UNIPOL-GC

An R package for fitting unipolar log-logistic IRT models to graded and continuous item responses

USER'S GUIDE

Prepared by:

Ana Hernández-Dorado

Pere J. Ferrando

Fabia Morales-Vives

Please reference this document as:

Hernandez-Dorado, A., Ferrando, P.J., & Morales-Vives, F. (2024). *UNIPOL-GC User's Guide. Technical Report*. Universitat Rovira i Virgili. Tarragona.

Available at: https://www.psicologia.urv.cat/en/tools/unipol-r-code/

Contents

-1. Theoretical Bases and features

The Graded Response model (GRM; Samejima, 1969) and the continuous response model (CRM; Bejar, 1977, Samejima, 1973, Wang & Zeng, 1998), are unidimensional Item Response Theory (IRT) models in common use in personality and attitude measurement. These models, however, were not initially designed for measuring non-cognitive traits. Rather, they were intended for abilities and cognitive traits, and they were directly transported into the non-cognitive domains in the hope that they would also be suitable (e.g. Reise & Moore, 2012). In our view, in most personality/attitude applications they are. In other cases, however, their appropriateness is more questionable.

When a non-cognitive measure is fitted by using the GRM or the CRM, the trait that is assessed is, implicitly or explicitly, assumed to be: (a) equally meaningful at both ends of the continuum, and (b) normally distributed in the population of interest (see e.g. Morales-Vives et al., 2023). For normal-range and clearly bipolar traits, such as extraversion-introversion, these assumptions seem quite reasonable. For other type of traits such as addictive behaviors, psychiatric disorders, clinical symptoms or maladaptive traits, however, it is more plausible to assume that the trait (a) takes only positive values (Lucke, 2013), (b) is more meaningful at the upper end of the dimension, and (c) has a rightly skewed distribution in the population of interest. A typical example of these assumed features is a psychiatric symptoms checklist administered in a community sample (see e.g. Morales-Vives et al., 2023). As for the first two assumptions, the low trait end is likely to reflect merely the absence of symptoms. So, a trait scale that only adopts positive values seems more natural here. Furthermore, if the low trait end only reflects absence of manifestations while the upper end reflects different degrees of severity, the trait is, clearly, more meaningful at its upper end. With regards to the distribution (point c), finally, most individuals will be expected not to suffer from clinical disorders or have very low levels, and they will be grouped at the lower end of the trait continuum. On the other hand, the far fewer individuals who do have disorders to a nonnegligible extent, will extend over the upper tail with a higher degree of heterogeneity. If this is so, the latent trait distribution in our example will have, intrinsically, a low mean and high variance both leading to a pronounced right skewness (e.g. Magnus & Liu, 2018; Morales-Vives et al., 2023).

One way to accommodate the alternative trait assumptions discussed above is to use unipolar (or positive trait) IRT models, which were initially proposed for binary-response items (Lucke, 2013, 2015). The UNIPOL-GC program implements log-logistic extensions of this type of models that are intended for graded and (approximately) continuous response items. They are: (a) the log-logistic graded response model (LL-GRM; Reise et al. 2021) and (b) the log-logistic continuous response model (LL-CRM; Ferrando et al., 2023). Essentially, these models can be viewed as transformations of the standard GRM and CRM in which the trait scale is changed so as to take only positive values and have a rightly skewed distribution in the population of interest. More specifically, in the LL-GRM and the LL-CRM formulations, the trait is assumed to have a lognormal distribution. This change of the trait scale, however, has profound consequences on how the models behave in comparison to their standard counterparts (Lord, 1975, Yen, 1986).

-1.1 Description of the LL-GRM and the LL-CRM: main features

In this section we shall only provide a basic, mostly conceptual summary of the models implemented in the program. More technical presentations, to which the interested reader is referred, are provided in Reise et al. (2021) for the LL-GRM, and Ferrando et al. (2023) for the LL-CRM.

The most common approach for fitting IRT models is two-stage: Calibration and Scoring (e.g. Mislevy & Bock, 1990). In the calibration stage, the parameters of the items are estimated and model-data fit at the structural level is assessed. In the scoring stage, provided that model-data fit is judged to be acceptable, the item parameter estimates are taken as fixed and now, and used to obtain individual score estimates and accompanying measures of (conditional) score accuracy. We shall use this schema to describe the main features of the two models implemented in UNIPOL-GC.

Consider first that the items have *m* ordered categories scored with successive integers: 1,2…*m*. These categories are assumed to be separated by *m*-1 thresholds *k* (*k*=1…*m*-1). Now, according to the LL-GRM, the probability of responding above the *k* threshold in item *j* for a fixed trait level θ_U, is: (the *U* stands for unipolar):

$$P^*(X_{jk}|\theta_U) = \frac{\epsilon_{jk}{\theta_U}^{\alpha_j}}{1+\epsilon_{jk}{\theta_U}^{\alpha_j}}. \qquad (1)$$

As a function of level θ_U, equation (1) describes the Threshold Response Function (TRF) for this item and threshold category. The trait θ_U is assumed to follow a lognormal distribution with parameters μ_U=0 and σ_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.

The ε_{jk} and α_j item parameters are both restricted to have positive values, and are "easiness" and slope/discrimination parameters, respectively. The ε_{jk} values are related to the marginal proportions of endorsement, so that, a high value means that a large proportion of people responds above the corresponding *k* threshold or category boundary. As for the slopes, at low α_j values the expected TRC is flatter at almost all trait levels, and, the higher the α_j value becomes, the more the TRC increases at low trait levels.

