Current Psychology
https://doi.org/10.1007/s12144-022-03224-6

Should suicidal ideation be regarded as a dimension, a unipolar trait or a mixture? A model-based analysis at the score level

Fabia Morales-Vives[1,2] · **Pere J. Ferrando**[1,2] · **Jorge-M. Dueñas**[1,2]

Accepted: 12 May 2022

Abstract

Screening questionnaires administered in community samples may allow to early identify suicidal ideation (S.I.). Although the results found in these samples suggest that S.I. behaves like a unipolar trait or a quasi-trait, it is routinely assessed using procedures developed for bipolar traits. Therefore, the main aim of this study is to determine whether there is a basis for modelling S.I. as a bipolar trait, a unipolar trait, or a quasi-trait with two classes of individuals (symptomatic and asymptomatic). In a community sample and mainly at the scoring level, we compare the results provided by fitting three models based on different assumptions: GRM (bipolar traits), LL-GRM (unipolar traits) and FMA (quasi-traits). 773 Spanish participants answered a S.I. and a life satisfaction questionnaires. GRM and LL-GRM provided equivalent results at the structural level, but not at the scoring level, especially in the conditional and marginal accuracy of the estimated scores. While the GRM scores are highly accurate only in a narrow range well above the mean, the LL-GRM scores are highly accurate in a much wider range around the mean. They also have different implications for the prediction of life satisfaction. FMA results suggest that an asymptomatic and a symptomatic class could not be clearly differentiated. In conclusion, LL-GRM would make it possible to accurately measure a larger number of subjects in a community sample than GRM, leaving fewer cases of vulnerable people unidentified. These results should be considered by researchers and professionals when deciding which modellings to use for screening purposes.

Keywords Suicidal ideation · Unipolar trait · Quasi-trait · Graded-Response model · Log-Logistic model · Factor Mixture Analysis

According to the World Health Organization (2021), every year more than 700,000 people commit suicide. In fact, suicide was the fourth leading cause of death among adolescents and young people in 2019 (World Health Organization, 2021). For this reason, suicidal behaviour is considered to be a serious health problem and a global challenge throughout the life course, including youth.

When addressing this problem, suicidologists and public health officials tend to focus on suicides and suicidal behaviours, which means that people with suicidal ideation (S.I.) receive less attention than they need in prevention research, clinical treatments and health care policies (e.g. Jobes & Joiner, 2019). However, thoughts of suicide are the first sign of possible future suicidal behaviour (Nock et al., 2008). Furthermore, according to Franklin et al. (2017), the main risk factor for presenting S.I. is having previously suffered from S.I. Therefore, the early identification of those people with S.I. would help in reducing suicidality in adolescence and youth, and preventing S.I. and suicidal behaviours in the long term.

Instead of waiting for the first signs of psychopathology or suicidal behavior to manifest, one way of identifying S.I. early is to use screening tools such as questionnaires on a community sample from the general population (as opposed to a clinical sample). In this type of scenario, many clinical instruments (not only S.I. questionnaires) have been observed to behave as if they were measuring a type of construct different from those measured by cognitive tests or normal-range questionnaires. More specifically, many instruments of this type appear to assess what Reise and Waller (2009) called unipolar traits or "quasi-traits", which means

✉ Fabia Morales-Vives
fabia.morales@urv.cat

[1] Psychology Department, Universitat Rovira i Virgili, Tarragona, Spain

[2] Research Center for Behavior Assessment (CRAMC), Tarragona, Spain

Published online: 28 May 2022

traits that are only (or mostly) relevant in one direction, with the variation at the other end of the scale being less informative from both a substantive and a psychometric perspective. Thus, in clinical traits of this type, one end of the scale represents severity while the other end represents absence of the pathology: for example, depression at one pole and lack of depression (which is something very different from happiness) at the other pole. In contrast, bipolar or dimensional traits have meaningful variation at both poles of the trait (for example, extraversion at one pole and introversion at the other). Furthermore, the distribution of the clinical unipolar or quasi traits tend to be positively skewed in the general population (e.g. Magnus & Liu, 2018a, b), because most people have no symptoms (or very low levels in the construct) and so are grouped at the lower end of the scale, with no, or very little, variability between them. In contrast, the upper end of the scale reflects variability at the higher levels of severity of the pathology (Reise & Waller, 2009).

So far, unipolar traits and quasi-traits have been used as essentially interchangeable terms. However, Reise et al. (2018) proposed a distinction which we consider to be highly relevant here: The term "quasi-trait" refers to a construct which is **only** relevant at one end (generally the upper end) of the continuum whereas the other end only reflects absence of the disorder and is totally irrelevant. In contrast, the term "unipolar' trait refers to a construct that is **far more** meaningful at one end of the continuum than at the other end.

Although the view summarized so far seems quite plausible, screening clinical instruments of this type (and particularly those that deal with S.I.) are typically fitted and scored using the standard methodology that was initially developed for measuring bipolar traits, an approach that Reise et al. (2018) referred to as "business-as-usual" analysis. However, if the trait under study truly behaves as a unipolar or a quasi-trait, the use of the standard bipolar approach might have important implications for the estimation of the individual scores and, therefore, for further decisions based on these scores, such as determining appropriate cut-off values for identifying at-risk individuals. In this respect, it seems that most studies about clinical questionnaires tend to focus on the estimation of the structure underlying the data and the marginal reliability of the scores (e.g., Elhai et al., 2012; Otani et al., 2021), but not on the distribution of the trait estimates or on whether the measurement precision of the scores (conditional reliability) varies across the different levels of the trait (Murray et al., 2017).

The most reported result is that the distribution of S.I. scores in community samples is highly and positively skewed (e.g. Norr et al., 2016), but it is usually ignored, and the S.I. construct has routinely been treated as bipolar (e.g., Chang & Chang, 2016; Reynolds & Mazza, 1999). It is submitted here, however, that S.I. should be more properly modeled as a unipolar trait or a "quasi-trait". If the former, the most plausible approach would be to use a basic psychometric model that conceptualizes the construct in this way. If the latter, a plausible approach would be to model it as a mixture of two classes of respondents: one without S.I., and one with high, differentiated levels of S.I. (e.g. Smits et al., 2020).

Neither of the possible modelling approaches above seems to have been used to date in the S.I. domain, and, according to the literature review, the psychometric analyses of S.I. item responses across different measures have mainly been based on three types of model: Classical Test Theory (CTT; e.g., Cotton et al., 1995; Reynolds & Mazza, 1999), Linear Factor-analysis (LFA; Pinto et al., 1997; Sánchez-Álvarez et al, 2020; Zhang et al., 2014) and non-linear FA models or Item Response Theory (IRT) models that can be parameterized as FA models (see Ferrando & Lorenzo-Seva, 2013), particularly one- and two-parameter models (e.g., Díez Gómez et al., 2020; Núñez et al., 2019) and the graded response model (GRM; e.g., De Beurs, et al., 2014; Fitzpatrick et al., in press; Nugent, 2005, 2006). All these models are intended for measuring dimensional traits, especially under the standard prior assumption that the trait under study has a normal distribution (e.g. Mislevy, 1984).

With regards to the mixed approach, the only study we are aware of is the taxometric analysis by Liu et al. (2015). Their results suggests that S.I. is dimensional, no categorical. However, their study was based not on a heterogeneous sample (the scenario considered here) but on a clinical sample of depressed, treatment-seeking adolescents, a setting that makes it more difficult to identify a profile of individuals without S.I. Moreover, their approach did not take into account that S.I. could be both dimensional and categorical, the view adopted in the current study.

Plausible Modeling Approaches for SI Responses

This paper discusses two plausible approaches for modeling the responses to an S.I. instrument made up of graded-response items, and they are compared with each other and with the standard modeling that views S.I. as a dimensional construct with a normal latent distribution. As discussed above, this modeling is intended for a scenario in which the S.I. measure is administered in a community sample from the general population for screening purposes. Below a brief non-technical review of the models is provided, together with a summary of how they are expected to function when applied to the scenario considered here. Because there are practically no previous studies in which the two new modellings above have been used with S.I., most of the expectations are based on results obtained in similar scenarios but

different domains, such as depression, addictions, or prevalence of mental disorders in general.

