Measuring Unipolar Traits with Continuous-Response Items: Some Methodological and Substantive Developments

Abstract

In recent years, some models for binary and graded format responses have been proposed to assess unipolar variables or "quasi-traits". These studies have mainly focused on clinical variables that have traditionally been treated as bipolar traits. In the present study, we have made a proposal for unipolar traits measured with continuous-response items. The proposed log-logistic continuous unipolar model (LL-C) is remarkably simple, and is more similar to the original binary formulation than the graded extensions, which is an advantage. Furthermore, considering that irrational, extreme or polarizing beliefs could be another domain of unipolar variables, we have applied this proposal to an empirical example of superstitious beliefs. The results suggest that, in certain cases, the standard linear model can be a good approximation to the LL-C model in terms of parameter estimation and goodness of fit, but not trait estimates and their accuracy. The results also show the importance of considering the unipolar nature of this kind of trait when predicting criterion variables, since the validity results were clearly different.

Keywords: unipolar traits, log-logistic unipolar model, continuous response format, criterion validity, superstitious beliefs

In the conventional unidimensional factor-analytic (FA) and Item Response Theory (IRT) modeling of typical-response items (personality, attitudes and beliefs), the measured trait is assumed to be a bipolar dimension which is (approximately) normally distributed in the population of interest (Molenaar & Dolan, 2018, Reise et al., 2018, 2021). We understand here that a bipolar trait is equally meaningful (see below) at both ends of the continuum. In the personality domain, for example, this is the case of extraversion (with introversion at one pole and extraversion at the other) or emotional stability (with neuroticism at one pole and emotional stability at the other pole). But there are many examples in other psychological domains, such as happiness (with happiness at one pole and unhappiness at the other pole) or optimism (with optimism at one pole and pessimism at the other pole). Now, it should be clarified that the term "meaningful" is used here in terms of interpretation and differentiating power of the items and test scores. So, in a bipolar trait, (a) both, low and high scores reflect the relative standing (in terms of distance) of the individual below or above the mean; and (b) individuals can be equally well differentiated at both poles of the continuum.

The bipolarity and approximate normality assumptions are submitted to be reasonable in many typical-response applications, particularly when normal-range traits are measured in general populations (Reise et al., 2018, 2021, van der Maas et al., 2011). In other scenarios, however, they might be rather questionable. Consider, for example, an instrument intended to measure a clinical trait whose items refer to symptoms. Further, suppose that this instrument is administered in a community population for screening or detection purposes (e.g. Morales-Vives et al., 2022). First of all, the bipolarity assumption fits very uneasily here, as the lower pole of the continuum mostly reflects the absence of pathology. So (a) low scores would be difficult to interpret in terms of relative standing below the mean, and (b) it will be also difficult to

differentiate between the individuals located at the lower end. Second, as for normality, it is an unlikely assumption, as many individuals either do not suffer from clinical disorders or have very low trait levels. So, the item distributions are expected to be positively skewed (e.g. Magnus & Liu, 2018; Morales-Vives et al., 2022). This would be the case of variables such as drug addictions, depression or suicidal ideation. In these examples, one pole of the continuum would include the majority of people, who do not have the disorder or have very few symptoms, while the other pole would include those who do suffer from the pathology to varying degrees. Overall, it is expected that (a) many people will be grouped at the lower end of the continuum, with very little trait variability among them, and (b) far fewer people will be at the upper end but they will have a higher degree of trait-level heterogeneity.

If, for the type of trait under study (possibly a clinical trait), the bipolarity assumption is questionable, it may be worthwhile and more plausible to consider it as a "quasi-trait" or a unipolar trait instead of a bipolar trait (Reise & Rodriguez, 2016, Reise et al., 2018, 2021). As for more specific distinctions, the term "quasi-trait" means that only one end of the continuum is truly relevant (the pole reflecting the different levels of severity of those people who suffer from the pathology), while the other end only reflects the absence of the pathology, which makes it irrelevant (Reise et al., 2018). On the other hand, the term unipolar refers to a trait for which one end of the continuum is far more relevant or meaningful than the other (Lucke, 2013, 2015, Reise et al., 2018). So, the unipolar formulation means that there is a single population in which all individuals are scalable on the trait but not to the same degree. In contrast, the quasi-trait formulation means that the test scores reflect a mixture of two populations (symptomatic vs. asymptomatic) and that only the individuals in the symptomatic sub-population are truly scalable in the trait (see Morales-Vives et al., 2022). As for the

modeling, models intended for unipolar traits are labelled as “unipolar models” or “positive trait models” (Lucke, 2013, 2105), while the most well-known models for quasi-traits belong to the family of the “zero-inflated mixture models” (e.g. Magnus & Liu, 2018, Reise et al. 2021), and, essentially, aim to provide unbiased trait estimates for the subset of the population to which the trait is applicable. Finally, other developments combine both modeling formulations (Magnus & Liu, 2018). However, in this article we shall limit ourselves to unipolar modeling.

Continuing with our example, assume that the unipolar formulation is the most appropriate choice, but a standard bipolar-normal model is fitted instead: the business-as-usual approach (Reise et al., 2021). In this case, some distortions might be expected. To start with, the item distributions are likely to be strongly skewed. Although some skewed item distributions can be expected to appear also with normal-bipolar traits for various reasons (see below), this result should raise immediate concerns (Reise et al., 2018). Next, further potential distortions, which are discussed in greater detail below, may emerge when the model is fitted and the individuals scored. Psychometricians have been aware of these problems since the seminal study by Reise & Waller (1990) but, even today, they are largely overlooked in applications. So, skewed distributions are either ignored or transformed for purely statistical purposes (i.e. to make scores more amenable for linear modelling). And, at the applied level, no one seems to be too concerned about what the true or the most plausible distribution of the trait under study is. By way of example, in such widely examined clinical traits as depression or suicidal ideation, most studies routinely use models that assume normality (e.g., Bottesi et al., 2015; Sánchez-Álvarez et al., 2020; Zhang et al., 2014).