A useful auxiliary item location parameter can be defined from (1) as:

$$\delta_{jk} = \left(\frac{1}{\epsilon_{jk}}\right)^{\frac{1}{\alpha_j}}. \tag{2}$$

The δ_{jk} location parameter, which again is always positive, can be interpreted as follows: For each category threshold, δ_{jk} is the trait value at which the probability of responding above this threshold is 0.50. Overall, the interpretation of the item parameters so far described will (hopefully) become clearer in the illustrative example below.

We turn now to the LL-CRM. Suppose that the item scores can be treated as (approximately) continuous. For practical and interpretative purposes, we shall scale the item scores to have values between 0 and 1. With this scaling, for fixed θ_U, the expected score in item *j* is given by:

$$E(X_j|\theta_U) = \frac{\epsilon_j {\theta_U}^{\alpha_j}}{1+\epsilon_j {\theta_U}^{\alpha_j}}. \tag{3}$$

As a function of θ_U, the conditional expectation (3) is now the Item Response Function (IRF) of the LL-C model. Overall, the interpretation of the parameters in (3) is essentially the same as in the LL-GRM case. Again, θ_U , is assumed to follow a (0,1) lognormal distribution, and the ε_j and α_j item parameters (again both restricted to be positive) retain their interpretation as "easiness" and slope/discrimination parameters respectively. In this case, however, ε_j can be viewed as a general (not threshold-related) "easiness" parameter. So, other things equal, the higher ε_j is, the higher the expected score for item *j* becomes. Finally, an item location parameter δ_j can be also defined as

$$\delta_j = \left(\frac{1}{\epsilon_j}\right)^{\frac{1}{\alpha_j}}. \tag{4}$$

And is interpreted as the θ_U trait level at which the expected item score is 0.5. So, this is the trait level that corresponds to the midpoint of the response scale, which, conceptually, is the scale value that marks the transition from a tendency to disagree with the item to a tendency to agree with it (see Ferrando, 2009).

We shall now provide a graphical illustration of the type of item functions implied by the LL models. More specifically, because the LL-CRM is the simplest of the two models, we shall provide the LL-CRM-based IRFs corresponding to two hypothetical items (the TRFs have essentially the same shape, but, within each item, one curve would be needed for each category threshold).

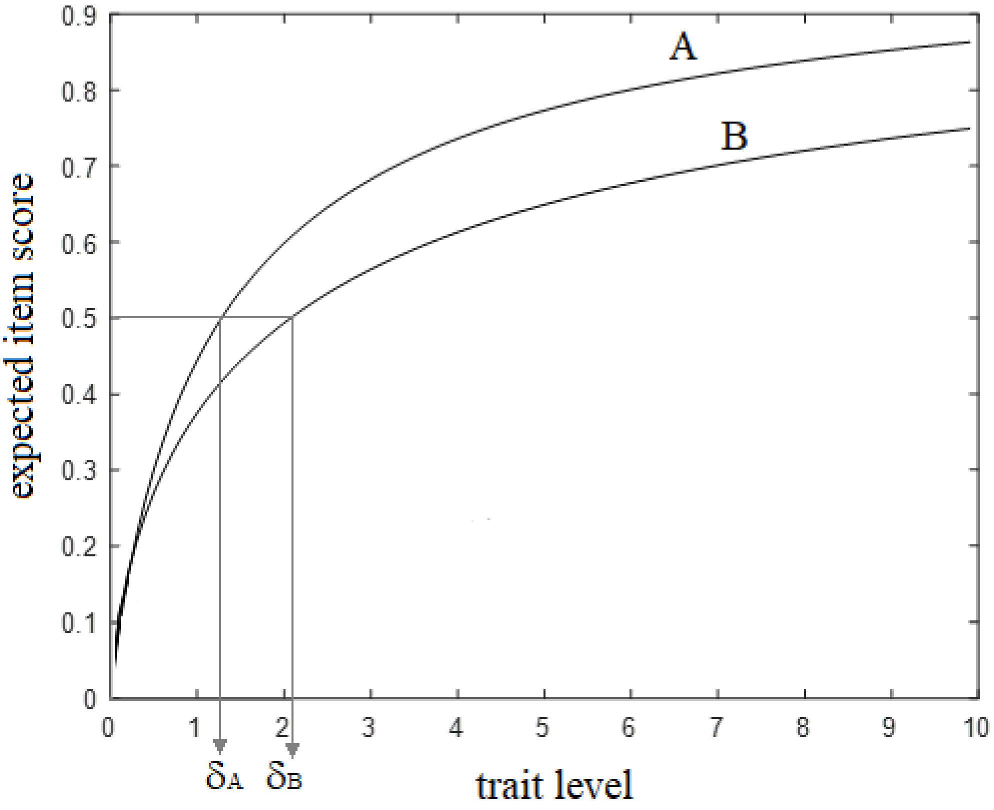


Figure 1. Item response functions (curves) in the LL-CRM

To start with, the IRFs in figure 1 clearly depart from the usual CRM (or GRM) ogives. Here, each curve is a power function (e.g. Stevens, 1975), and its general trend is that the curve is concave downward, and its slope tends to increase more strongly for trait values close to zero and flattens as θ_U increases. Conceptually, this trend implies that the item score becomes progressively less sensitive to the trait level as this level increases.