The standard IRT model used for calibrating clinical measures is Samejima's (1969) graded-response model (GRM), which, in the case of S.I., has been used in previous studies by De Beurs, et al. (2014), Fitzpatrick et al. (in press), and Nugent (2005, 2006). In its logistic version, the probability of obtaining a graded score of k or greater on item j, (denoted by the cumulative operating characteristic; COC) for a given trait level θ_1 is

$$P\left(X_{ij} \geq k \mid \theta_1\right) = \frac{1}{1 + \exp\left(-\alpha_j(\theta_1 - \beta_{jk}\right)}. \quad (1)$$

In Eq. (1) the trait θ_1 is assumed to follow a standard normal distribution, so it ranges from $-\infty$ to $+\infty$ (i.e. a bipolar dimension). The item parameters α_j and β_{jk} are the discrimination or slope parameter, and the location or threshold parameter for category k, respectively (more discussion is provided below).

Although the GRM with the latent normality assumption would clearly be inappropriate if S.I. behaves as a unipolar trait, the practical relevance of using the "wrong" GRM in this type of scenario is far from clear. Based on the literature review, our impression is that the use of the GRM and other standard IRT models with clinical measures has generally led to acceptable fits and meaningful interpretations but with some odd results. At the calibrations level, there are mainly two. First, a good range of item locations cannot be generally found, and most items have location parameters concentrated at the upper end of the scale (Reise & Waller, 1990, 2009; Reise et al., 2021; Smits et al., 2020). Second, in many items, the discrimination parameter estimates are unusually high compared to those of normal-range personality measures (Magnus & Liu, 2018a, b; Reise & Waller, 2009; Smits et al., 2020).

The impact of fitting the GRM at the score level has been less studied but some consequences can be readily deduced from the item results above. A narrow range of item locations together with very high discriminations implies that the information function will be very peaked around a narrow trait range. In turn, this result implies that the score estimates will only provide accurate measurement around a narrow range of trait levels and will tend to display end (floor or ceiling) effects.

The alternative model to the GRM considered here, and which explicitly treats the construct as a unipolar trait, is the polytomous version of the Log-Logistic (LL) model proposed by Lucke (2013, 2015). It is denoted as LL-GRM and it can be viewed as a psychometric adaptation of Stevens's (1975) power law. It was derived by Lucke from substantive theory and developed specifically for measuring addiction (Lucke, 2013). Here, the use of the LL-GRM is based on more pragmatic grounds. First, it is regarded as a potentially versatile and useful model for fitting a variety of clinical measures (not only addictive disorders) that can be thought to measure narrow-bandwidth unipolar traits with a positively skewed latent distribution (Magnus & Liu, 2018a, b; Reise & Rodriguez, 2016; Reise et al., 2018, 2021). Second, the general functioning of the LL-GRM can be related to that of the standard GRM (Lucke, 2015) with a different COC and a different latent trait distribution. In more detail, the COC of the LL-GRM is given by

$$P\left(X_{ij} \geq k \mid \theta_2\right) = \frac{\varepsilon_{jk}\theta_2^{\alpha_j}}{1 + \varepsilon_{jk}\theta_2^{\alpha_j}}. \quad (2)$$

In Eq. (2), trait θ_2 is now assumed to follow a log-normal distribution, so it only takes positive values and has a right-skewed distribution. The item parameters α_j and ε_{jk} are, respectively, the discrimination or slope parameter, and an "easiness" parameter that reflects the endorsement rate for category k. The discrimination parameter is defined in the same way as in (1).

As mentioned above, the LL-GRM does not appear to have been used to date for measuring S.I. However, information is available about how it is expected to function in similar clinical scenarios (Reise & Rodriguez, 2016; Reise et al., 2018, 2021). At the calibration level, the slopes are expected to be the same as those of the GRM. However, if the easiness parameters are transformed to make them equivalent to the GRM locations, results vary. Unlike the GRM locations, most LL-GRM locations are expected to concentrate at the lower end of the scale while a few are found more expanded at the upper end. At the scoring level, and in agreement with the location results, most of the information is expected to be found in the lower trait range, and the estimated scores are expected to be compressed at the low end of the trait level and expanded at the high end (Magnus & Liu, 2018a, b; Reise & Rodriguez, 2016; Reise et al. 2018).

Reise and coworkers (Reise & Rodriguez, 2016; Reise et al., 2018, 2021) have shown that, at the calibration level, the parameters of the LL-GRM can be obtained by transforming the GRM parameters. This suggests a simple procedure for fitting the LL-GRM from the calibration results obtained from the GRM. More important in this section, however, is that the equivalence above implies that the structural fit of both models to the inter-item correlation matrix is expected to be the same. This result, in turn, explains why the use of the 'wrong' GRM with clinical measures still generally leads to acceptable levels of model-data fit (e.g. Lucke, 2013) and further justifies placing the focus of attention on the score estimates, as it is intended to do here.

Finally, the second modeling approach considered here is Factor Mixture Analysis (FMA) modelling (e.g. Lubke & Miller, 2015; Lubke & Muthén, 2005), which uses a

hybrid view that is both categorical and continuous. The FMA approach assumes that the measure under study has a dimensional structure that can be modelled using FA but also that different clusters or classes of people can be distinguished within this structure. Graphically, the factor solution in FMA provides a dimensional basis on which individuals can be represented by points or vectors depending on their estimated scores. Now, if different classes truly exist, these points will tend to cluster in different clouds (classes) that can be well differentiated (i.e. non-overlapping clusters and a good distance between centroids).

The FMA solution considered here has two basic specifications. First, it considers a two-class solution: one with no symptoms (no S.I.) or with very low levels in the construct, and another with symptoms that has high and differentiated levels of S.I. Second, the basic FA model is non-linear and based on an underlying-variables approach (UVA, e.g. Muthén, 1993). So, the FMA modelling considered here is, in fact, the normal-ogive version of the graded-response IRT model in Eq. (1) (see Ferrando & Lorenzo-Seva, 2013) from which two classes of respondents can be distinguished.

As mentioned above, the FMA does not appear to have been used in S.I. applications. However, related approaches have been used in similar scenarios. Wall et al. (2015), for instance, proposed an IRT-based zero-inflated mixture model which (a) identifies essentially the same two classes as above, and (b) estimates the IRT parameters based only on the 'symptomatic' class of interest. The model by Wall et al. (2015), however, was developed for binary responses. Along the same lines, Magnus and Liu (2018a, b) proposed a mixture model for polytomous responses based on the LL-GRM (see also Smits et al., 2020). Like Wall et al. (2015), Magnus and Liu (2018a, b) aimed to calibrate the IRT model only in the 'symptomatic' class. So, both approaches explicitly treat the construct under study as a quasi-trait and are only interested in modelling and scoring the individuals for which the construct is relevant. Our interest in using FMA here, however, is more exploratory. Our main aim is to determine whether there is a basis for modelling S.I. as a quasi-trait and, therefore, whether the two hypothesized classes can be well differentiated.

Purposes of the Present Research

As pointed out above, the aim of the present study is to compare the results of fitting the GRM, the LL-GRM and the FMA solutions to the (graded) responses of an S.I. inventory. As has been mentioned above, hardly any previous studies compare these modellings in the field of suicidal ideation and assess their different results and implications, which justifies the need for this study. Furthermore, the focus of the current study is not on the structural results in the calibration stage but on the scoring results. This level of analysis was chosen, not only because it has received less attention, but also because most clinical decisions are made on the basis of an individual's scores. The following examples show the importance of this issue for S.I. Reynolds (1988) proposed to use a cutoff score of 41 in the Suicidal Ideation Questionnaire (SIQ, Reynolds, 1988) to identify those individuals who should be referred for further suicidal evaluation, and this cutoff was established by the frequency distribution of S.I. raw scores in a nonclinical sample. Likewise, Joiner et al. (2002) suggested that individuals with a score equal to or higher than 0 on the Depressive Symptom Inventory-Suicidality Subscale (DSI-SS, Metalsky & Joiner, 1997) might require further assessment, considering the frequency distribution of the suicidality scores in a community sample of adolescents and young adults. Other authors have proposed cutoff scores based on the receiver operating characteristic (ROC) curve, also taking as a starting point individual scores on suicide ideation questionnaires (e.g., Holi et al., 2005; Osman et al., 2001). Therefore, it is important to focus on the implications of the different models on S.I. score estimates, since researchers and screening professionals use the scores to make decisions and to differentiate between individuals with and without severe S.I.

The present research focuses on three main properties of SI score estimates: (a) their distribution, (b) their quality, determinacy and accuracy, at both the conditional and marginal levels, and (c) their relation to particular external variables. Properties (a) and (b) can be viewed as "internal" and are particularly relevant to assessing which type of respondent will be accurately differentiated on the basis of their scores. Property (c) is "external", and can provide further evidence about the nature of the SI construct in terms of differential predictability at different levels and for different individuals (Ghiselli, 1956).