Within the unipolar formulation, the starting point of this paper is the log-logistic unipolar model for binary items initially proposed by Lucke (2013, 2015),

which we shall denote by LL-B. Magnus and Liu (2018) proposed estimation and implementation procedures for the LL-B and extended it to include the quasi-trait formulation mentioned above. Reise et al. (2021) extended Lucke's formulation to the graded-response case by using a re-parameterization of Samejima's (1969) graded response model, so we shall denote their proposal as LL-GRM. As far as we know, however, no extensions of the original model to the continuous-response case have been proposed to date, and we believe that they are worth exploring for both substantive and methodological reasons. At the substantive level, typical-response measurement is considered a natural application for continuous formats (Bejar, 1977) and also simple, easily understood by any type of respondent, and less time consuming (see e.g. McCormack et al., 1988). At the methodological level, the main claim is that the format is expected to provide more information and sensitivity than the discrete alternative to differentiate among individuals (Bejar, 1977, Samejima, 1969, 1973). Our experience, however, suggests that the gains in accuracy are minor (see also Dolnicar et al., 2011) and that the main methodological advantages lie in the relative simplicity and attractive features of the resulting models.

We turn finally to the application domains. So far unipolar models have only been used with clinical traits, particularly addictions (e.g., Lucke, 2015). However, we submit that they might be quite useful in other arenas, one of which is irrational, extreme or polarized beliefs that are uncommon in the general population. There is a growing social and academic interest in such beliefs, more so since the beginning of the COVID-19 pandemic, which highlighted the problem of the rapid spread of false ideas that contradict scientific evidence. Although internet and social networks facilitate the spread of these extreme or irrational ideas, most of the population does not believe in them, or at least not to a great extent (e.g., Huete et al., 2022), so they could plausibly

be formulated as unipolar traits (see below for a stronger rationale), behaving in the same way as the clinical variables above, in which one pole of the continuum is hardly meaningful, since it involves the absence of symptomatology (in this case, the absence of belief in an idea that goes against science and empirical evidence). Specifically, the present study focuses on superstitious beliefs and attempts to determine whether they can be fruitfully modeled as unipolar traits. As expected, extreme beliefs have been routinely modeled so far as if they were bipolar-normal constructs (e.g., Fasce et al., 2021; Renken et al., 2015).

At this point, some further discussion in terms of applicability may be relevant. The argument used in the article is that a unipolar trait (which is an intrinsic trait characteristic) is expected to lead to a skewed item distribution (which is an empirical result), and more so if many people is located near the zero-point origin (e.g. asymptomatic people in a community sample). However, skewed item distributions may arise due to many other causes compatible with bipolar traits (see Molenaar & Dolan, 2018) being faulty test construction and/or inadequate sample collection the most commonly considered. So, it would be very difficult to univocally determine that the skewed distributions are due to the unipolar trait nature, and so that the model proposed here is the most appropriate. Therefore, theoretical assessment of the trait nature (as we shall attempt to do here) is a first basic requirement. Apart from that, certain evidence that would support the appropriateness of the unipolar formulation exists. In spite of the criticisms that the skewed distributions observed in many clinical instruments are due to faulty test construction, it has been consistently found that, to construct items able to distinguish well between individuals anywhere along the trait continuum is an impossible task with this type of traits (e.g. Reise et al., 2021). And this is exactly the result that would be expected if the trait was unipolar

Purpose of the Present Study

The present article has three main aims. First, we propose to extend the original LL-B unipolar model to the continuous response case, which we shall denote as LL-C. Procedures for fitting the model, addressing model-data fit, scoring respondents and assessing score accuracy are proposed and implemented in a comprehensive R program. However, the emphasis is not on proposing sophisticated estimation methods, but rather on understanding the rationale, properties, functioning, and appropriateness of the model.

The second aim is to compare the functioning and interpretation of the LL-C model to those of (a) Samejima's (1973) continuous response model, and (b) the standard linear FA model, which is the most used in applications based on continuous formats. We assume that the LL-C is the 'true' model in order to assess: (a) the conditions in which these other models are expected to behave as a reasonable approximation, (b) the expected distortions when this is not the case, and (c) the differential interpretations to which the models lead.

The third aim is the most substantive: to explore the applicability, viability and interest of the proposed model in domains other than clinical symptoms and in which, to the best of our knowledge, unipolar models have never been considered before. More specifically, as discussed above, we shall undertake an empirical study focused on superstitious beliefs (for example, believing that if you break a mirror you will be unlucky, or that if your ears are ringing, someone is criticizing you).

The LL-C Proposal

Consider a test made up of *n* items with an approximately continuous format that aims to measure a unipolar trait θ_U. For interpretative purposes, the item scores are scaled to values between 0 and 1. The trait θ_U is assumed to follow a lognormal distribution with parameters $\mu_U=0$ and $\sigma_U=1$. So: (a) θ_U is anchored to zero and has no upper limit, and (b) $ln(\theta_U)$ is normally distributed with zero mean and unit variance.

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

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

The α_j and β_j item parameters are both restricted to having positive values, and are location and form-curvature parameters, respectively (see below). Expression (1) is the same as the one originally proposed by Lucke (2015, eq. 13.6) for the LL-B model.

As a function of θ_U, the conditional expectation (1) is the Item Response Function (IRF) of the LL-C model, which is, essentially a 0-1 scaled power function. For different values of α_j and β_j, figure 1 shows the resulting IRFs.

INSERT FIGURE 1 ABOUT HERE

The curves in figure 1 clearly depart from the usual ogives considered in standard IRT models. Their general trend is that they are concave downward, and their slope tends to increase more strongly for trait values close to zero and flattens as θ_U increases. This result agrees with most power formulations (e.g. Stevens, 1975) and means here that the item score becomes progressively less sensitive to the trait level as this level increases. The α_j parameter mainly determines the height of the curve. The β_j parameter can be interpreted as a discrimination parameter: at low β_j values the curve is

flatter at almost all trait levels and, the higher the β_j value becomes, the more the curve increases at low values. This functioning can be assessed in more detail by deriving the slope of the curve at each point of θ_U. It is given by

$$Slope(\theta_U) = \frac{dE(X_j \mid \theta_U)}{d\theta_U} = \frac{\alpha_j \theta_U^{\beta_j - 1}}{(1 + \alpha_j \theta_U^{\beta_j})^2} = (\frac{\beta_j}{\theta_U}) E(X_j \mid \theta_U)(1 - E(X_j \mid \theta_U)) \quad (2)$$

and is essentially the slope of a typical ogive (see Ferrando, 2002) but divided by the trait level. So, according to the numerator in (2), the slope is directly proportional to β_j (as it should be) and reaches its maximum at the expected score of 0.5 (see below). However, its magnitude is mostly dominated by the trait level in the denominator: as can be deduced from figure 1, the slope increases abruptly as θ_U approaches the 0 lower bound, and approaches zero as θ_U increases.

