InDisc

2019

D. Navarro-González & P.J. Ferrando

An R program for assessing person and item discrimination in typical-response measures

Contents

1. THEORETICAL BASES

Typical-response (personality and attitude) measures intended to measure a single dimension, are generally fitted using conventional factor-analytic (FA) models, or FA-related item response theory (IRT) models, such as the graded-response model (GRM) and the two-parameter model (2PM). In this type of modelling, the observed item response can be viewed as the outcome of an encounter between an item that has two characteristics: location/s and discrimination, and an individual that has a single characteristic: location or trait level. No individual feature is considered that can be viewed as equivalent to the item discriminating power. More technically, no 'dual' person discrimination parameter is considered in this type of models.

"Item discrimination" is a term that has multiple meanings in item analysis (Ferrando, 2012). For the present purposes, however, we shall adopt the original Thurstonian view, based on the concept of item discriminal dispersion (IDD), or fluctuation around the item position on the trait continuum (Edwards & Thurstone, 1952). An item that is interpreted in the same way for all the respondents, or across hypothetical repeated presentations to the same respondent, has small IDD, which means that: (a) has a well-defined position on the trait continuum (small fluctuation); (b) is able to consistently differentiate between individuals with different trait levels; and so, (c) is an accurate and consistent indicator of this trait.

The person counterpart of IDD, would be a parameter that models the fluctuation of the perceived trait level (i.e. person location) of this individual around its central value across the test items or hypothetical replications of the same item. This parameter has received different names in the literature, such as "person fluctuation" (Ferrando, 2007; Levine & Rubin, 1979; Lumsden, 1980), "person reliability" (Ferrando, 2004; Lumsden, 1977, 1978), or "slope of the person response function" (Trabin & Weiss, 1983, Jackson, 1986), and determines the consistency of the response pattern of this individual in terms of sensitivity of his/her responses to the different item locations. In InDisc we denote this parameter as "individual discriminal dispersion" or "person discriminal dispersion" (PDD), which makes it clear that it is the person counterpart of the IDD. The same as in the item case, low PDD means high individual discrimination, thus, a highly discriminating individual will have a well-defined trait level with small PDD, and so, will respond with high consistency regarding the different item locations, leading to response patterns that approach Guttman patterns (Coombs, 1948, Ferrando, 2004, 2013, Fiske, 1968). At the other end, an individual with low discrimination will have a poorly defined trait level, with high PDD, and so will be largely insensitive to the item locations and his/her response pattern will be almost at random. Both, evidence and literature review, suggests that individuals do indeed differ in the sensitivity of the responses they provide to personality and attitude questionnaires (Tellegen, 1988). So, dual models (see Fiske, 1968) in which both items and persons differ in terms of discriminating power seem to be the most plausible approach for fitting typical responses.

Apart from its plausibility, assessing PDD is submitted to have both, substantive and practical interest. At the substantive level, conventional models have explored the interpretation of the IDD, and have related this characteristic mainly to the degree of item ambiguity, type of stem and average stem length (e.g. DeFleur & Catton, 1957, Ferrando, 2013, Lumsden, 1980, Taylor, 1977). As expected, the meaning of PDD has been far less studied, but it has been hypothesized to be related to the relevance and degree of clarity and strength with which the trait is internally organized in the individual (Traitedness; e.g. Markus, 1977, Reise & Waller, 1993, Taylor, 1977, Tellegen, 1988). Evidence in this respect suggests that the PDD estimates can be effectively used to reflect traitedness (LaHuis et al. 2017, Reise & Waller, 1993). Finally, recent studies also suggest that PDD is related to general intelligence so that the more capable individuals tend to respond to personality questionnaires with smaller amounts of PDD (Escorial et al. 2019).

At a more practical level, the PDD estimates provide additional information about the consistency of the respondent's answering behaviour as well as the accuracy with which the trait location of this individual can be estimated. This information, in turn, can be of use in individual assessment, and has also been used in exploratory person-fit research (see Conijn et al., 2013). Finally, the PDD estimates are expected to have a moderating role in validity assessment (Ferrando, 2004, 2013) mainly for two reasons. First, the person location estimates of the less discriminating individuals are less reliable, which means that attenuated validity estimates are expected in this population. Second, those individuals for whom the trait is relevant are expected to be more likely to display a stronger correspondence between trait self-description and external trait-relevant variables (Markus, 1977, Paunonen, 1988).

1.1 Models implemented in InDisc

The models discussed in this section are "Dual Thurstonian Models" (DTMs) because (a) they include discrimination parameters for both items and individuals, and (b) these parameters are modelled using Thurstone's concept of discriminal dispersion.

In his comprehensive presentation of DTMs, Ferrando (2019) considered three different models: The DTM for continuous responses (DTCRM), the DTM for graded responses (DTGRM) and the DTM for binary responses (DTBRM). However, the DTBRM is simply the particular case of the DTGRM when the number of ordered response categories reduces to two, with no other distinctive feature that the usual 0-1 scoring typical in binary responses is replaced by the usual integer graded scoring (1-2 in this case). For this reason, only the DTCRM and the DTGRM will be discussed here. Essentially, both models can be formulated as extended unidimensional factor-analytic models with an extra discrimination parameter for each individual.