Method

Participants

Participants were 773 Spanish individuals (58.9% women) aged between 15 and 29 years old (M = 19.3, S.D. = 3.0). Of this sample, 47.2% were secondary education students, 12.3% were undergraduate students and 40.5% were workers. A total of 14.1% of workers had finished primary education, 41.9% had finished secondary education and 44.0% had finished university studies.

Instruments

The scale used was the modified Spanish version of the *Scale for Suicide Ideation* (SSI) (Beck et al., 1979) developed by

Villardón (1993) and specifically intended for young people over 14. This version is intended to measure a single factor of desire for suicide. It is made up of 10 items with four response choices ranging from 1 to 4. More specifically, each item consists of four statements that reflect various levels of S.I., with option 1 being the lowest (lack of wishes, thoughts or plans related to suicide) and option 4 the highest. The estimated reliability (Cronbach's alpha) of the SSI total scores in our study was $\alpha = 0.89$.

Satisfaction with Life Scale (SWLS) (Diener et al., 1985). The Spanish adaptation developed by Atienza et al. (2000) was used. This questionnaire evaluates satisfaction with life, understood as the cognitive judgements that people make about the satisfaction with their own life as a whole, based on their objectives, expectations, values and interests. It is intended to measure a single dimension, and it is made up of 5 items with a Likert response format (1 = strongly disagree, 5 = strongly agree). The estimated reliability of the SWLS raw scores in our study was $\alpha = 0.83$.

Procedure

The project and the protocol of this study was approved by the Ethical Committee of the Faculty of Educational Sciences and Psychology of the Universitat Rovira i Virgili. This study was carried out in accordance with the recommendations of Spanish Organic Law 3/2018, of 5 December, on the Protection of Personal Data and Guarantee of Digital Rights and the Spanish Agency for Data Protection, which regulate the fundamental right to the protection of data. According to this legislation, the personal data of minors can be processed with their own consent when they are older than 14, so parental consent is not required for minors who are 15–17 years old. As the questionnaires were administered online, through a survey designed for this purpose, and all the participants were 15 years of age and older, participants only needed to give their own informed consent. In fact, the exclusion criteria were being under 15 or over 29 years old, not resident in Spain, and not providing informed consent. Participants had to accept the conditions of the study before participating and could decide to drop out at any time.

The survey was anonymous, and confidentiality and data protection were guaranteed. It included information about the main goals of the study, and instructions on how to answer each questionnaire. It also included information on the voluntary nature of participation. The survey was disseminated in several ways: 1) through WhatsApp groups, Facebook and Twitter, 2) through several Spanish associations, which sent the survey to their members, 3) through high schools from different regions in Spain (teachers were asked to send it to their students who were 15 and older). Once the participants had finished the questionnaire, the website allowed them to share it with other people on the social networks who met the inclusion criteria (e.g. WhatsApp and Facebook). This is known as a non-probabilistic "snowball" sampling procedure.

This research was part of a wider study, and for this reason the survey included other questionnaires in addition to SSI and SWLS. Although the SSI was present in all cases, the other questionnaires appeared randomly, so not all the participants answered all the questionnaires and the survey did not become too long and tiring. For this reason, only 363 participants out of 773 answered the SWLS questionnaire.

Data Analysis

It should first be taken into account that there were no missing values, because the online survey did not allow any item to be left without response. The items appeared one at a time, and it was not possible to move on to the next item if the item on display remained unanswered. Furthermore, in order to reduce outliers and anomalous response patterns, those subjects who answered the questionnaires in less than 5 min were not included in the sample, since this was considered insufficient time to have paid adequate attention to the items, and responses could be random or not very serious. Therefore, after the responses of those participants who had answered the questionnaire too quickly had been removed, the resulting sample was 773 subjects.

Analyses were conducted in four stages. In the first, preliminary stage, basic descriptive statistics were obtained at the sample, item, and raw-score levels. In the second and third stages, the SSI items were first calibrated with the three types of modeling solutions discussed above (second stage). Next, score estimates were obtained for each participant, and the internal properties of the score estimates were assessed (third stage). Finally, in the fourth stage, evidence of the external validity of the estimated scores was obtained by using the SWLS scores as measures of a relevant, theoretically related variable.

As outlined in the paragraph above, the three solutions to be compared were fitted by using a random-regressors two-stage (calibration and scoring) estimation approach (McDonald, 1982). In the calibration stage, (second general stage above) the structural item parameters corresponding to the three solutions were estimated, and the model data-fit was assessed. In the scoring stage, the item parameter estimates were taken as fixed and known, and used to estimate the individual trait levels for each respondent.

A unified calibration approach was used in which the three solutions were fitted by using a limited-information factor-analytic (FA) underlying-variables approach (UVA; e.g. Muthén, 1993). For the GRM and LL-GRM this approach entails fitting the unidimensional FA model to the first-order (marginal endorsement proportions) and

second-order (polychoric correlation matrix) data. The resulting output consists of the item threshold and item loading estimates (see Ferrando & Lorenzo-Seva, 2013) that can be transformed to the corresponding GRM and LL-GRM parameter estimates. The common estimation procedure was robust unweighted least squares with mean and variance corrections of the fit statistics and standard errors (ULS-MV), as implemented in the Mplus version 8.6 program (Muthén & Muthén, 2017). Goodness of model-data fit was the same for both the GRM and the LL-GRM solutions, as they were obtained by re-parameterizing the common-FA solution. Assessment of fit was appraised using: (a) the RMSEA as a measure of relative fit; (b) the CFI as a measure of comparative fit with respect to the null independence model, and the root mean squared residual (RMSR) as a measure of absolute fit (e.g. Tanaka, 1993). We acknowledge that there are more flexible and sophisticated procedures for calibrating the LL-GRM (Lucke, 2013, 2015; Magnus & Liu, 2018a, b), but we submit that the results provided by the FA-UVA approach can be taken as essentially correct (e.g. Reise & Rodriguez, 2016).

With regards to the FMA solution, the general calibration above was also used but with the estimation criterion and measures of model-data fit typical of this type of model (e.g. Muthén & Asparouhov, 2006). The estimation criterion was robust maximum likelihood (MLR), and the relative fit when comparing one and two classes used three types of indicator: (a) the BIC parsimony information criterion, which provides a trade-off between simplicity and goodness-of-fit, (b) the normed entropy criterion, which indicates the extent to which individuals can be differentiated in terms of the class they belong to, and (c) the Lo-Mendel-Rubin (LMR) test, which assesses whether adding a second class to the single class baseline solution significantly improves model-data fit.

The Mplus scripts for fitting the GRM and the FMA models can be obtained from the authors under request. The programs for re-parametrizing the GRM and transforming it to the LL-GRM at the calibration level were written in MATLAB code by the authors.

For the three calibrated solutions, the score estimates obtained in the third stage were Bayes expected a posteriori (EAP, Bock & Mislevy, 1982). This type of scoring has two interesting properties in the present case. First, it ensures that all the individual estimates fall within reasonable values. Second, it allows us to assess the impact of the different priors (see below) on the internal and external properties of the resulting score estimates. For the GRM solution, the EAP estimates were computed using FACTOR (Lorenzo-Seva & Ferrando, 2006). For the remaining solutions, they were obtained using MATLAB programs written by the authors. The priors were (a) standard normal in the GRM solution and (b) lognormal for the LL-GRM. In each class, the two-class FMA solution assumed a normal prior with unit variance, and fixed the general mean for the whole group at zero.

The internal properties of the EAP scores derived from each solution were assessed as follows. First, the corresponding histograms were plotted, together with the corresponding Kernel estimates of the probability density function (see Ferrando, 2003). Second, the conditional accuracy of the scores at different trait levels was assessed by plotting the conditional reliability estimates as a function of the trait estimated values. It is noted that the usual measure of local accuracy is the amount of information (Lord, 1980). However, we decided to use the conditional reliability for two reasons. First, it is a unitless index, which allows the results for the different solutions to be directly compared. Second, it is a normed 0–1 index that is easy to interpret and familiar to applied researchers.

Finally, the overall (marginal) accuracy and determinacy of the score estimates was assessed by using two indicators: (a) marginal reliability (as a measure of overall accuracy) and (b) the factor determinacy index (FDI, a measure of the determinacy of the score estimates) (see Ferrando & Lorenzo-Seva, 2018). All the internal properties discussed so far were computed using MATLAB programs written by the authors.

