the Wb-CRM to, (a) the graded response format case, and (b) the multidimensional case according to the template provided in this article. These developments would correspond to those made for the log-logistic model: Reise et al. (2021) for the graded extension, and Ferrando et al. (2026), multidimensional extension.

Conclusion

In spite of its acknowledged limitations, we consider the proposal as a step forward and an original contribution to the field of unipolar constructs, which provides responses to assessment needs in this area. In this sense, the results of the illustrative example (limited as it is) not only show that the model behaves appropriately with real data, but also reinforces previous evidence obtained with unipolar models (Ferrando et al., 2024, 2026; Morales-Vives et al., 2023, 2025). If future evidence confirms this good functioning, the model would be quite useful for professionals and researchers working with non-cognitive unipolar-skewed constructs, and would fill a gap in this field. In particular, when applied to community samples, it would enable adequate cut-off points to be established for accurately differentiating between individuals with irrelevant construct levels and individuals who clearly have construct manifestations, and so for identifying individuals who require preventive assistance.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: No data has been made available for this article.

Supplementary Materials

Type of supplementary material	Availability/Access
Data	
No data provided	—
Code	
No code available	—
Material	
No material provided	—
Study/Analysis preregistration	
No preregistration	—
Other	
No other material has been made available	—

References

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

Asparouhov, T., & Muthén, B. (2010). *Plausible values for latent variables using Mplus. technical report*. www.statmodel.com

Bazán, J. L., Branco, M. D., & Bolfarine, H. (2006). A skew item response model. *Bayesian Analysis, 1*(4), 861–892. https://doi.org/10.1214/06-BA128

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

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*(2), 205–229. https://doi.org/10.1177/0049124192021002004

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/

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

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

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*(3), 234–243. https://doi.org/10.1177/01466210122032046

Ferrando, P. J., & Morales-Vives, F. (2023). Is it quality, is it redundancy, or is model inadequacy? Some strategies for judging the appropriateness of high-discrimination items. *Anales de Psicología / Annals of Psychology, 39*(3), 517–527. https://doi.org/10.6018/analesps.535781

Ferrando, P. J., Morales-Vives, F., Casas, J. M., & Navarro-González, D. (2026). A multidimensional continuous response model for measuring unipolar traits. *Applied Psychological Measurement, 50*(1–2), 3–20. https://doi.org/10.1177/01466216251360311

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

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *The British Journal of Mathematical & Statistical Psychology, 33*(2), 234–246. https://doi.org/10.1111/j.2044-8317.1980.tb00610.x

Hernández-Dorado, A., Ferrando, P. J., & Morales-Vives, F. (2024). *UNIPOL-GC user's guide: Technical report*. Universitat Rovira i Virgili.

Huang, H. Y. (2025). Exploring the influence of response styles on continuous scale assessments: Insights from a novel modeling approach. *Educational and Psychological Measurement, 85*(1), 178–214. https://doi.org/10.1177/00131644241242789

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Lucke, J. F. (2014). Positive trait item response models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt & 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 & 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

McCool, J. I. (2012). *Using the Weibull distribution: Reliability, modeling, and inference*. John Wiley & Sons.

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

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*(3), 354–368. https://doi.org/10.1037/1082-989X.9.3.354

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359–381. https://doi.org/10.1007/BF02306026

Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika, 80*(3), 625–644. https://doi.org/10.1007/s11336-014-9406-0

Molenaar, D., Cúri, M., & Bazán, J. L. (2022). Zero and one inflated item response theory models for bounded continuous data. *Journal of Educational and Behavioral Statistics, 47*(6), 693–735. https://doi.org/10.3102/10769986221108455

Monroe, S. (2021). Testing latent variable distribution fit in IRT using posterior residuals. *Journal of Educational and Behavioral Statistics, 46*(3), 374–398. https://doi.org/10.3102/1076998620953764

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*, 21397–21411. https://doi.org/10.1007/s12144-022-03224-6

Morales-Vives, F., Ferrando, P. J., & Hernandez-Dorado, A. (2025). Modelling maladaptive personality traits with unipolar item response theory: The case of Callousness. *The Journal of General Psychology, 152*(3), 375–402. https://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

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*(10), 421–425. https://doi.org/10.1016/S1364-6613(02)01964-2

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

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611–630. https://doi.org/10.1007/BF02294494

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

Raykov, T., & Marcoulides, G. A. (2012). *A first course in structural equation modeling*. Routledge.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*(1), 45–58. https://doi.org/10.1177/014662169001400105

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

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

Shim, H., Bonifay, W., & Wiedermann, W. (2023). Parsimonious asymmetric item response theory modeling with the complementary log–log link. *Behavior Research Methods, 55*(1), 200–219. https://doi.org/10.3758/s13428-022-01824-5

Shuford, E. H., Jr., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika, 31*(2), 125–145. https://doi.org/10.1007/BF02289503

Sijtsma, K., & van der Ark, L. A. (2022). Advances in nonparametric item response theory for scale construction in quality-of-life research. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 31*, 1–9. https://doi.org/10.1007/s11136-021-03022-w

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C. B., & Chalmers, R. P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research, 29*(4), 1030–1048. https://doi.org/10.1177/0962280220907625

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

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*(2), 211–226. https://doi.org/10.1177/014662168300700209

Tutz, G., & Jordan, P. (2024). Latent trait item response models for continuous responses. *Journal of Educational and Behavioral Statistics, 49*(4), 499–532. https://doi.org/10.3102/10769986231184147

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. https://doi.org/10.1037/a0022749

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

Wells, C. S., & Hambleton, R. K. (2016). Model fit with residual analyses. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume 2* (pp. 395–413).

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement, 67*(1), 73–87. https://doi.org/10.1177/0013164406288163

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*(2), 281–301. https://doi.org/10.1007/s11336-004-1175-8

Methodology (METH) is the official journal of the European Association of Methodology (EAM).

PsychOpen GOLD is a publishing service provided by the Leibniz Institute for Psychology (ZPID), Germany.

Methodology
2026, Vol. 22(3), 225–251
https://doi.org/10.5964/meth.20097