Further insight can be gained into the functioning of (1) if we define an extremeness or "difficulty" parameter as the θ_U trait level at which the expected item score is 0.5. Conceptually this is the trait level that corresponds to the midpoint of the response scale or the threshold scale value that marks the transition from a tendency to disagree with the item to a tendency to agree with it (see Ferrando, 2009). The parameter is:

$$\delta_j = \left(\frac{1}{\alpha_j} \right)^{\frac{1}{\beta_j}}. \quad (3)$$

And, in terms of the extremeness parameter, model (1) can be written as:

$$E(X_j \mid \theta_U) = \frac{\left(\frac{\theta_U}{\delta_j}\right)^{\beta_j}}{1+\left(\frac{\theta_U}{\delta_j}\right)^{\beta_j}}. \quad (4)$$

Expressed in the form (4), the LL-C model is a "quotient" model as considered by Rasch (1966), Lord (1975), Ramsay (1989), and van der Maas et al. (2011). The expected score increases with θ_U at a rate that is inversely proportional to the difficulty. When the item difficulty is close to zero, the IRF increases abruptly. For difficult items, the transition is more gradual. To appraise this point, note that at $\theta_U = \delta_j$, the slope of the IRF is: $\beta_j / 4\delta_j$

The functioning summarized so far differs considerably from that of the conventional IRT ogive models. In the latter, the IRFs are point-symmetric curves in which the extremeness/difficulty index is both (a) the point of symmetry and (b) the trait level at which the IRF has its maximum slope. In contrast, the LL-C curves are not symmetric around any trait level and reach their maximal slope as θ_U approaches zero.

The LL-C expectation can be linearized (Mellenbergh, 1994, Wang & Zeng, 1998) by taking natural logarithms on both sides of (1). That is to say,

$$\ln\left(\frac{E(X_j \mid \theta_U)}{1-E(X_j \mid \theta_U)}\right) = \ln(\alpha_j) + \beta_j \ln(\theta_U) = \mu_j + \beta_j \theta_B. \quad (5)$$

expression (5) is indeed the conditional expectation in Spearman's linear factor analysis (FA) model (e.g. Mellenbergh, 1994) when applied to transformed responses (e.g. Tutz & Jordan, 2022). Note that θ_B is now a normal variable that ranges from minus to plus

infinity. So we use the *B* subscript for 'bipolar'. On the basis of (5), the linearized form of the LL-C we propose here for respondent *i* when answering item *j* is:

$$Y_{ij} = \ln(\frac{X_{ij})}{1-X_{ij})} = \ln(\alpha_j) + \beta_j \ln(\theta_U) + \varepsilon_{ij} = \mu_j + \beta_j \theta_{Bi} + \varepsilon_{ij}\,. \qquad (6)$$

When residual variance is constant and does not depend on θ_B

$$\sigma^2_{\varepsilon_j} = Var(Y_j) - \beta^2{}_j\,. \qquad (7)$$

Because θ_B is a monotonic transformation of θ_U, the conditional distribution of X_j for a fixed value of θ_B is the same as its conditional distribution under the corresponding θ_U value (e.g. Lord, 1975). It is given by:

$$f(X_j \mid \theta_U = \exp(t)) = \frac{1}{\sigma_{\varepsilon j}\sqrt{2\pi}} \frac{1}{X_j(1-X_j)} \exp\left\{ -\frac{1}{2} \left[\frac{\ln(\frac{X_j}{1-X_j}) - (\ln(\alpha_j) + \beta_j t)}{\sigma_{\varepsilon j}} \right]^2 \right\}. \qquad (8)$$

(see Ferrando, 2002 and Wang & Zeng, 1998). Distribution (8) is the S_B distribution in Johnson's system (Johnson, 1949), and is also known as the four-parameter log-normal distribution (Aitchison and Brown, 1957).

We turn now to the properties of the scores derived from the structural model described so far. By using (8), we find that maximum likelihood (ML) score estimates for each respondent can be obtained in closed form. They are given by

$$\hat{\theta}_{Ui}(ML) = \exp\left[\frac{\sum_{j=1}^{n}\frac{\beta_j \ln(\frac{X_{ij}}{\alpha_j(1-X_{ij})})}{\sigma^2_{\varepsilon j}}}{\sum_{j=1}^{n}\frac{\beta^2{}_j}{\sigma^2_{\varepsilon j}}}\right]. \quad (9)$$

The accuracy with which score estimates in (9) allow the trait level of an individual to be assessed can be quantified by deriving the amount of test information, which is:

$$I(\theta_U) = -E(\frac{\partial^2 \ln L(\theta_U)}{\partial \theta^2{}_U}) = \frac{1}{\theta^2{}_U}\sum_{j=1}^{n}\frac{\beta^2{}_j}{\sigma^2_{\varepsilon j}}, \quad (10)$$

The interpretation is clear. The amount of item information depends directly on the signal-to-noise ratio (the term in the denominator of (9)) and inversely on the square of the trait value. So, information increases as the item becomes less noisy (i.e. of higher quality) and decreases as the trait level increases, which means again that, other things constant, the item score provides more information as the trait level approaches the zero lower bound.

Finally, we shall discuss the rationale and foundations of the LL-C model. If the trait can be conceived as bipolar, then (6) will simply be the linearized expression for the logistic version of Samejima's (1973) continuous response model (CRM, see also Bejar, 1977, Ferrando, 2002, Wang & Zeng, 1998). Now, because the trait is modeled as unipolar, the linearization (6) is obtained by using a double transformation: (a) the unipolar trait levels are log transformed to make them amenable for the CRM assumptions, and (b) the CRM is proposed as the model for the transformed θ_B trait levels. With this background, the LL-C model can be viewed as a change in the trait scale from the basic CRM. In further detail, the trait scale can be defined as the scale at

which all IRFs have a specific form, and this scale can be one-to-one monotonically transformed (Lord, 1975). Furthermore, transformations of this type can be made either for mathematical or practical convenience (Lord, 1975, Yen, 1986) or for more substantive and interpretative reasons (Ramsay, 1989), which is what we attempt here. First, the lognormal is possibly the most realistic choice for modeling a unipolar trait of the type described above: an intrinsically positive variable that has a low mean and high variance, both of which lead to pronounced skewness. Furthermore, a stronger substantive basis can be invoked if the measured trait can be conceived as a variable that has evolved according to the law of proportional effects (e.g. Aitchison & Brown, 1957): i.e. when the increase in the trait level at any stage of its evolution is a random proportion of the previous level. Lucke (2013, 2015) considered that addictive disorders grow according to this law.