The external properties in the fourth stage were assessed with graphical regression procedures and kernel-smoothed nonparametric regression (e.g. Härdle, 1990). The main interest here was to assess not the overall relation between the SSI estimated scores and the SWLS scores, but whether this relation was the same in different regions of the trait estimates (i.e. differential validity). More specifically, if SI behaves mainly as a unipolar trait or a quasi-trait, then the validity relation will be expected to be stronger at the upper end of the trait values (the region in which SI behaves more effectively as a dimension). This type of assessment is best carried out using the procedures discussed above. The kernel-smoothing regression, in particular, obtains an estimated regression line by taking a weighted average at different evaluation points and not imposing a priori any functional form. The specific procedure in the study was the Nadaraya-Watson kernel estimator (see e.g. Härdle, 1990).

Results

Preliminary Analyses

Descriptive statistics for the 10 items of the SSI questionnaire are shown in Table 1. More specifically, this table shows means, standard deviations and skewness for each item and for the raw sum scores. It also shows the percentage of participant responses on each response option. As can be seen, the means of the items ranged between 1.09 and 1.54,

Table 1 Means, standard deviations, skewness, kurtosis and percentage of participant responses for the SSI items

	Mean	Standard deviation	Skewness	Percentage of responses			
				1	2	3	4
Item 1	1.40	0.65	1.57	67.3	26.0	5.7	1.0
Item 2	1.45	0.72	1.63	65.6	25.8	6.3	2.3
Item 3	1.09	0.37	4.99	93.4	4.9	1.2	0.5
Item 4	1.26	0.58	2.50	80.5	14.5	3.8	1.2
Item 5	1.54	0.94	1.56	70.2	12.9	9.7	7.2
Item 6	1.33	0.56	1.66	71.8	24.1	3.7	0.4
Item 7	1.36	0.56	1.41	67.7	28.9	3.0	0.4
Item 8	1.28	0.60	2.23	78.4	15.8	4.9	0.9
Item 9	1.43	0.85	1.63	78.7	2.2	16.8	2.3
Item 10	1.31	0.55	1.67	73.0	23.3	3.4	0.3
Raw sum scores	13.45	4.67	1.78				

and standard deviations ranged between 0.37 and 0.94. All the items had skewness values above 1, which means that they are all positively skewed. The skewness value for the raw sum scores was also higher than 1. Items 3, 4 and 8 had the highest skewness values (above 2.0). As can be seen in Table 1, the first response option of each item (which involves lack of wishes, thoughts or plans related to suicide) was chosen by most participants, especially in items 3 (93.4%), 4 (80.5%), 8 (78.4%), and 9 (78.7%).

Item Calibration

As discussed above, at the structural level the GRM and the LL-GRM unidimensional solutions provide the same fit results: (a) RMSEA (90% confidence interval): 0.038 (0.025; 0.050); (b) CFI = 0.994 and (c) RMSR = 0.029. Under any of the usual recommendation guides (e.g. Schermelleh-Engel et al., 2003) these results indicate an excellent model-data fit. Furthermore, by using the approach proposed by Lee et al. (2012), power analysis gave an estimate of $\beta = 0.99$. So, there should be ample power to detect any misfit.

As for the estimated item parameters, the results agreed with those reported in the literature (Magnus & Liu, 2018a, b; Reise & Rodriguez, 2016; Reise et al., 2021, Smits et al., 2020). Given the strongly skewed item distributions reported above, the estimated thresholds were generally positive and rather high, which means that, under the GRM modeling, all the items can be regarded as very extreme (i.e. "difficult"). In fact, the item locations under the GRM were concentrated between the mean and two standard deviations above the mean in the trait scale assumed in this model. In contrast, the 'easiness' parameters in the LL-GRM were mostly located at the lower end of the trait scale assumed in this model. These results determine the different profiles of conditional accuracy that were obtained and which are discussed below.

Table 2 Initial factor loading estimates and corresponding transformed slopes

Item	Loading	Slope
i1	0.711	1.0111
i2	0.827	1.471
i3	0.686	0.9428
i4	0.885	1.9008
i5	0.698	0.9747
i6	0.933	2.5926
i7	0.852	1.6274
i8	0.914	2.2528
i9	0.891	1.9625
i10	0.655	0.8668

Table 3 Fit indices for the mixture analyses based on the one-factor model

Nº of classes	BIC	Δ	LMR test	*p*
1	8460.64	-	-	-
2	8472.23	.47	1.59	.37

Criteria: Bayesian Information Criterion (BIC), Entropy value (Δ), Lo-Mendel-Rubin (LMR) difference test with associated probability

The slope or discrimination parameter is the same under the GRM and the LL-GRM, and the estimates obtained here also agree with other reported results when the GRM is fitted to clinical items. Table 2 shows the initial factor loading estimates and the corresponding transformed slopes. These estimates are far higher than those usually obtained in normal-range personality assessment, and those for items 6 and 8 are particularly high. Note also that there is a wide range of discrimination estimates: the estimate of item 6 is three times higher than that of item 10.

Regarding the FMA results, they are summarized in Table 3.

The BIC parsimony index is slightly lower in the single class solution, and so seems to be the best option. This result shows that the LMR test is not significant, and it need not have been included. Note also that the entropy estimate is quite low, which means that it would be difficult to clearly differentiate between classes by using only internal information. So, the decision regarding the appropriateness of the single-class solution seems clear here. However, as the estimates provided by the two-class FMA solution and the further validity results were interesting, we decided to report the results derived from the two-class solution for purposes of illustration and completeness. At this internal level, the solution distinguished between one class that comprised 19% of the subjects and another that comprised the remaining 81%. In standard scaling (unit variance) and if the general mean is set to 0, the estimated means were 1.43 for class-1, and -0.335 for class-2. The higher mean of the first class suggests that it comprises those participants with higher levels of S.I., which is a relatively small group. In contrast, the second class is much larger and it comprises those participants with lower levels of S.I.

Individual Scoring: Internal Properties

Figure 1 shows the histograms of the GRM- and LL-GRM-based EAP estimates together with the kernel-smoothed non-parametric estimated densities.

The distribution of the GRM-based EAP estimates (panel a) clearly departs from the prior normal assumed in the estimation, which suggests that this prior is inappropriate for modeling the SI construct. Overall, the use of this "inappropriate" prior is expected to produce a type of 'ensemble bias' phenomenon (Mislevy, 1986) in which the distribution of the trait estimates is incorrectly pushed towards normality. Even so, the information provided by the item responses clearly outweighs the normal prior, so the posterior distribution is still positively skewed (the skewness coefficient was 0.67). In contrast, the distribution of the LL-GRM estimates in Fig. 1 (panel b) is far more extreme (the skewness coefficient here is 3.08). This is only to be expected given that, in this case both the information provided by the item responses and the lognormal prior are consistent with each other. As a result of this mutual consistency and, as expected (Magnus & Liu, 2018a, b; Reise et al., 2018, 2021) the LL-GRM distribution contracts most of the cases at the lower end and expands the remaining cases at the upper end or upper tail.

For the sake of completeness, we have also included the graphs for the two-class FMA-based score estimates (panel c). To add more information, the kernel smoothed densities were plotted separately for the participants in each class according to the posterior probabilities.

The distribution is, again, virtually the same as that obtained based on the standard GRM solution in panel a), as it should be. Note also that the smaller distribution presumably corresponding to class 1 appears to fit the upper tail of the distribution. So, Fig. 3 provides support for the fit results in Table 3: that is, there is still not enough evidence to suggest that the two distinguishable classes are present.

Regarding the conditional accuracy results, the conditional reliability estimates across the range of scores is displayed in Fig. 2 for the GRM (panel a) and the LL-GRM (panel b).

As expected, the profiles of conditional accuracy are quite different under the two models. Because items are "difficult" according to the GRM, accuracy is maximum at high trait levels (the levels that match the difficulty of the test). Thus, in accordance with the reliability profile in panel a, the GRM scores could be used to distinguish between individuals with high levels of severity, but not between those who have SI and those who do not. This result agrees with the results reported by Reise and coworkers (Reise et al., 2018, 2021).