Turning now to comparisons, the general difference between both items is that item A is easier and more discriminating than B. First, with regards to easiness, note that δ_A is about 1, whereas δ_B is about 2, which means that a higher trait level is required in B to attain an expected score equal to the scale midpoint. Also, the area under the curve corresponding to A (which is proportional to the easiness value ε_A) is larger than the area under the curve of B. As far as discrimination is concerned, note that, in the IRF of item A, the slope increases more abruptly at very low, near zero θ_U values. This result implies that the α_A value is higher than the α_B value.

By making appropriate (exponential) transformations of the person and the item easiness parameters, the TRFs and IRFs in equations (1) and (3) can be transformed to the corresponding functions of the standard GRM and CRM respectively. Details are provided in Reise et al. (2021, p.12) and Ferrando et al. (2023, equations 6 to 8). Here we shall only discuss the practical and substantive implications of these results. First, the result that the standard GRM and CRM can be obtained as re-parameterizations of the LL-GRM and the LL-CRM provides a basis for considering the latter as transformations of the former in which the scale of the trait is changed. Second, the result also shows that, at the calibration stage, both type of models (standard and LL) are expected to reach the same degree of model-data fit when calibrated in the same dataset. Thus, we have a case of alternative models that fit the data equally well but that are based on different principles and philosophies (see e.g. Samejima, 1996). Finally, because the GRM and CRM can be parameterized as factor-analytic models, the result implies that so are the LL-GRM and the LL-CRM. This third implication is highly relevant in practice, as it means that well-established procedures can be used to calibrate the new models considered here.

We shall now discuss the results concerned with the scoring stage. For both, the LL-GRM and the LL-CRM the outcomes of this stage are: (a) the individual trait point estimates (i.e. individual scores), and (b) the conditional measures of accuracy corresponding to each individual estimate. Because reliability is a unitless and bounded measure that applied researchers are familiar with, in our proposal we have chosen to use reliability as a measure of score accuracy. So, for each respondent, in addition to the individual score estimate, we shall provide the corresponding conditional reliability estimate.

As discussed above, the score (point) estimates obtained from the LL models, are, essentially, non-linear (exponential) transformations of the score estimates that would have been obtained based on the standard counterparts of these models. So, if an IRT model is fitted with the sole purpose of ranking the individuals according to their trait levels, then it is not really worth using the models proposed here, because the rank order would be the same as if the conventional model had been used.

In terms of conditional reliabilities, however, the LL models considered here and their standard counterparts will generally function in diametrically opposite ways. Under the conventional models (GRM or CRM) the items would be modelled as "difficult" (recall that the proportions of item endorsements when these models are appropriate are very low). So, the conditional reliability of the scores will be maximal at high trait levels. Conceptually, these results mean that the test will be considered to be highly appropriate for differentiating between the (few) individuals who have very high levels but will not be sensitive enough to differentiate between individuals at lower levels. On the other hand, under the models considered here (LL-GRM or LL-CRM) the conditional reliability curve as a function of the trait level will be decreasing, reaching its highest values at very low trait levels. This result implies that, according to these models, the test will accurately differentiate between those individuals who have no, or virtually no trait manifestations and those who clearly do have them. However, it will not be sensitive enough to make finer differentiations between those with high levels.

The opposite predictions made by the standard and LL models (in this case the GRM and the LL-GRM) are illustrated in figure 2, the functions having been obtained in the same dataset. (Note also the different trait scaling in the abscissae axis). Assuming that the LL-GRM was the more correct model for this data, if the standard GRM was fitted instead to this data, it would then provide a very wrong picture of how the score accuracy functions across trait levels.