In this article, we shall try to justify that other constructs can also grow according to the law of proportional effects, particularly uncommon, extreme or polarized beliefs that are not supported by science. Firstly, what could give rise to this growth in these beliefs is the confirmation bias, according to which individuals tend to select and consume contents that are in line with their pre-existing ideas and their hypothesis in hand (e.g., Nickerson, 1998; Westerwick et al., 2017). Secondly, this increase may also be due to the phenomenon known as "echo chambers". According to Quattrociocchi et al. (2016), people with polarized ideas, for example about conspiracies or against science, tend to interact online with like-minded individuals and ignore other communities. So, the more polarized people are, the more like-minded friends they have on social networks. This would make the internet act as an echo chamber for these individuals, reinforcing pre-existing ideas and making them more and more extreme. Therefore, as a working hypothesis, we shall submit that this increase is

in line with the law of proportional effects: i.e. the partly-random increase in the extremeness of these ideas at a given moment depends on the pre-existing level which individuals have because it is this level that leads them to behave in such a way that they reinforce their beliefs.

Fitting the LL-C and R Implementation

We propose to fit the LL-C using a simple, conventional two-stage, limited-information conditioned procedure (McDonald, 1982), with a first calibration stage in which the item parameters are estimated and the goodness of model-data fit is assessed, and a second scoring stage in which trait level estimates and accompanying errors are obtained for all the individuals. Conditional and marginal reliability estimates are also obtained in this second stage. As for overall implementation, the procedure has been implemented in the LILAC-R program, which consists of a main function that, in turn, calls other sub-functions for (a) input setting, (b) calibration and goodness-of-fit assessment, and (c) scoring and score accuracy assessment. LILAC-R is available at: https://www.psicologia.urv.cat/ca/utilitats/lilac-r-code/

Item Calibration

Structural estimation of the LL-C uses the fact that the model can be linearized in the form (6) and consists simply of (a) fitting the standard unidimensional FA to the logit-transformed item scores and (b) re-parameterizing some of the estimates obtained (see Reise et al., 2021 for a graded-response related approach). First, the original item scores in the 0-1 interval are logit transformed, and, to obtain finite transformed scores, the lowest and highest endpoints are fixed at .01 and .99, respectively. Second, the unidimensional FA model is fitted to the transformed item scores. The item structural estimates obtained are: the intercepts (μ_j), the loadings (β_j), and the residual variances (

$\sigma^2_{\varepsilon_j}$). The last two are directly parameters of the original non-transformed LL-C model. As for the location parameters, they are obtained as: $\alpha_j = \exp(\mu_j)$. Fitting procedures and structural goodness-of-fit assessment are conventional and common to any structural model based on continuous variables (e.g. Raykov & Marcoulides, 2012).

In our implementation, the structural model can be fitted with any of the estimation procedures available for continuous variables. The calibration step is implemented through a sub-function that, in turn, needs the 'cfa' function from the lavaan package (Rosseel, 2012), which requires a model syntax object to be specified. In this case, only the single common factor and all the items that act as indicators need to be defined. The output includes a summary table with the LL-C parameter estimates (α_j, β_j, and $\sigma^2_{\varepsilon_j}$), and a list of the following goodness-of-fit measures ($\chi^2, with\ d.f.$, GFI, CFI, RMSR, and RMSEA).

Scoring

Two scoring schemas can be chosen by users: (a) maximum likelihood (ML) and (b) Bayes expected a posteriori (EAP, Bock & Mislevy, 1982). The pros and cons of both are discussed in detail in several articles (e.g. Lord, 1986), and here we shall make only a few comments specifically related to the present proposal. First, the ML point estimates can be obtained in closed form so the typical problems of instability and out of bound estimates generated by iterative schemas (e.g. Lord, 1986) are avoided or at least greatly minimized. And EAP estimation has the advantage that the additional information provided by the prior is not arbitrary but based on the theory from which the model emerges.

The ML point estimates are obtained for each individual using (9). The corresponding standard errors are obtained by using the square roots of the reciprocal of the amount of information (e.g. Samejima, 1977 as obtained from (10).

EAP estimation is totally standard (e.g. Bock & Mislevy, 1982), and the only specifics are that (a) the prior is lognormal ($\mu_U = 0$,$\sigma_U = 1$) and (b) the likelihood function is obtained from (8). For each individual, the output consists of the trait θ_{Ui} point estimate and the corresponding posterior standard deviation (PSD), which serve as standard error estimate (e.g. Bock & Mislevy, 1982).

Finally, we shall consider conditional and marginal reliability estimates (see e.g. Mellenbergh, 1996) derived from both scoring schemas. A conditional reliability estimate for the ML score $\hat{\theta}_{Ui(ML)}$ is obtained as

$$\rho_{\hat{\theta}_{Ui}(ML)} = \frac{Var(\theta_U)}{Var(\theta_U) + (\frac{1}{I(\hat{\theta}_{Ui(ML)})})}. \quad (11)$$

where $Var(\theta_U)$ is the variance of the prior lognormal distribution and $I(\hat{\theta}_{Ui(ML)})$ is the amount of information obtained from (10). The corresponding conditional estimate for the EAP scores is obtained as

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

The marginal reliability estimates are:

$$\rho_{(ML)} = \frac{Var(\theta_U)}{Var(\theta_U) + E(\frac{1}{I(\hat{\theta}_{Ui(ML)})})}. \quad (13)$$

and:

$$\rho_{(EAP)} = 1 - \frac{E(PSD(\hat{\theta}_{Ui})^2)}{Var(\theta_U)}. \quad (14)$$

Provided that the standard errors (ML) or the PSDs (EAP) remain relatively uniform, the marginal reliabilities in (13) and (14) are representative of the overall precision of the estimates in the population of respondents.

The R implementation of the scoring process in LILAC-R is specific to this proposal, and requires users to choose ML or EAP. The ML approach requires the transformed original data matrix (0-1 item scores; dimension respondents × items) and the parameter estimates obtained in the calibration step. With this information, the ML scores, the standard errors and conditional reliabilities are computed and reported in the output. The EAP option needs the same information as ML as well as a file containing a matrix with the quadrature points and nodes. The output consists of a summary table that includes the EAP scores with their corresponding standard errors (PSD), and conditional reliabilities.

Finally, using the mean and variance of the prior lognormal distribution, the marginal reliability is obtained for the chosen option: ML or EAP.