For the LL-GRM solution in panel b, accuracy is maximum around the mean of the trait distribution. Therefore, the LL-GRM profile could be used to differentiate between those who have severe SI and those who do not. However, it would not be so useful for differentiating between degrees

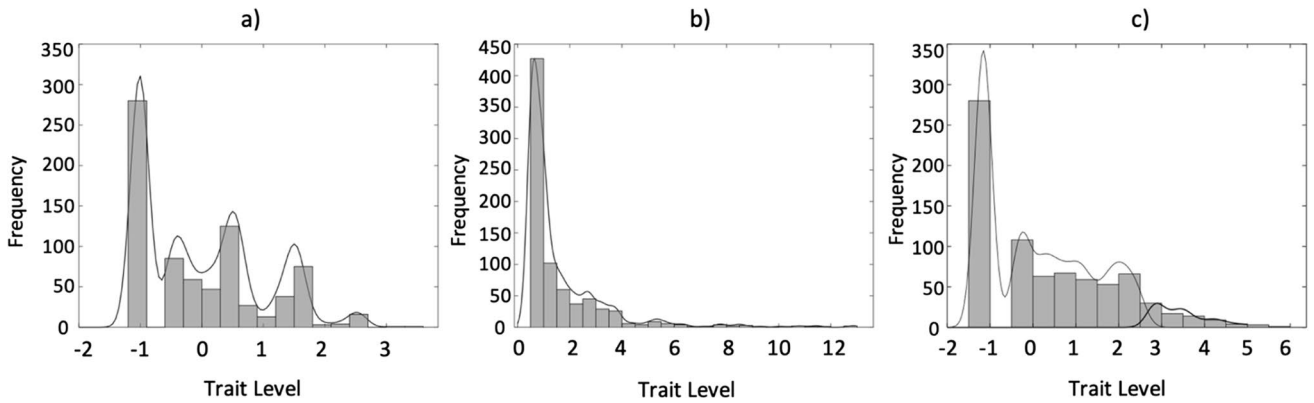

Fig. 1 Distribution of the S.I. trait estimates with kernel density curve: **a**) GRM, **b**) LL-GRM, **c**) Two-class MFA solution

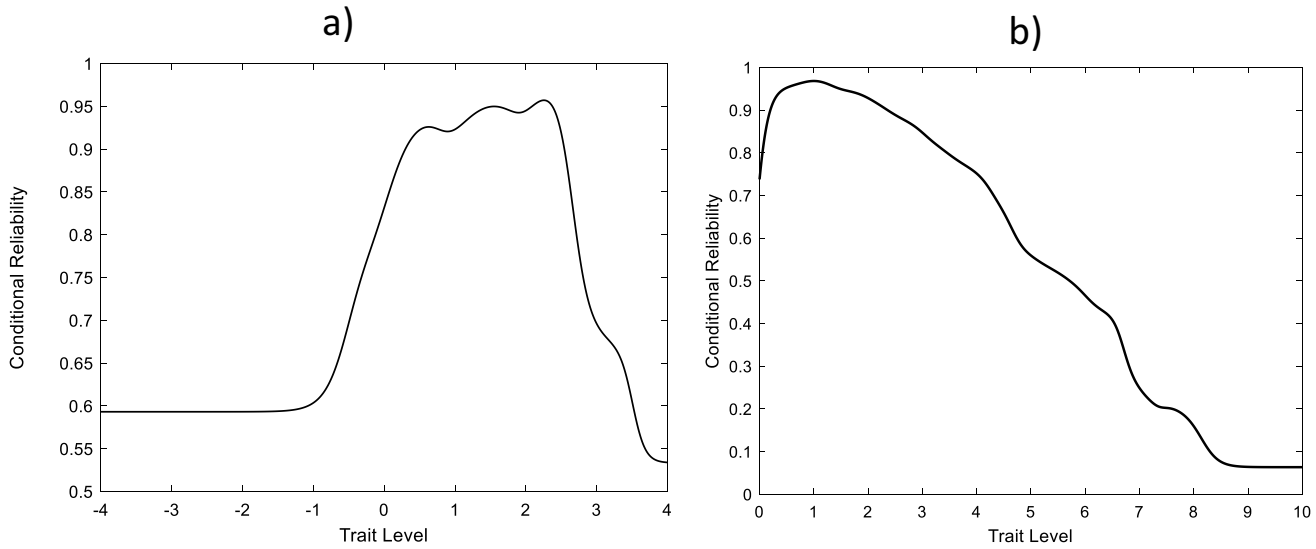

Fig. 2 Conditional reliability of the S.I. trait estimates at different trait levels: **a**) GRM, **b**) LL-GRM

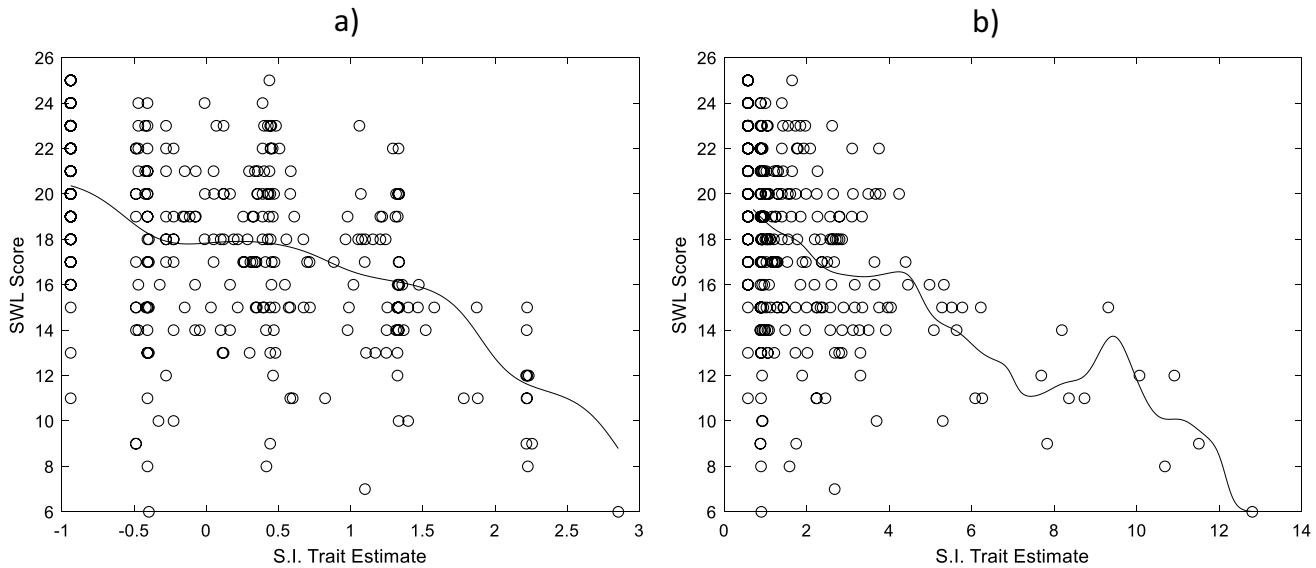

Fig. 3 Bivariate scatterplot of the SWL scores against S.I. estimates with fitted kernel line: **a**) GRM, **b**) LL-GRM

of SI severity. Again, this result matches other results in the literature with different clinical constructs (Magnus & Liu, 2018a, b; Reise et al., 2018, 2021).

Regarding the overall measures of accuracy and determinacy, the estimated marginal reliability of the GRM-based score estimates and the corresponding determinacy index were $r_{xx} = 0.78$, and $FDI = 0.88$. For the LL-GRM score estimates they were $r_{xx} = 0.90$ and $FDI = 0.95$. These values reflect the different ranges and densities of the trait distribution at which the scores provide high accuracy (see Ferrando et al., 2019). Thus, the conditional accuracy curve for the GRM is more peaked (narrower range) and more accurate at high levels, which is where subjects are fewest. In contrast, the conditional accuracy for LL-GRM is higher across a broader range at the lower level where there are more participants. Overall, although both sets of scores can attain a high degree of accuracy and determinacy, the LL-GRM scores are more accurate and determinate for most respondents.

Individual Scoring: External Validity Evidence

For the GRM (panel a) and the LL-GRM (panel b), Fig. 3 shows the bivariate scatterplot between (a) the SWLS scores on the ordinate axis, and (b) the corresponding EAP score

estimates on the abscissa axis, together with the fitted kernel regression line. Overall, it is apparent from both graphics that the regression of the SWLS scores on the SI-EAP estimates is neither linear nor homoscedastic. Thus, at the lower end of the trait estimated values, the regression curve is flatter, and the conditional variance is larger. In fact, in both panels, the scatter clearly approaches the so called "twisted-pear" contour (Fisher, 1959).