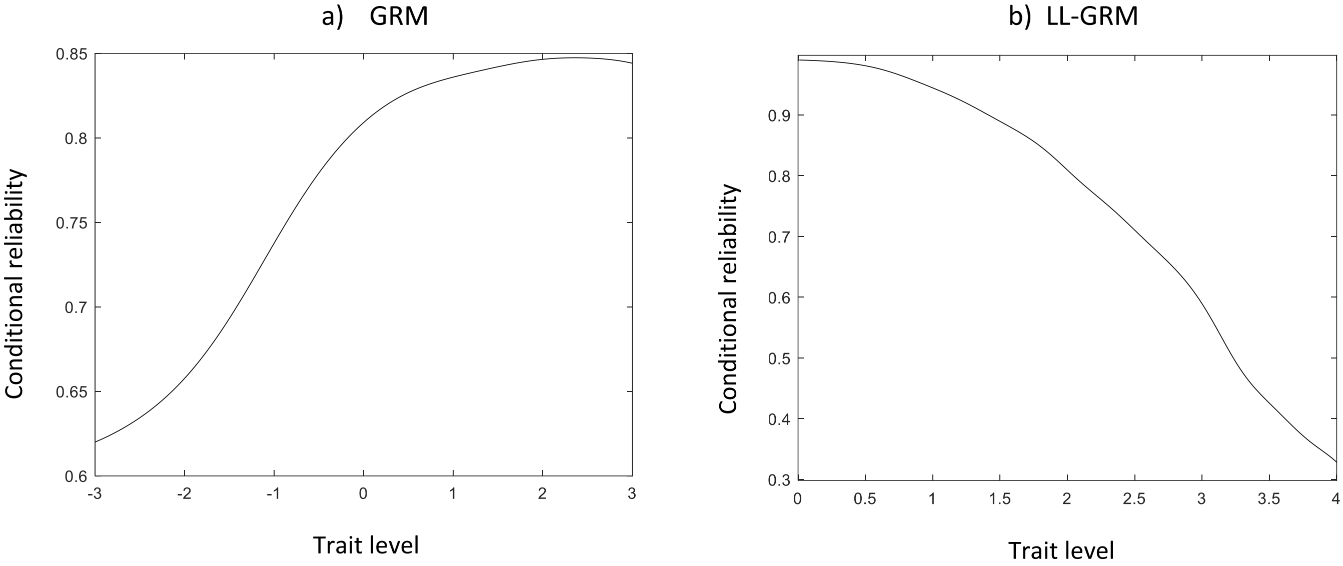


Figure 2. Conditional reliabilities as a function of trait level in the GRM and LLGRM

-1.2 Calibration and scoring procedures used in UNIPOL-GC

As advanced above, by using appropriate "linearizing" transformations, both the LL-GRM and the LL-CRM can be formulated as factor-analytic (FA) models and fitted using a simple and robust two-stage limited-information procedure. The calibration stage in UNIPOL-GC is based on this approach. So, in the LL-GRM case, the used transformation is a nonlinear step function governed by thresholds and based on the 'probit' function. This mechanism implies that the item scores are fitted by using the underlying-variables-approach (UVA) FA for ordered-categorical variables (e.g. Muthén, 1993). In the LL-CRM case, the 0-1 scaled item scores are logit transformed (see Ferrando et al. 2023, equation 6) and the transformed item scores are fitted using the linear FA model for continuous variables. Further details on estimation procedures and goodness-of-fit assessment are provided below. This first stage of calibration is carried out using the cfa function from the lavaan package, which requires the specification of the model and the selection of an estimation method. Once the initial FA estimates have been obtained, UNIPOL-GC carries out the reparameterization, and transforms them to the TRF or IRF item parameter estimates in equations (1) to (4) (i.e. ε's , α's and δ's). Model-data fit at the structural level is also assessed at this stage.

In the scoring stage, the item parameter estimates above are taken as fixed and known (see Mislevy & Bock, 1990) and individual point estimated scores and accompanying reliability estimates are obtained. At this stage, we wanted procedures that were robust and able to provide finite and plausible estimates for all the respondents in the sample under analysis, which led us to choose Bayes expected a posteriori (EAP, Bock & Mislevy, 1982) score estimation for both LL-GRM and LL-CRM. In both cases, and as discussed above, the prior distribution for θ_U was lognormal (0,1).

In the LL-CRM (but only in this model) maximum likelihood (ML) score estimates for each respondent can be obtained in closed form (Ferrando et al. 2023, equation 9). This result means that, for this model, finite ML score estimates can be obtained for all the respondents. For this reason, ML scores (see Lord, 1986) and accompanying conditional reliability estimates are also available at the user's request for the LL-CRM.

-1.2 Substantive and practical considerations for using the LL models

As discussed above, the two models implemented in UNIPOL-GC are expected to attain the same degree of model-data fit at the calibration stage than their standard counterparts. In this respect, we concur with Samejima (1996) in that an acceptable degree of fit is indeed a necessary condition for considering a model as appropriate, but is not sufficient. So, in the present case, if the fit results are acceptable, both the standard and the unipolar versions of the model fulfill the initial necessary conditions, and the choice of one or the other will have to be guided mostly by substantive and interpretative reasons

In principle, the LL-GRM or the LL-CRM can be considered as 'a priori' appropriate choices when the trait under study can be convincingly conceived as theoretically unipolar. This means that the low end of the dimension is more naturally interpretable as lack of trait manifestations rather than as a "symmetrical" opposite pole of the upper end. In terms of scalability, this trait conception implies that individuals are expected to be more scalable at the upper trait end (e.g. different degrees of severity) than at the lower end (absence of trait manifestations and undifferentiation).

Empirically, the 'a priori' bases above, imply that the distribution of the item scores in the scale under study are expected to be markedly and consistently asymmetrical (rightly skewed). Again, this is a necessary but no sufficient condition, because items that measure a bipolar trait can also have this type of distribution if they are chosen to be sufficiently "difficult". Ferrando et al. (2023) and Reise et al. (2021) however, noted that, if the trait under study is better conceived as unipolar, then, finding "easy" items that tap the lower trait end is far more difficult than finding "difficult" items that better measure at the upper end in which individuals with high levels are more differentiable. Both, Ferrando et al. (2023) and Reise et al. (2021) provide further empirical checks for supporting the appropriateness of the models implemented here. These proposals are outside the objectives of this manual, but their reading is recommended for potential users interested in unipolar models.

Finally, the calibration and scoring procedures implemented in UNIPOL-GC have been chosen to be simple and robust. So, no estimation problems are expected to appear even if small samples were used. However, given the nature of the traits modeled with the LL-GRM or the LL-CRM, in which the majority of the sample is expected to remain very undifferentiated at the lower trait end, the choice of large samples is highly recommended if these models are to be used.