Comparison with the CRM and Linear FA: Expected Distortions and Differential Interpretations

As discussed above, the CRM is the model most closely related to LL-C, and linear FA is by far the most used model in applications based on continuous responses. Given this scenario, in this section we shall (a) assume that LL-C is the correct model for a given set of data, and (b) study the expected consequences of fitting the CRM or linear FA instead of the "true" LL-C. The consequences will be studied at the two stages of calibration and scoring.

Comparisons at the Calibration Stage

From the calibration approach we propose, if the LL-C is the correct model and the CRM is fitted to the data, then it is clear that: (a) after the appropriate transformations, the obtained item LL-C estimates will be unbiased, and (b) the goodness-of-fit results will be correct. In this regard, we can consider the CRM and the LL-C to be indistinguishable at the structural level.

When the linear FA model is fitted directly to the observed item scores, however, it can no longer be considered to be a transformation but an approximation. Essentially, what linear FA will do is approximate the curvilinear IRFs by using straight lines (e.g. Tutz & Jordan, 2022). So, some distortions in terms of model-data fit and biased parameter estimates are expected to occur (e.g. McDonald & Alhawat, 1974, Mooijaart, 1983). Here, we shall study these consequences by using Taylor expansions (Wolter, 2007), which obtain relations that are relatively simple and accurate.

Assume that the observed item scores in the 0-1 metric are directly fitted with the linear FA model. The structural equation is now:

$$X_{ij} = \tau_j + \lambda_j \theta_{Bi} + \omega_{ij} . \qquad (15)$$

The conventional assumption of the FA model is that the trait or common factor is distributed as a standard, usually normal, variable (mean zero and unit variance). So, we have used the term θ_B in (15).

Now, if the covariance between items *j* and *k* is approximated by using a bivariate Taylor expansion, the following relation is obtained:

$$\begin{aligned} Cov(X_j, X_k) &\cong \beta_j \sqrt{\frac{Var(Y_j)}{Var(X_j)}} \beta_k \sqrt{\frac{Var(Y_k)}{Var(X_k)}} \\ &- \frac{1}{4}\left[\frac{(1-\alpha^2{}_j)}{\alpha_j} Var(X_j) \frac{(1-\alpha^2{}_k)}{\alpha_k} Var(X_k) \right] = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} \end{aligned} \quad (16)$$

So, if the unidimensional LL-C model is correct and a direct linear FA solution is fitted, then at least two factors are expected to be needed to fully reproduce the inter-item covariances: a main "content" factor, and a second artifactual, or differential curvature factor (see McDonald & Ahlawat, 1974). If the loadings on the second factor are negligible and a unidimensional linear solution is fitted, then (a) the unidimensional solution will fit the data well, and (b) unbiased estimates of the β parameters of the LL-C model can be readily obtained. If they are substantial, then (a) the unidimensional solution will have a bad fit, and (b) the estimated loadings will be biased: essentially a weighted average of the first and second factor loadings in (16) (see Ferrando, 2009).

Although interpreting the relevance of the second, artifactual loadings is not immediate, it can be shown that the magnitude of the second-factor loading is mainly related to the extremeness and discrimination of the item score: As the item becomes more extreme and discriminating, the magnitude of the loading (regardless of the sign)

increases. This result makes sense, and agrees with related developments (Mooijaart 1983, McDonald & Ahlawat, 1974).

Finally, we shall consider the location parameter. Under all the usual assumptions made with the linear model, the expectation of (15) is the intercept parameter τ_j. Now, if the LL-C is correct, the expected score for X_j will be (approximately):

$$E(X_j) = \tau_j \cong \frac{\alpha_j}{1+\alpha_j} + \frac{Var(X_j)(1-\alpha^2{}_j)}{2\alpha_j}. \qquad (17)$$

So, the intercept estimate will be a function of (a) the α_j location parameter in the LL-C, and (b) and the variance of the observed item scores. This result means that a plausible estimate of α_j can be obtained from the linear model solution.

Comparisons at the Scoring Stage

Only ML trait estimates are considered here so the derived standard errors and reliability estimates will be based on the amount of information. To start with, if the LL-C is correct and the CRM is fitted to the data, the ML score estimates obtained from the latter will be nonlinearly related to the θ_U ML estimates *(i.e.* $\hat{\theta}_B = \ln(\hat{\theta}_U)$*)*. So the θ_B ML estimates will be "stretched" at the lower values of θ_U (i.e. near zero) and compressed at higher values. However, the relation between both sets of estimates will be one to one, and the rank order identical in both cases.

As for the θ_B estimates obtained from the linear model (15), again we are no longer dealing with a transformation but an approximation. What is expected now is that (a) the nonlinear relation between the linear θ_B estimates and the θ_U ML estimates

will be stronger, and (b) this relation will not be one-to-one but there will be some scatter around the regression curve. The relation between the linear θ_B estimates and the θ_U estimates is very similar to that between the simple sum scores and the θ_U estimates.

We turn finally to the comparisons in terms of the conditional accuracy of the estimated scores, which is where the differences become more important. The amount of information in the LL-C model for estimating θ_U is that provided by equation (10). The amount of information in the CRM for estimating θ_B is given by

$$I(\theta_B) = \sum_{j=1}^{n} \frac{\beta^2{}_j}{\sigma^2{}_{\varepsilon j}}, \tag{18}$$

which does not depend on θ_B (see Bejar, 1977, and Wang & Zeng, 1998). Results (18) and (10) agree: the information of the LL-C (10) is that of the CRM (18) divided by the square of the derivative of the transformation of θ_B into θ_U (Lord, 1975).

Finally, the corresponding amount of information provided by the linear model is (e.g. Mellenbergh, 1994)

$$I(\theta_B) = \sum_{j=1}^{n} \frac{\lambda^2{}_j}{\sigma^2{}_{\omega j}}, \tag{19}$$

This is also a constant value. Furthermore, up to the second-order Taylor approximation we are using here, the amount of information in (18) and (19) is expected to be the same (the attenuation terms in the numerator and in the denominator cancel out). The implications of results (10), (18) and (19) are discussed in detail in the empirical study.

Assessing the Appropriateness of the LL-C Model

An appropriate degree of goodness of model-data fit (GOF) is a basic requisite if an IRT model is to be considered appropriate. However, it is not the only requisite (García-Perez, 1999). This is particularly clear here because the two models – LL-C and the CRM –function very differently and are indistinguishable in GOF terms.

Model appropriateness is, then, a comprehensive, multi-faceted concept that includes GOF but which goes far beyond it. In this initial proposal, we suggest that the researcher should take three facets (as well as GOF) into account to judge the appropriateness of the LL-C model. The first facet is the consistency between the functioning of the model and the theory from which the model is derived (Ramsay, 1989, 1996, Reise et al., 2018, 2021). The second is the "internal" properties of the scores derived from the LL-C model as related to their intended use. And the third is the predictive properties of the scores, and how they relate to other theoretically relevant measures.