For further analytical evidence for these graphical results, a cubic polynomial line was first fitted that approached the kernel line displayed in the graphics. The fit in both cases was significantly better than the linear fit (the R^2 increases were 0.05 for GRM and 0.03 for LL-GRM). Second, the product-moment correlation was computed between the SWLS and the SI-EAP scores in the lower and upper regions of the EAP continuum using threshold values of 1.5 (GRM) and 4.5 (LL-GRM) (see panels (a) and (b) of Fig. 3). Results for the GRM were r = -0.28 (lower range) and r = -0.47 (upper range). The difference was statistically significant. The corresponding results for the LL-GRM were r = -0.24 and r = -0.54, which were also significant.

Finally, the stability and generalizability of the results above were assessed with a Bootstrap resampling analysis. As a summary, the regression curves proved to be quite stable under resampling, and there were no substantial outliers that could have given rise to spurious interpretations. This last result can be anticipated from visual inspection of the scatterplots in Fig. 3.

Substantively, the results above tend to support the hypothesis that SI behaves like a unipolar trait. Furthermore, the standard interpretation of the twisted-pear effect (Fisher, 1959) seems quite plausible here. Thus, the absence (or very low levels) of SI makes it difficult to predict SWLS: individuals with little SI will probably be more or less satisfied with their life. At the other pole, however, increasing levels of SI are more strongly associated to decreasing levels of SWLS.

In terms of differences between the two models, the predictive behavior of the GRM and the LL-GRM scores can be seen from the distributions of the trait estimates in Fig. 1, and the conditional reliability curves in Fig. 2. The changes between the flatter slope at low levels and the steeper slope at high levels are more apparent in Fig. 3, panel (a), because the GRM tends to spread the low scores more evenly, as discussed above. For the same reason, the twisted-pear effect is more apparent in panel (b), because the low SI values are more condensed at the lower end and the conditional dispersion is much more marked at this end.

For the sake of completeness, Fig. 4 also shows the scatterplot based on the two-class results. It is essentially the same as the one for the GRM but differentiated by the classes of conditional probability: class 1, circles; class 2, stars. The result makes sense, as it implies a large, undifferentiated class with almost no SI and poor predictive power, and a small class with SI and much greater predictive power. However, this type of regression plot can also be obtained in a simpler way by using a unipolar formulation (note that

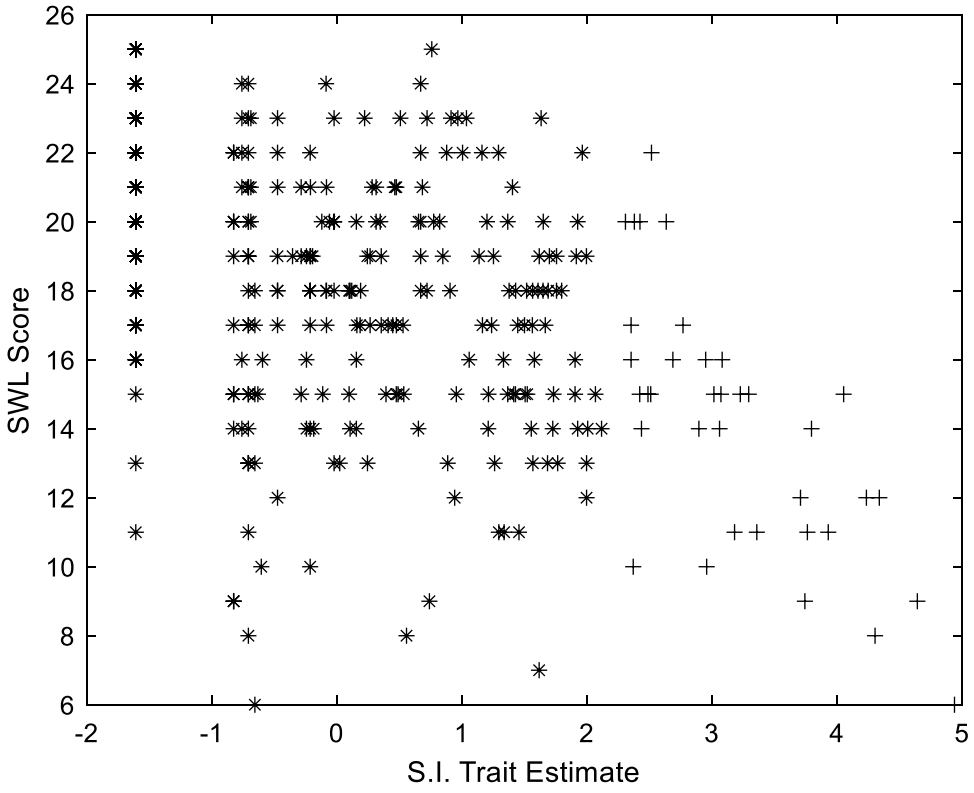

Fig. 4 Bivariate scatterplot of the SWL scores against S.I. estimates. Two-class FMA solution

class-1 corresponds to the upper-tail of the distribution in Fig. 1, panel c).

Discussion

In general, the studies carried out in community samples show that the distribution of S.I. scores tend to be highly and positively skewed (e.g., Norr et al., 2016), because most people have low levels of S.I. or do not suffer from this problem at all. However, procedures originally developed for bipolar traits are routinely used with this variable (e.g.,Cotton et al., 1995; Díez Gómez et al., 2020; Núñez et al., 2019; Sánchez-Álvarez et al., 2020), and so ignore this issue completely. According to Reise et al. (2018), while bipolar traits have meaningful variation at both poles of the trait, unipolar traits are far more meaningful at one end of the continuum than the other, and "quasi-traits" are only relevant at one end of the continuum, the other end only reflecting absence of the disorder. Therefore, S.I. seems to behave more like a unipolar trait or a "quasi-trait" than a bipolar trait, like other clinical constructs such as depression.

Given the discrepancy between the plausible behavior above and the routine psychometric treatment of the S.I. scores, the main goal of the current study was to compare the results obtained with three different psychometric models (GRM, LL-GRM and FMA) in a community sample. Furthermore, and considering that many clinical decisions are based on individual's scores, the study has focused on scoring results, which have received far less attention than structural results to date. In fact, this study observes that the main differences between GRM and LL-GRM are at the score level not the structural level. With regard to the structural level, both the GRM and the LL-GRM unidimensional solutions provided the same very good fit results, even though the GRM assumes that the latent variable follows a standard normal distribution while the LL-GRM treats the construct as unipolar. As discussed above, this result is congruent with the literature, as at the structural level both the GRM and the LL-GRM can be viewed as re-parameterizations of the same model (Reise & Rodriguez, 2016; Reise et al., 2018, 2021). What is of interest at this level, however, is that, although S.I. does not appear to behave like a bipolar trait, using a procedure such as GRM, which assumes a bipolar trait with a normal distribution, does not negatively affect the fit of the model and can lead to significant and interpretable results at the calibration level.

If S.I. behaves more like a quasi-trait than a unipolar trait, it might be possible to differentiate between classes of individuals (in our case, individuals with severe levels of S.I. and individuals without S.I. or low levels). The study by Liu et al. (2015) suggests that this variable is dimensional, not categorical, but it does not take into account that a variable may be both dimensional and categorical at the same time. For this reason, an FMA analysis was carried out, the results of which suggest that a two-class solution is less parsimonious than a single class solution. Furthermore, the addition of the second class does not significantly improve model-data fit. Despite this, the two-class solution provides plausible and interpretable results that are worth considering. More specifically, one class comprises most of the sample (81%), as expected, because most people have no suicidal ideation or very low levels. The other class comprises 19% of the sample, with a much higher mean than the other group, which suggests that the levels of S.I. are much higher, also as expected. However, the results suggest that it is not possible to differentiate clearly and unambiguously between an asymptomatic and a symptomatic class so, in structural terms, a single class unipolar model may be considered as more appropriate than a two-class quasi-trait model.