-2. Installation and setup

The package is available through website (https://www.psicologia.urv.cat/en/tools/lilac-r-code/), and can be downloaded directly form the website and installed manually. To do so, extract all the files from the downloaded zip and save them in a folder. Next, install the lavaan package (Rosseel, 2012), and then open the llgm.R file and execute the function.

-3. Program Usage I: Entering and importing data

The function usage is the following:

```
> llgm(data, type, model, estimator, nodes)
```

Where the description of the input arguments are:

Data	a data matrix of raw item scores
Type	defines the type of data that has been entered. Item scores can be treated as "continuous" or "graded". If scores are defined as continuous data, they should be scaled from 0 to 1. If they are not, the program will rescale them to the 0-1 format.

	For continuous data, the user can choose the scoring method they prefer: Maximum Likelihood "ML" or Bayesian Expected A Posteriori "EAP."
model	model description using lavaan syntax. https://cran.r-project.org/web/packages/lavaan/lavaan.pdf (p.55 – lavOptions – Estimator)
estimator	All lavaan estimators are available for calibration: ML, GLS, WLS, DWLS, ULS, DLS. PML as well as their robust variants.
nodes	Matrix of 20 quadrature points and nodes where the first column represents points on a lognormal 0,1 distribution and the second column represents the probabilities associated with each quadrature point on the lognormal trait scale. This matrix is already provided.

-3. Program Usage II: Output

The output will be printed in the console just as shown in the empirical example below. All the data is stored in a list where the user can find the indices with the following arguments:

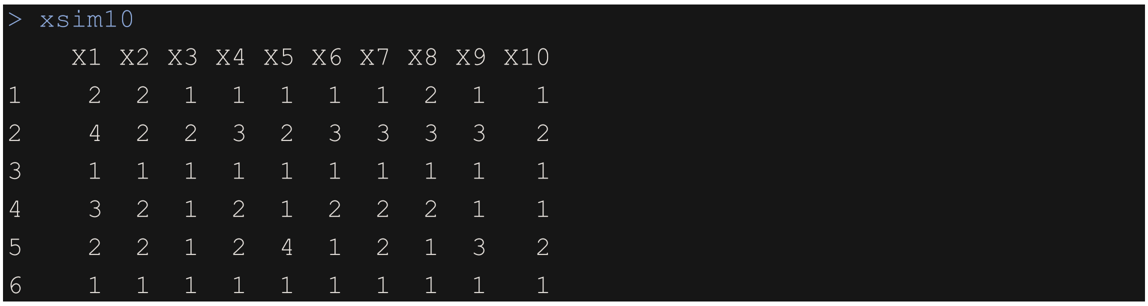
Parameters	Contains the item parameter estimates of the LL-CRM or LL-GRM model. For continuous responses, they are: intercept/easiness, slope/discrimination, standard error of the slope and difficulty (delta) index. For graded data, for each item they are the

	slope/discrimination, the threshold easiness estimates and difficulty (delta) index.
GOF	All available fit indices in the lavaan package are computed, allowing the user to decide which ones to choose. Kline (2015) suggests that, at a minimum, the reported indices should include the chi-square, the RMSEA, the CFI, and the SRMR.
Scoring	For continuous data, the matrix structure includes: • Score point estimates corresponding to each respondent. • Amount of Information or PSD: The conditional amount of information provided by the scores (ML), or the Posterior Standard Deviations (EAP), depending on the type of scoring chosen by the user. • Conditional Reliability across the estimated trait levels. For graded data, the matrix structure includes: • Score point estimates corresponding to each respondent. • Confidence intervals around each point estimate. • Posterior standard deviations (PSD) :
Reliability	Conditional Reliability across the estimated trait levels

-5. Illustrative example

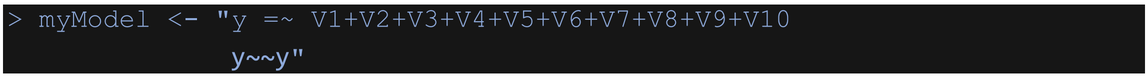
This example illustrates how to use UNIPOL-GC . The database used is located in the file named **xsim10.dat**. The data consists of 10 graded response items with a 1 to 4 response format (which means that there are 3 thresholds per item). It is a simulated dataset generated from the item parameters taken from a clinical scale. So, the LL-GRM is expected to fit rather well the data.

Once you confirm that you are in the working directory containing the procedure codes, import the files **xsim10.dat** and **nodos20log.dat**, install the *lavaan* package (Rossell, 2012), and load the main function *llgm.*



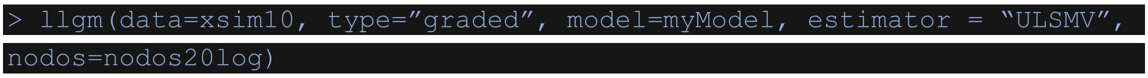
```
> xsim10
    X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
1    2  2  1  1  1  1  1  2  1   1
2    4  2  2  3  2  3  3  3  3   2
3    1  1  1  1  1  1  1  1  1   1
4    3  2  1  2  1  2  2  2  1   1
5    2  2  1  2  4  1  2  1  3   2
6    1  1  1  1  1  1  1  1  1   1
```

The syntax used for model specification is similar to that used in the Lavaan package. In this example:



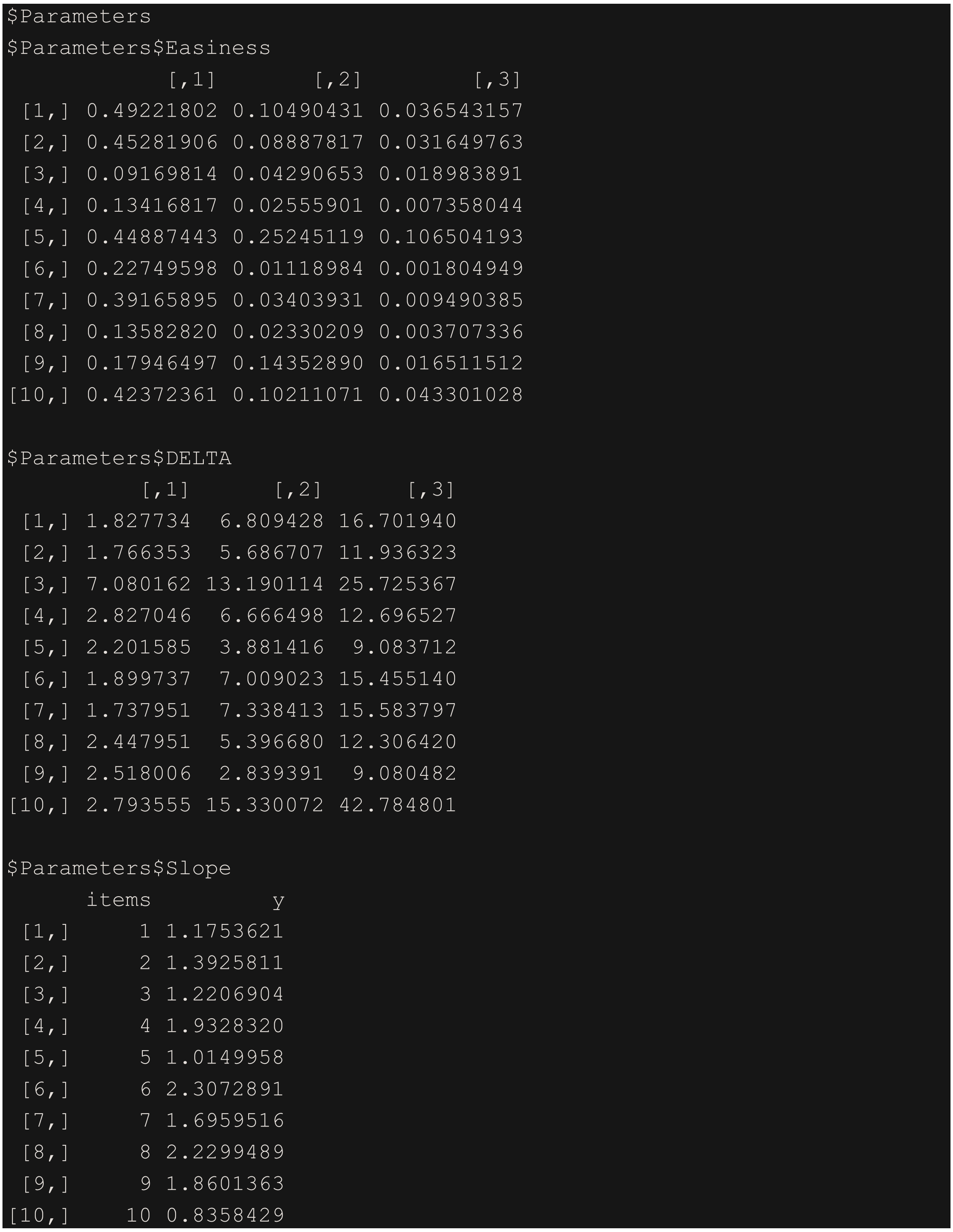
```
> myModel <- "y =~ V1+V2+V3+V4+V5+V6+V7+V8+V9+V10
             y~~y"
```

The function usage requires specifying the database to work with, the type (whether items are treated as continuous or graded response), the desired estimator, and the node matrix.



```
> llgm(data=xsim10, type=”graded”, model=myModel, estimator = “ULSMV”,
nodos=nodos20log)
```

1. Calibration: Item parameter estimates



```
$Parameters
$Parameters$Easiness
            [,1]       [,2]        [,3]
 [1,] 0.49221802 0.10490431 0.036543157
 [2,] 0.45281906 0.08887817 0.031649763
 [3,] 0.09169814 0.04290653 0.018983891
 [4,] 0.13416817 0.02555901 0.007358044
 [5,] 0.44887443 0.25245119 0.106504193
 [6,] 0.22749598 0.01118984 0.001804949
 [7,] 0.39165895 0.03403931 0.009490385
 [8,] 0.13582820 0.02330209 0.003707336
 [9,] 0.17946497 0.14352890 0.016511512
[10,] 0.42372361 0.10211071 0.043301028

$Parameters$DELTA
          [,1]      [,2]      [,3]
 [1,] 1.827734  6.809428 16.701940
 [2,] 1.766353  5.686707 11.936323
 [3,] 7.080162 13.190114 25.725367
 [4,] 2.827046  6.666498 12.696527
 [5,] 2.201585  3.881416  9.083712
 [6,] 1.899737  7.009023 15.455140
 [7,] 1.737951  7.338413 15.583797
 [8,] 2.447951  5.396680 12.306420
 [9,] 2.518006  2.839391  9.080482
[10,] 2.793555 15.330072 42.784801

$Parameters$Slope
      items         y
 [1,]     1 1.1753621
 [2,]     2 1.3925811
 [3,]     3 1.2206904
 [4,]     4 1.9328320
 [5,]     5 1.0149958
 [6,]     6 2.3072891
 [7,]     7 1.6959516
 [8,]     8 2.2299489
 [9,]     9 1.8601363
[10,]    10 0.8358429
```