The initial step in the first facet is to analyze the content to appraise whether the trait under study fits better (in theoretical terms) into the bipolar conception or into the unipolar conception, as discussed below (Reise & Rodriguez, 2016, Reise et al., 2018). At the more empirical level, it is of interest to assess the distributions of the raw item scores (e.g. Bejar, 1977, Reise et al., 2018). For the LL-C model to be appropriate, the raw distributions should generally be right skewed, with a sizable number of cases piled-up at the lower end.

With regard to the internal properties of the LL-C-derived scores, if the researcher or the practitioner is interested only in the relative standings of the respondents on the trait, then the scores estimated by the LL-C model will be as good as those estimated by the CRM. And the scores derived from the simple linear model may

be perfectly appropriate in terms of rank ordering. However, if they are interested in other things – for example, assessing the regions of the trait in which the scores are most accurate, maximizing the differentiation of individuals in the range where they matter most, or establishing cut-off values for diagnostic purposes – then the choice of the model is of the utmost importance (e.g. Santor et al., 1995). We propose three indicators for assessing this facet. The first is the information/conditional reliability curve plotted against both the percentiles (Lord, 1975) and the estimated trait level. The first curve displays the information provided by the LL-C together with the constant information provided by the CRM and the linear model. The second indicator is also a graphic proposed by Ramsay (1989) and Reise et al. (2021): it is the plot of the estimated θ_U scores against the simple sum or raw scores. This graphic assesses the extent to which the relations between the two scores are nonlinear and, if they are, whether the nonlinear trend agrees with theoretical expectations (see below). The third indicator, finally, is the marginal reliability (13) or (14) which informs us about the overall accuracy of the scores across all trait levels.

Finally, we shall discuss the criterion-validity evidence. If the scores derived from the LL-C model were more strongly related to other relevant variables than competing scores (CRM and linear in our case), this result would provide strong support for the LL-C appropriateness (Reise et al., 2022). However, although the choice of the scale can affect the validity correlations, mainly in terms of non-linearity and end effects (Yen, 1986), if this result is observed, we believe that its magnitude will be rather small. In our opinion, it is more interesting to study whether the LL-C scores show differential validity effects with respect to the relevant external variable or criterion. More in detail, if the trait is really unipolar and has most meaning at the upper end, then the relation with the external variable should be stronger at upper trait levels

than at lower trait levels. This hypothesis can be assessed by inspecting the dispersions around the regression line (see Morales-Vives et al., 2022).

Empirical Study: Extreme Beliefs

In this empirical study, we used a set of 7 items referring to superstitious beliefs, which were taken from the general pool used in the study by Huete-Pérez et al. (2022) about beliefs not supported by science. The item contents are in Table 2. These items were administered to 1097 individuals aged between 18 and 69 years old (59.6% women). Of these, 2% had an elementary education, 15% had a secondary education, 65% were undergraduates or master's students, and 18% had finished a postgraduate course. Participants were asked to rate each item on a continuous scale, indicating the extent to which they agreed with the belief or behavior described in the item by placing a mark on a continuum that ranged from 0 (totally disagree) to 100 (totally agree).

The second and third columns in table 2 show the medians and the skewness coefficients of the superstitious beliefs item scores. Except for two items (1 and 5), the medians clearly show that a substantial number of cases are concentrated at the low end of the 0-1 response scale, so the items are positively skewed. The two items that do not follow this trend refer to behaviors that are relatively common and well accepted by the general population: many people use some kind of good luck charm or cross their fingers to wish for luck, even if they do not necessarily fully believe in these behaviors.

For both the LL-C and the linear model, item parameters were estimated using robust (mean and variance corrected) ML estimation as implemented in LILAC-R. Goodness of fit results are in Table 1. Although the fit of the LL-C is consistently better across all indicators, the differences are clearly negligible: both solutions fit the data quite well and are very similar. What these results suggest is that in this study the

second differential curvature factor in (16) is negligible so the linear approximation to the LL-C IRFs is consistently good for all of the items.

INSERT TABLE 1 ABOUT HERE

Table 2 shows the calibration results based on the LL-C solution and the approximate β estimates obtained from the linear model using (16) in the last column. In general, the items are quite discriminating (high βs) and "difficult", which is expected (Reise et al., 2018, 2021). The range of discrimination, however, is rather wide and the most discriminating item (4) is about 3.5 times more discriminating than the "worst" item (3). As for location, the values can be interpreted by noting that the mean and mode of the prior lognormal θ_U distribution are 1.65 and 1, respectively. So, except for item 1, all difficulties are clearly above the trait mean. Finally, the most "odd" item is item 3, with a rather low β estimate, which in turn implies exceedingly large difficulty. To understand this result, the IRF of this item would be like the lowest curve in Figure 1: a curve that flattens quickly and which has great difficulty reaching the 0.5 threshold. One possible explanation is that, unlike other items, this one seems to be more related to impatience and impulsivity than to superstitious beliefs: If a person feels impatient and bored waiting for the lift, he/she may start pressing the button to have a sense of physical control over the situation, which does not necessarily involve a superstitious idea.

Finally, note that the linear β estimates are quite close to those obtained from the LL-C solution, a result which agrees with the results in table 1: when the linear approximation is good, both the model-data fit and unbiased item parameter estimates are expected to be acceptable.

INSERT TABLE 2 ABOUT HERE

We turn now to scoring. ML score estimates were chosen in this study, and we shall now discuss their internal properties. Figure 2 shows the conditional reliability (information) curves plotted against the percentile (panel a) and the trait estimates (panel b). The common percentile metric can be used to plot the constant reliability provided by the linear and the CRM scores (thin straight line) and the LL-C curve (thick curve). Two results are apparent in panel (a). First, up to about the 70th percentile the reliability of the LL-C scores is quite high. And, second, up to about the 65th percentile, the reliabilities of the ML scores derived from the LL-C model are consistently better than the constant reliability derived from the CRM and the linear model. So the relative efficiency of the LL-C scores is greater than that of the competing scores until well past the first half of the percentile scale. Conceptually, these results mean that the LL-C scores measure individuals more efficiently in the lowest-to-medium end of the trait level. In a unipolar trait, these are the vast majority. In contrast, the competing scores are more accurate for the few individuals in the upper percentile. Again, in agreement with the discussion above, the LL-C scores are more useful for differentiating between "asymptomatic" respondents (in this case, participants without superstitious beliefs or with low levels of such beliefs) and "symptomatic" respondents (superstitious participants, with high scores on these items). Panel (b) also provides complementary information: the LL-C scores are highly reliable only between 0 and about 5 units. However, the range of estimated scores extends beyond 65 points. Finally, to complete these results, the marginal reliability of the LL-C scores was .97, slightly higher than the constant .95 of the competing scores.