The fact that both the GRM and the LL-GRM lead to equivalent results at the calibration level does not mean that they provide equivalent results at the scoring level. In fact, although the rank order of individuals is the same in the GRM and the LL-GRM solutions, the properties of the scores are sufficiently differentiated to give rise to considerable differences in clinical practice. First, as the GRM assumes a normal distribution, the distribution of the estimated scores obtained by this procedure is less skewed than that obtained by the LL-GRM, because the GRM pushes the distribution towards normality. Despite this, the GRM-based distribution is still far from normality, which is further evidence that this model does not conform to the characteristics of the S.I. construct. Second, and more importantly, there are clear differences in the conditional and marginal accuracy of the estimated scores. Thus, the GRM scores are highly accurate only in the narrow range between 1.5 and 2 standard deviations above the mean, which means that this procedure is more accurate at high levels. In contrast, the LL-GRM scores are highly accurate in a much wider range around the mean, so providing accurate estimations for a larger number of participants. If cut-off values were to be set purely in internal terms (i.e. based on the conditional accuracy curves), the use of the GRM would require a cut-off value to be established about two deviations above the mean, and this value would serve to accurately differentiate between individuals who already have high levels of S.I. In contrast, the use of the LL-GRM would call for a cut-off around the trait mean, and would allow us to differentiate more accurately individuals who suffer from severe S.I. from individuals who do not. In some S.I. studies, cut-off points are based on the distribution of the scores in a community sample (e.g., Met al.,sky & Joiner, 1997; Reynolds, 1988). In this case, if the objective is to identify subjects who present suicidal ideation and differentiate them from those who do not, so that they can be provided with the support and

attention they need to decrease their emotional suffering and the possible risk of suicidal behaviour, the LL-GRM procedure would be more appropriate. However, many studies rely on external criteria and calculate cut-off points on the basis of the receiver operating characteristic (ROC) curve (e.g., Holi et al., 2005; Osman et al., 2001). So, further studies are needed to determine how the GRM and LL-GRM procedures affect the establishment of cut-off points when these external criteria are used.

Given the conditional accuracy results above, the overall (marginal) precision is higher for LL-GRM scores, which thus makes it possible to accurately measure a larger number of subjects in a community sample. This is an advantage for the LL-GRM, as it means that a larger number of cases can be accurately identified and more people given the help they need. Therefore, this advantage should be taken into account by researchers and screening professionals when they decide which procedure and modelling to use, because screening studies in community samples should provide accurate results for as many people as possible, so that vulnerable people suffering from suicidal ideation are not left without attention and help. With LL-GRM modelling, fewer cases of vulnerable people will remain unidentified. This is of considerable importance for such a sensitive issue as suicide, which is a health problem that costs the lives of many people around the world (World Health Organization, 2021) and also causes suffering in the people close to them. In fact, the accurate assessment of negative emotional states may help to improve the life quality of those people who need it (e.g., Aki et al., 2020; Ladakis & Chouvarda, 2021). For this reason, it would be advisable for applied researchers and test developers to use both GRM and LL-GRM, so that the manuals of S.I. questionnaires and guidelines for practitioners can give the general results, the scoring schemas, and the cut-off points they can determine, and explain the range of precision in each case. Then, professionals carrying out screening studies in community samples (for example, professionals working in educational counseling centers) can decide which option, type of scoring, and cut-off points are most suitable for their objectives. They should also consider whether they aim to differentiate between high levels of severity, or to identify subjects who suffer suicidal ideation, because GRM and LL-GRM have different implications at this level.

Finally, another aim of the study was to assess the effects of the various models on the prediction of an external variable: satisfaction with life. Differential validity effects were observed on both GRM and LL-GRM scores in different regions of the S.I. continuum, with higher predictive power at high levels of SI. In fact, the results suggest that there is a weak relationship between the two variables at low levels of SI, which means that people with little or no suicidal ideation may experience different levels of life satisfaction, whereas people with high SI tend to experience low levels of life satisfaction. This result, in turn, supports that SI behaves like a unipolar trait, and is more meaningful at one end of the continuum than at the other end. As well as the differential validity effects, however, there are some differences between GRM and LL-GRM results. In the GRM, the regression curve is flatter at lower levels and steeper at higher levels, which is congruent with the distributions of the trait estimates and the conditional reliability curves obtained, as this procedure tends to spread low scores in order to push the distribution through normality. In contrast, in the LL-GRM, heteroscedasticity is higher at lower levels of SI, which better shows the difference between individuals with high SI and individuals with low or no SI in terms of their life satisfaction. Several previous studies show a negative relationship between SI and satisfaction with life (e.g., Morales-Vives & Dueñas, 2018; You et al., 2014) but, as far as we know, none of them assessed the implications of GRM and LL-GRM modelling for the prediction of an external variable such as life satisfaction, or even for other kinds of external variable, Therefore, these results are new in this field, and they provide further evidence about the unipolar nature of suicidal ideation and its implications for external validity.

Of course, the study has some limitations. Even though the questionnaires were disseminated in a variety of environments, one of them being high schools (where there are a wide range of students from different socio-economic levels), it is still not possible to ensure that the sample is fully representative of the community. Additional information such as psychological history, suicide attempts, child abuse, etc., would have been useful to resolve this issue, but this information was not collected, as it would have made data administration prohibitively long and tiring. In spite of this potential limitation, however, the low percentage of participants with suicidal ideation fully agrees with the expected characteristic of community samples, which suggest that the sample is a community one and not a clinical one.

To sum up, the current study shows that the GRM and LL-GRM have different implications at the scoring level that should be considered by researchers and screening professionals, especially in relation to the conditional and marginal accuracy of the estimated scores. Furthermore, they also have different implications for the prediction of an external variable. As far as we know, this has not been assessed in previous studies on unipolar traits and "quasi-traits". However, further studies should be made about the implications of both modellings for establishing cut-off points based on external variables.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was supported by a grant from

the Spanish Ministry of Science and Innovation (Grant number PID2020-112894 GB-I00) and a grant from the Catalan Ministry of Universities, Research and the Information Society (Grant number 2017 SGR 97).

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval This study was performed in line with the principles of the Declaration of Helsinki. The project and the protocol of this study was approved by the Ethical Committee of the Faculty of Educational Sciences and Psychology of the Universitat Rovira i Virgili.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Conflict of interest statement The authors do not have a financial or personal relationship with a third party whose interests could be positively or negatively influenced by the article's content.

Competing interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Aki, B. D., Lamptey, E., Hembah, S. N., Oibiokpa, O. M., & Tachin, T. R. (2020). Covid-19 lockdown: Psychological implications on life quality. *Journal of Human, Earth, and Future, 1*(2), 78–86. https://doi.org/10.28991/HEF-2020-01-02-04

Atienza, F., Pons, D., Balaguer, I., & Garcia-Merita, M. (2000). Propiedades Psicométricas de la escala de satisfacción con la vida en adolescentes. *Psicothema, 1*(2), 314–319.

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: The Scale for Suicide Ideation. *Journal of Consulting and Clinical Psychology, 47*(2), 343–352. https://doi.org/10.1037/0022-006X.47.2.343

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444.

Chang, E. C., & Chang, O. D. (2016). Development of the Frequency of Suicidal Ideation Inventory: Evidence for the validity and reliability of a brief measure of suicidal ideation frequency in a college student population. *Cognitive Therapy and Research, 40*(4), 549–556. https://doi.org/10.1016/10.1007/s10608-016-9758-0

Cotton, C. R., Peters, D. K., & Range, L. M. (1995). Psychometric properties of the suicidal behaviors questionnaire. *Death Studies, 19*(4), 391–397. https://doi.org/10.1080/07481189508252740

De Beurs, D. P., de Vries, A. L., de Groot, M. H., de Keijser, J., & Kerkhof, A. J. (2014). Applying computer adaptive testing to optimize online assessment of suicidal behavior: A simulation study. *Journal of Medical Internet Research, 16*(9), e207. https://doi.org/10.2196/jmir.3511

Diener, E., Emmons, R., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Díez Gómez, A., Pérez Albéniz, A., Ortuño Sierra, J., & Fonseca Pedrero, E. (2020). SENTIA: An adolescent suicidal behavior assessment scale. *Psicothema, 32*(3), 382–389. https://doi.org/10.7334/psicothema2020.27

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., Chan, P. K., Slembarski, R., Liberzon, I., Galea, S., & Calabrese, J. R. (2012). The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Research, 199*(3), 169–173. https://doi.org/10.1016/j.psychres.2012.05.018

Ferrando, P. J. (2003). A Kernel density analysis of continuous typical-response scales. *Educational and Psychological Measurement, 63*(5), 809–824. https://doi.org/10.1177/0013164403251323

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*(5), 762–780. https://doi.org/10.1177/0013164417719308

Ferrando, P. J., & Lorenzo-Seva, U. (2019). An external validity approach for assessing essential unidimensionality in correlated-factor models. *Educational and Psychological Measurement, 79*(3), 437–461. https://doi.org/10.1177/0013164418824755

Ferrando, P. J., & Lorenzo-Seva, U. (2013). *Unrestricted item factor analysis and some relations with item response theory*. Technical Report. Department of Psychology, Universitat Rovira i Virgili. Retrieved December 15, 2021, from https://psico.fcep.urv.cat/utilitats/factor/documentation/technicalreport.pdf

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

Fitzpatrick, S. S., Morgan-López, A. A., Saraiya, T. C., Back, S. E., Killeen, T. K., Norman, S. B., López-Castro, T., Ruglass, L. M., Saavedra, L. M., & Hien, D. A. (in press). Graded response item response theory in scaling suicidal thoughts and behaviors among trauma-exposed women with substance use disorders. *Psychology of Addictive Behaviors*. https://doi.org/10.1037/adb0000757

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143*(2), 187–232. https://doi.org/10.1037/bul0000084

Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology, 40*(6), 374.