The set of item parameter estimates above must be interpretated jointly to get a picture on how the items function (see figure 1). Furthermore, it is useful to note that the lognormal (0,1) distribution has a mean of 1.65 and a standard deviation of 2.16. Finally, it should be taken that the slope/discrimination values are in the same scale as in the conventional GRM. So, values above, say 1.5 indicate a high discriminating power (e.g. Reise & Moore, 2012 and Reise et al., 2021).

With this information, let us take a look at two items with different functioning. Item 3 is both highly difficult and fairly high discriminating. To establish the first feature, note that the row of the 3 easiness estimates for this item contains very low values, which means that most of the respondents no longer even pass the first threshold (in other words, they obtain the lowest 1 item score). Furthermore, the trait level that is required to have a 50% chance of scoring at least in the second category (i.e. the first delta value corresponding to this item) is 7.08 that is, almost two and a half standard deviations above the mean in the trait scale. The third delta estimate (25.72) means that, scoring in the highest 4 category requires an improvably high trait level. Turning now to the discrimination, the slope estimate for this item is 1.22, and indicates that this item is fairly discriminating, so that it will measure accurately but only in a narrow range of trait values.

Consider now, a more balanced, “progressive” item: item 5. Given the easiness and location parameters, it follows that, to respond in the second category or above in this item does not require an exaggeratedly high trait level. Furthermore, reaching higher categories becomes progressively more difficult, as it should be. However, there are no abrupt jumps here, and a 50% chance of scoring in the highest category requires a trait standing of 9.08 (3.4 deviations above the mean). High, but not exaggeratedly high in a skewed distribution with a long right tail. Finally, the discriminating power of this item is medium, which suggests that the item will provide acceptably accurate measurement across a wider range of trait values than item 3.

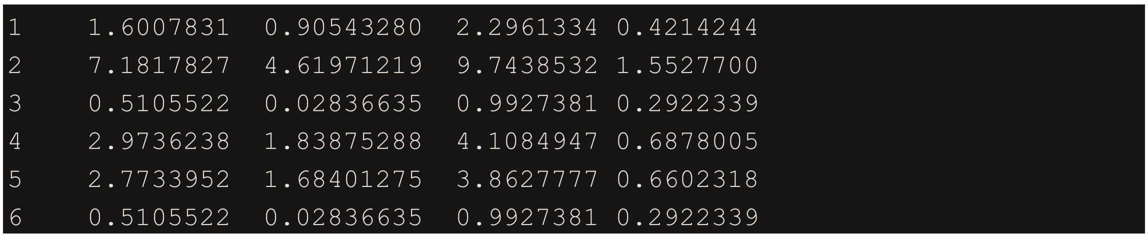
2. Calibration: Goodness of fit results

```
$GOF
               npar                fmin               chisq
             40.000               0.010              20.916
                 df              pvalue      baseline.chisq
             35.000                  NA           20334.979
        baseline.df     baseline.pvalue                 cfi
             45.000                  NA               1.000
                tli                nnfi                 rfi
              1.001               1.001               0.999
                nfi                pnfi                 ifi
              0.999               0.777               1.001
                rni               rmsea      rmsea.ci.lower
              1.001               0.000               0.000
     rmsea.ci.upper        rmsea.pvalue                 rmr
              0.000               1.000               0.018
         rmr_nomean                srmr        srmr_bentler
              0.020               0.020               0.018
srmr_bentler_nomean                crmr         crmr_nomean
              0.020               0.020               0.022
         srmr_mplus   srmr_mplus_nomean               cn_05
              0.018               0.020            2379.625
              cn_01                 gfi                agfi
           2739.759               1.000               1.000
               pgfi                 mfi
              0.467               1.007
```

No surprises here. Because the simulated data conforms to the model, the fit is near perfect, as it should be. Note that the chi-square value (20.916) is even below the number of degrees of freedom (35; its expected value).

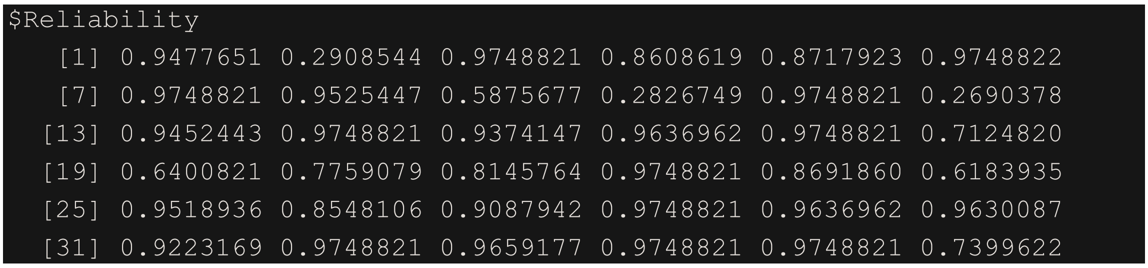
3. Scoring

```
$Scoring
            th       th_li      th_ls        se
```



```
1    1.6007831  0.90543280  2.2961334 0.4214244
2    7.1817827  4.61971219  9.7438532 1.5527700
3    0.5105522  0.02836635  0.9927381 0.2922339
4    2.9736238  1.83875288  4.1084947 0.6878005
5    2.7733952  1.68401275  3.8627777 0.6602318
6    0.5105522  0.02836635  0.9927381 0.2922339
```