INSERT FIGURE 2 ABOUT HERE

Still within the "internal" facet, figure 3 plots the θ_U ML estimates against the raw scores. The relation is clearly non-linear and, at the low end of the scale, the

changes in raw score units hardly change the trait estimates. However, as the number of endorsed superstitious beliefs increases, the average trait estimate increases more and more. Note that this behavior is consistent with the law of proportional effects: the increase in the trait level estimate as more beliefs are endorsed is proportional to the number of previous beliefs endorsed.

INSERT FIGURE 3 ABOUT HERE

We turn finally to the third, criterion validity facet. A sub-sample of 137 respondents was administered a single-survey question asking them to report their degree of trust in the government on a 0-100 scale (no trust at all - complete trust). The product moment correlation between the θ_U estimates and the trust ratings was negative but very low (r = -.14). However, the scatterplot in figure 4 provides interesting information. First, the relation is clearly heteroskedastic, with a large dispersion at lower θ_U values and a small dispersion at higher values (the twisted-pear effect; Fisher, 1959). Second there is clearly a differential validity effect, as there is virtually no relation at low θ_U values and a far stronger relation as θ_U increases. Furthermore, this relation seems to be non-linear. To assess these trends, we first computed the product-moment correlation only for the cases with θ_U estimates above 5 (see figure 4). The correlation was now r = -.56. Next, we fitted an exponential curve to this subset (displayed in the figure) and obtained a non-linear correlation estimate (square root of the eta correlation ratio) of r = -.58).

INSERT FIGURE 4 ABOUT HERE

Conceptually, the results above make sense and provide support for the appropriateness of the LL-C. The lack of superstitious beliefs does not predict at all the level of trust in the government: people with few or no irrational beliefs may have

different levels of trust. As the number of superstitious beliefs increases, however, people tend to distrust the government more and more. Therefore, it seems that those people who tend to have irrational beliefs (in this case, superstitious beliefs) are also likely to distrust official messages and all those entities that represent this "official" discourse, such as governments. In contrast, non-superstitious people probably trust in the government for many reasons, such as their political, sociological or economic viewpoint, which may or may not coincide with that of the governing party.

Discussion

This article proposes and develops an IRT model intended for continuous item responses and unipolar traits. As for the first feature, the continuous format is submitted to be a natural application for continuous responses that has advantages in terms of the administration and the methodological properties of the resulting models. Furthermore, the development of computerized questionnaires has made this response format more attractive to researchers and practitioners, since they enable these items to be more easily scored than with paper-and-pencil questionnaires. Considering also that data is increasingly being collected online, this response format is expected to be used more and more in the future. The LL-C model we propose is also remarkably simple, and more closely resembles the original binary formulation than the graded extensions. Finally, at the substantive level, the present study focuses on superstitious beliefs, a variable that could plausibly be considered unipolar, although it has traditionally been studied as if it were bipolar (e.g., Fasce et al., 2021; Renken et al., 2015).

The LL-C model can be made equivalent to a variation of the CRM in which the trait scale is transformed, and it can also be linearized to adopt the structure of the standard FA model. For this reason, we have compared its functioning to that of the

other two models when it is used in the standard form: i.e. it assumes that the trait is bipolar (CRM) and fits the model directly to the observed item responses (linear FA). Admittedly, this is only an initial step. In recent years, there has been renewed interest in continuous responses, and new IRT models have been proposed. It would be of interest to study whether these models can be adapted to the unipolar case and, if so, what advantages they may have.

Apart from the methodological contributions, we would like to emphasize that this article has attempted to provide a substantive rationale for what we are proposing, and also includes an empirical example to illustrate the results obtained with different models. In this example, we used seven items of superstitious beliefs, most of which, as expected, had distributions consistent with a unipolar formulation. Furthermore, the items were generally quite discriminative (high βs), which is also consistent with the unipolar formulation. In addition, the linear model fitted the data well and allowed the LL-C parameters to be closely approximated, which means that the impact of the curvature factor was very low. Therefore, in this particular example, the linear model is a good approximation of the LL-C model. This does not mean, however, that these two models will always be made equivalent when all the items have an unequivocally unipolar nature (in the present example this was not the case, since the means of two items were higher than expected). In solutions with a more important second curvature factor, the linear model is expected not to provide as good a fit as the LL-C model. This result might be expected with some irrational beliefs that are more extreme or polarizing than superstitious beliefs (for example, flat-eartherism). Further studies are required to assess this issue.

Where the alternative models clearly differ is in the trait estimates and their accuracy. In the linear model, as in the CRM, the reliability value of the estimated

scores is constant for different levels of the trait, while the results obtained with the LL-C model show that it provides more accurate estimations at medium and low levels. These results are consistent with those observed for clinical variables (e.g. Morales-Vives et al., 2022), suggesting that the LL-C is preferable to the other models when the aim is to discriminate 'symptomatic' from “non-symptomatic” cases, rather than to discriminate between different high levels of the trait. This may make a lot of sense when the aim is to carry out campaigns to counteract hoaxes or false beliefs with possible consequences on society (as was the case with the pandemic and the vaccines). In these cases, the important point is to identify those people who may be prone to irrational beliefs, in order to target them, instead of differentiating between high levels of severity.

Another important result is the differential predictability or criterion validity effect, expected if the variables are truly unipolar (i.e. more meaningful at one end of the continuum). This effect was clear in our example: superstitious people tend to distrust the government (probably because they distrust the prevailing discourses and ideas, some of which are represented by the government), but no such negative relation is observed for the other participants (probably because in these cases trust or distrust in the government depends on many variables).

Like many initial proposals of this type, this one has its share of limitations so further research is required. As mentioned above, some promising models for continuous variables could be transformed to make them unipolar. Likewise, the estimation procedures we have proposed (specially at the calibration stage) are extremely basic, and can be considerably refined. However, although further refinements would no doubt improve estimates, they would not change anything essential. Furthermore, the validity evidence is only based on a single criterion, and we

fully acknowledge that this is a limitation of the study. Further studies would be clearly needed in order to obtain further evidence and a better understanding about the impact the unipolar variables may have on criterion validity relations, and, if this impact is consistent, about the applicability of these results in empirical settings.