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

Holi, M. M., Pelkonen, M., Karlsson, L., Kiviruusu, O., Ruuttu, T., Heilä, H., Tuisku, V., & Marttunen, M. (2005). Psychometric properties and clinical utility of the Scale for Suicidal Ideation (SSI) in adolescents. *BMC Psychiatry, 5*(8), 1–8. https://doi.org/10.1186/1471-244X-5-8

Jobes, D. A., & Joiner, T. E. (2019). Reflections on suicidal ideation [Editorial]. *Crisis: The Journal of Crisis Intervention and Suicide Prevention, 40*(4), 227–230. https://doi.org/10.1027/0227-5910/a000615

Joiner, T. E., Jr., Pfaff, J. J., & Acres, J. G. (2002). A brief screening tool for suicidal symptoms in adolescents and young adults in general health settings: Reliability and validity data from the Australian National General Practice Youth Suicide Prevention Project. *Behaviour Research and Therapy, 40*(4), 471–481. https://doi.org/10.1016/s0005-7967(01)00017-1

Ladakis, I., & Chouvarda, I. (2021). Overview of biosignal analysis methods for the assessment of stress. *Emerging Science Journal, 5*(2), 233–244. https://doi.org/10.28991/esj-2021-01267

Lee, T., Cai, L., & MacCallum, R. (2012). Power analysis for tests of structural equation models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 181–194). Guilford Press.

Liu, R. T., Jones, R. N., & Spirito, A. (2015). Is adolescent suicidal ideation continuous or categorical? A taxometric analysis. *Journal of Abnormal Child Psychology, 43*(8), 1459–1466. https://doi.org/10.1007/s10802-015-0022-y

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

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88–91. https://doi.org/10.3758/BF03192753

Lubke, G. H., & Miller, P. J. (2015). Does nature have joints worth carving? A discussion of taxometrics, model-based clustering and latent variable mixture modeling. *Psychological Medicine, 45*(4), 705–715. https://doi.org/10.1017/S003329171400169X

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21

Lucke, J. F. (2013). 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

Magnus, B. E., & Liu, Y. (2018a). 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

Magnus, B. E., & Liu, Y. (2018b). A Zero-Inflated Box-Cox normal unipolar item response model for measuring constructs of psychopathology. *Applied Psychological Measurement, 42*, 571–589. https://doi.org/10.1177/0146621618758291

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*(4), 379–396.

Metalsky, G. I., & Joiner, T. E., Jr. (1997). The hopelessness depression symptom questionnaire. *Cognitive Therapy and Research, 21*(3), 359–384. https://doi.org/10.1023/A:1021882717784

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359–381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177–195.

Morales-Vives, F., & Dueñas, J. M. (2018). Predicting suicidal ideation in adolescent boys and girls: The role of psychological maturity, personality traits, depression and life satisfaction. *Spanish Journal of Psychology, 21*(e10), 1–12. https://doi.org/10.1017/sjp.2018.12

Murray, A. L., Eisner, M., & Ribeaud, D. (2017). Can the Social Behavior Questionnaire help meet the need for dimensional, transdiagnostic measures of childhood and adolescent psychopathology? *European Journal of Psychological Assessment, 35*(5), 674–679. https://doi.org/10.1027/1015-5759/a000442

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Sage Publications.

Muthén, B., & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31*(6), 1050–1066. https://doi.org/10.1016/j.addbeh.2006.03.026

Muthén, L. K., & Muthén, B. (2017). *Mplus: Statistical analysis with latent variables: User's Guide (Version 8)*. Muthén & Muthén.

Nock, M. K., Borges, G., Bromet, E. J., Cha, C. B., Kessler, R. C., & Lee, S. (2008). Suicide and suicidal behavior. *Epidemiological Review, 30*, 133–154. https://doi.org/10.1093/epirev/mxn002

Norr, A. M., Allan, N. P., Macatee, R. J., Capron, D. W., & Schmidt, N. B. (2016). The role of anxiety sensitivity cognitive concerns in suicidal ideation: A test of the Depression-Distress Amplification Model in clinical outpatients. *Psychiatry Research, 238*, 74–80. https://doi.org/10.1016/j.psychres.2016.02.016

Nugent, W. R. (2005). The development and psychometric study of an ultra-short-form suicidal ideation measure. *Best Practices in Mental Health, 1*(2), 1–18.

Nugent, W. R. (2006). A psychometric study of the MPSI suicidal thoughts subscale. *Stress, Trauma, and Crisis, 9*(1), 1–15. https://doi.org/10.1080/15434610500506209

Núñez, D., Arias, V., Méndez-Bustos, P., & Fresno, A. (2019). Is a brief self-report version of the Columbia severity scale useful for screening suicidal ideation in Chilean adolescents? *Comprehensive Psychiatry, 88*, 39–48. https://doi.org/10.1016/j.comppsych.2018.11.002

Osman, A., Bagge, C. L., Gutierrez, P. M., Konick, L. C., Kopper, B. A., & Barrios, F. X. (2001). The Suicidal Behaviors Questionnaire-Revised (SBQ-R): Validation with clinical and nonclinical samples. *Assessment, 8*(4), 443–454. https://doi.org/10.1177/107319110100800409

Otani, M., Hiraide, M., Horie, T., Mitsui, T., Yoshida, T., Takamiya, S., Sakuta, R., Usami, M., Komaki, G., & Yoshiuchi, K. (2021). Psychometric properties of the Eating Disorder Examination-Questionnaire and psychopathology in Japanese patients with eating disorders. *International Journal of Eating Disorders, 54*(2), 203–211. https://doi.org/10.1002/eat.23452

Pinto, A., Whisman, M. A., & McCoy, K. J. (1997). Suicidal ideation in adolescents: Psychometric properties of the suicidal ideation questionnaire in a clinical sample. *Psychological Assessment, 9*(1), 63. https://doi.org/10.1037/1040-3590.9.1.63

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., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*(1), 45–58.

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., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to addressing non-normal distributions in the application of IRT models to personality measures. *Journal of Personality Assessment, 100*(4), 363–374. https://doi.org/10.1080/00223891.2017.1381969

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

Reynolds, W. M. (1988). *Suicidal ideation questionnaire (SIQ): Professional manual*. Psychological Assessment Resources.

Reynolds, W. M., & Mazza, J. J. (1999). Assessment of suicidal ideation in inner-city children and young adolescents: Reliability and validity of the Suicidal Ideation Questionnaire-JR. *School Psychology Review, 28*(1), 17–30. https://doi.org/10.1080/02796015.1999.12085945

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Psychometric Society. Retrieved December 15, 2021, from http://www.psychometrika.org/journal/online/MN17.pdf

Sánchez-Álvarez, N., Extremera, N., Rey, 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

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8*(2), 23–74.

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.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Sage.

Villardón, L. (1993). El pensamiento de suicidio en la adolescencia [The thought of suicide in adolescence]. Editorial Rontegui.

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement, 39*, 583–597. https://doi.org/10.1177/0146621615588184

World Health Organization. (2021). *Suicide worldwide in 2019: Global Health Estimates*. Geneva. Retrieved December 15, 2021, from https://www.who.int/publications-detail-redirect/9789240026643

You, Z., Song, J., Wu, C., Qin, P., & Zhou, Z. (2014). Effects of life satisfaction and psychache on risk for suicidal behaviour: A cross-sectional study based on data from Chinese undergraduates. *British Medical Journal Open, 4*(3), e004096. https://doi.org/10.1136/bmjopen-2013-004096

Zhang, Y., Yip, P. S. F., & Fu, K. W. (2014). Validation of the Chinese version of the Reynolds' suicidal ideation questionnaire: Psychometric properties and its short version. *Health and Quality of Life Outcomes, 12*(1), 1–9. https://doi.org/10.1186/1477-7525-12-33