4. Conditional reliability across trait levels

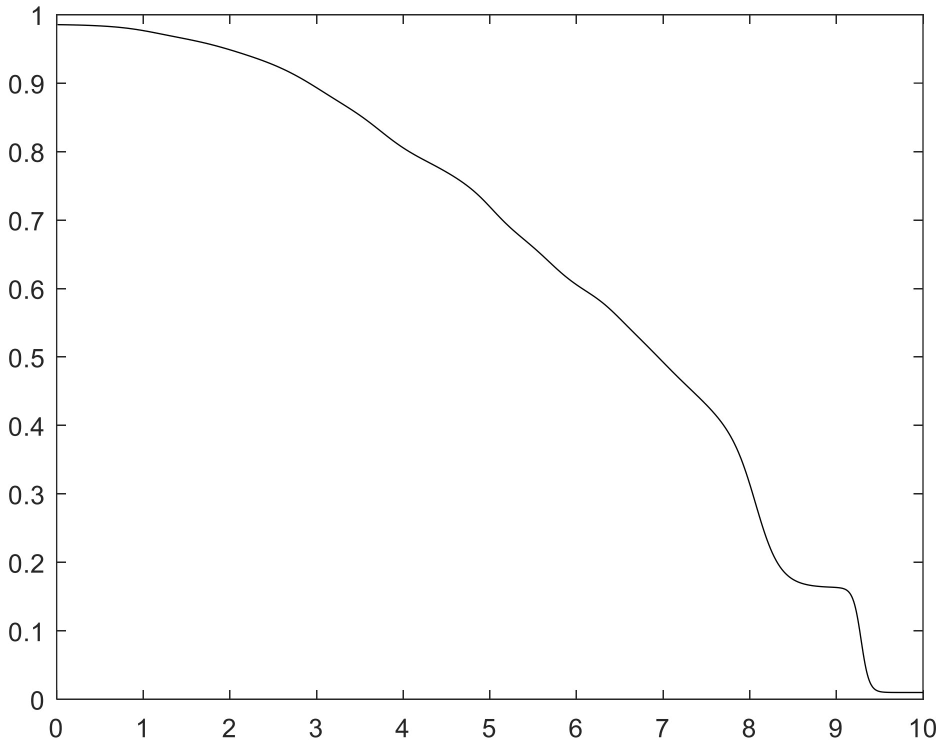


```
$Reliability
   [1] 0.9477651 0.2908544 0.9748821 0.8608619 0.8717923 0.9748822
   [7] 0.9748821 0.9525447 0.5875677 0.2826749 0.9748821 0.2690378
  [13] 0.9452443 0.9748821 0.9374147 0.9636962 0.9748821 0.7124820
  [19] 0.6400821 0.7759079 0.8145764 0.9748821 0.8691860 0.6183935
  [25] 0.9518936 0.8548106 0.9087942 0.9748821 0.9636962 0.9630087
  [31] 0.9223169 0.9748821 0.9659177 0.9748821 0.9748821 0.7399622
```

A joint interpretation (point estimates and conditional reliabilities) is also recommended here, and we shall interpret the scoring results of the first two participants. The point estimate of the first respondent (1.600) is below the 1.65 trait mean, and, at low trait values the LL-GRM is expected to measure with more accuracy. So, as expected, the accompanying reliability estimate (0.974) is very high, which means that the confidence interval around the trait estimate for this individual is relatively narrow.

The second respondent has a trait point estimate well above the trait mean (7.181) and so, the measure is expected to be rather inaccurate. Indeed, at this level, the reliability is very low (0.290) and so, the confidence interval around the point estimate is very wide, which evidences the uncertainty of the estimation.

(Extension note). To get a whole picture of the scoring results, if the conditional reliabilities in (4) were plotted against the estimated trait scores in (3), the following conditional information function would be obtained:



Clearly, the point estimates are accurate at low values, below the 1.65 mean, and marginally acceptable reliability (say above 0.70) can only be achieved up to one and a half deviation above the mean. For high levels, the scores are very imprecise, and that is how the model works: very accurate for differentiating between those individuals who have no trait manifestations and those who clearly do have them, but not sensitive enough to make finer differentiations between those with high levels.

References

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

Ferrando, P. J., Morales-Vives, F., & Hernández-Dorado, A. (2023). Measuring Unipolar Traits With Continuous Response Items: Some Methodological and Substantive Developments. *Educational and Psychological Measurement*, 00131644231181889.

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

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

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models.

Morales-Vives, F., Ferrando, P. J., & Dueñas, J. M. (2023). Should suicidal ideation be regarded as a dimension, a unipolar trait or a mixture? A model-based analysis at the score level. *Current psychology*, *42*(25), 21397-21411.

Reise, S. P., & Moore, T. M. (2012). An introduction to item response theory models and their application in the assessment of noncognitive traits. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 699–721). American Psychological Association. https://doi.org/10.1037/13619-037

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

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. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika*, *23*, 17-35.

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

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

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