In this article we have considered the bipolarity-unipolarity distinction with respect to the nature of the trait. However, a similar distinction has been made in the literature regarding the polarity of the item response scales, and, furthermore, it has been found that the use of unipolar or bipolar scales can impact the accuracy of the derived trait scores (Menold & Raykov, 2016). An interesting topic for further research would be then to assess whether the use of unipolar response scales that match the (assumed) unipolarity of the trait under study can lead to a better functioning of the model, especially in terms of accuracy/information and criterion validity. Hopefully, this and other topics will be covered by future research.

References

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

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

Bottesi, G., Ghisi, M., Altoè, G., Conforti, E., Melli, G., & Sica, C. (2015). The Italian version of the Depression Anxiety Stress Scales-21: Factor structure and psychometric properties on community and clinical samples. *Comprehensive psychiatry*, *60*, 170-181. https://doi.org/10.1016/j.comppsych.2015.04.005

Dolnicar, S., Grun, B., Leisch, F., & Rossiter, J. (8 – 11 February 2011). Three good reasons NOT to use five and seven point Likert items. CAUTHE 2011: 21st CAUTHE National Conference, Adelaide, Australia

Fasce, A., Avendaño, D., & Adrián- Ventura, J. (2021). Revised and short versions of the pseudoscientific belief scale. *Applied Cognitive Psychology*, *35*(3), 828-832. https://doi.org/10.1002/acp.3811

Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item response. *Multivariate Behavioral Research*, *37*(4), 521-542. https://doi.org/10.1207/S15327906MBR3704_05

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

Fisher, J. (1959). The twisted pear and the prediction of behavior. *Journal of Consulting Psychology*, *23*(5), 400-405. https://doi.org/10.1037/h0044080

García-Pérez, M.A. (1999). Fitting logistic IRT models: Small wonder. *The Spanish Journal of Psychology*, *2*, 74-94. https://doi.org/10.1017/S1138741600005473

Huete-Pérez, D., Morales-Vives, F., Gavilán, J., Boada, R., & Haro, J. (2022). PEUBI: Development and Validation of a Psychometric Instrument for Assessing Paranormal, Pseudoscientific and Conspiracy Beliefs in Spain. *Applied Cognitive Psychology*, *36*(6), 1260-1276. https://doi.org/10.1002/acp.4010

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*(1/2), 149-176. https://doi.org/10.2307/2332539

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. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2), 157-162. https://www.jstor.org/stable/1434513

Lucke, J. F. (2013). Positive trait item response models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, and C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 199–213). Springer.

Lucke, J. F. (2015). Unipolar item response models. In S. P. Reise and D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 272–284). Routledge/Taylor & Francis Group. 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

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of mathematical and Statistical Psychology*, *27*(1), 82-99. https://doi.org/10.1111/j.2044-8317.1974.tb00530.x

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223-237. https://doi.org/10.1207/s15327906mbr2903_2

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293–299. https://doi.org/10.1037/1082-989X.1.3.293

Menold, N., & Raykov, T. (2016). Can reliability of multiple component measuring instruments depend on response option presentation mode?. *Educational and Psychological Measurement*, *76*(3), 454-469. https://doi.org/10.1177/0013164415593602

Molenaar, D., & Dolan, C. V. (2018). Nonnormality in latent trait modelling. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, (pp. 347-373). John Wiley & Sons Ltd. https://doi.org/10.1002/9781118489772.ch13

Morales-Vives, F., Ferrando, P. J., & Dueñas, J. M. (2022). 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

Mooijaart, A. (1983). Two kinds of factor analysis for ordered categorical variables. *Multivariate Behavioral Research*, *18*(4), 423-441. https://doi.org/10.1207/s15327906mbr1804_5

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220. https://doi.org/10.1037/1089-2680.2.2.175

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo chambers on Facebook. *SSRN Electronic Journal, 2795110*, 1-15. https://doi.org/10.2139/ssrn.2795110

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. (1996). A geometrical approach to item response theory. *Behaviormetrika*, *23*(1), 3-16. https://doi.org/10.2333/bhmk.23.3

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*(1), 49–57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x

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

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., Mansolf, M., & Haviland, M. G. (2022). Bifactor Measurement Models. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 329-348). Guilford Press

Reise, S.P., & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: Some empirical and conceptual issues and challenges. *Psychological Medicine*, *46*, 2025–2039. https://doi.org/10.1017/S0033291716000520

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

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

Renken, M.D., McMahan, E.A., & Nitkova, M. (2015). Initial validation of an instrument measuring psychology-specific epistemological beliefs. *Teaching of Psychology, 42*(2), 126-136. https://doi.org/10.1177/0098628315569927

Rosseel, Y. (2012). lavaan: An R package for structural equation modelling. *Journal of Statistical Software, 48*, 1-36.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Psychometric Society. http://www.psychometrika.org/journal/online/MN17.pdf.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*(2), 203-219. https://doi.org/10.1007/BF02291114

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement*, *1*(2), 233-247. https://doi.org/10.1177/014662167700100209

Sánchez-Álvarez, N., Extremera Pacheco, N., Rey Peña, L., Chang, E. C., & Chang, O. D. (2020). Frequency of suicidal ideation inventory: Psychometric properties of the Spanish version. *Psicothema*, *32*(2), 253-260. https://doi.org/10.7334/psicothema2019.344

Santor, D.A., Zuroff, D.C., Ramsay, J.O., Cervantes, P., & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment, 7*(2), 131 –139. https://doi.org/10.1037/1040-3590.7.2.131

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

Tutz, G., & Jordan, P. (2022). Latent Trait Item Response Models for Continuous Responses. *arXiv:2204.03841*. https://doi.org/10.48550/arXiv.2204.03841

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*, 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

Westerwick, A., Johnson, B. K., & Knobloch-Westerwick, S. (2017). Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs*, *84*(3), 343-364. https://doi.org/10.1080/03637751.2016.1272761

Wolter, K. M. (2007). Taylor series methods. In K.M. Wolter, *Introduction to variance estimation*, (pp. 226-271). Springer.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299-325. https://doi.org/10.1111/j.1745-3984.1986.tb00252.x

Zhang, J., Sun, L., Liu, Y., & Zhang, J. (2014). The change in suicide rates between 2002 and 2011 in China. *Suicide and Life‐Threatening Behavior*, *44*(5), 560-568. https://doi.org/10.1111/sltb.12090