We shall start by defining the common features of both models. At the moment of answering item *j*, respondent *i* has a momentary trait (or perceived trait) value T_i whereas item *j*, has a momentary (perceived) location b_j , both values defined on the continuum of the trait that is measured. This trait is denoted by θ, and is assumed to have zero mean and unit variance in the population.

$$T_i \;=\; \theta_i + \omega_i \quad ; \quad b_j = \beta_j + \varepsilon_j . \tag{1}$$

The distribution of T_i over the test items is assumed to be normal with mean θ_i and variance $\sigma^2{}_i$, which are the parameters that characterize respondent *i*. θ_i is the central trait value, the single value that best characterizes the standing of this individual on the trait, while $\sigma^2{}_i$ is indeed the PDD of this individual. With regards to item *j*, the distribution of b_j, over respondents is assumed to be normal, with mean β_j, and variance $\sigma^2{}_{\varepsilon j.}$, where β_j, can be interpreted as the conventional location of this item, while $\sigma^2{}_{\varepsilon.}$ is the IDD.

We shall first consider the DTCRM. Let X_{ij} be the (approximately continuous) score of individual i in item j. For convenience, X_{ij} is scaled to have values between 0 and 1, which can be conveniently done with InDisc. The structural model for this score under the DTCRM is:

$$X_{ij} = 0.5 + \lambda_j (T_i - b_j) \tag{2}$$

The conditional distribution of X_j for fixed θ_i and $\sigma^2{}_i$ is normal, with expectation:

$$E(X_{ij} \mid \theta_i, \sigma^2{}_i) = 0.5 + \lambda_j (\theta_i - \beta_j) \tag{3}$$

The conditional expectation in (3) is a linear function of the weighted person-item distance $\lambda_j(\theta_i\text{-}\beta_j)$. When $\theta_i > \beta_j$, the expected score is above the 0.5 response scale midpoint (i.e. 0.5), and when the person location matches the item location, the expected item score is the midpoint. So, as proposed above, β_j can be interpreted as a standard IRT difficulty index: it is the point on the trait continuum that marks the transition from the tendency to disagree/not endorse the item to the tendency to agree/endorse.

The conditional variance is given by:

$$Var(X_{ij} \mid \theta_i, \sigma^2{}_i) = \lambda_j{}^2 (\sigma_i^2 + \sigma_{\varepsilon j}^2). \tag{4}$$

And clearly shows the role of the IDD and PDD in this model. Note that the expected score in (3) is the same regardless of the magnitude of these dispersions. However, the item and person dispersions determine the expected consistency of the responses. Thus, when both PDD and IDD are small, so is the conditional variance in (4) which means that the observed score will be close to the expected score.

We turn now to the DTGRM, which is obtained from the linear model so far discussed by using an Underlying-Variables Approach (UVA; Muthén, 1984). Consider now that the observed item score X_j is a categorical variable, and assume that there is a latent response variable Y_j that underlies X_j, so that the following model holds for Y_j

$$Y_{ij} = \alpha_j (T_i - b_j)\,. \qquad (5)$$

Model (5) is the same model as (2) without the midpoint intercept term and with the variance of Y_j fixed to 1. This restriction is made for identification purposes, and means that the scale parameter λ_j is now a standardized loading α_j.

As is usual in the UVA, the relation between Y_j and the observed score X_j is assumed to be a step function governed by a threshold mechanism. In InDisc we shall consider, for any ordered-categorical response (included the binary response), the integer-value scoring 1,2,…With this scoring, the step relation between X_j and Y_j is:

$$\begin{aligned} &X = 1 \quad if \quad Y < \tau_1 \\ &X = 2 \quad if \quad \tau_1 \le Y < \tau_2 \\ &X = 3 \quad if \quad \tau_2 \le Y < \tau_3\,. \\ &\ldots \\ &X = c \quad if \quad \tau_{c-1} < Y \end{aligned} \qquad (6)$$

Where τ_k is a threshold and c is the number of response categories. From this modeling it follows the usual UVA result that the product-moment correlation between Y_j and Y_k is the polychoric correlation between X_j and X_k .

The probability of scoring k in item j for fixed θ_i and $\sigma^2{}_i$ is now:

$$
\begin{aligned}
&P(X_{ij}=k\,|\,\theta_i,\sigma^2{}_i)= \\
&\Phi\left(\frac{1}{\sqrt{\sigma_i^2+\sigma_{\varepsilon j}^2}}(\theta_i-(\beta_j+\frac{\tau_{jk-1}}{\alpha_j}))\right)-\Phi\left(\frac{1}{\sqrt{\sigma_i^2+\sigma_{\varepsilon j}^2}}(\theta_i-(\beta_j+\frac{\tau_{jk}}{\alpha_j}))\right). \\
&=\Phi\big(\gamma_{ij}(\theta_i-\delta_{jk-1})\big)-\Phi\big(\gamma_{ij}(\theta_i-\delta_{jk})\big)
\end{aligned}
\qquad (7)
$$

Where Φ is the c.d.f. of the standard normal distribution. To see how the DTGRM functions, consider first that the central person location θ_i determines the response category that has the greatest probability of being endorsed by respondent *i*. As for the role of the IDD and the PDD, the smaller they are, the greater the probability of endorsing this category is, and the smaller the probability of endorsing the remaining categories become. To provide an example that focuses only on the PDD, consider a respondent whose person location is between $\delta_{j\ k-1}$ and $\delta_{j\ k}$. As his/her PDD approaches zero, the probability of endorsing category *k* increases, whereas the probability of endorsing the remaining categories decreases. So, the process of responding of this individual becomes more deterministic. At the opposite extreme, as the PDD increases, the probability of responding in different categories becomes progressively more undifferentiated.

In closing this section, we shall briefly discuss the relations between the models so far discussed, and the corresponding conventional models with no individual discrimination parameters. When the PDD is a constant for all the respondents, the DTCRM becomes the standard unidimensional linear FA model (this can be derived from equation 2; see Ferrando, 2012), whereas the DTGRM reduces to the normal-ogive version of Samejima's (1969) GRM (this result can be seen directly from equation 7). Thus, from the present framework, the standard existing models can be viewed as the limiting case of the dual models fitted by InDisc when the amount of PDD is the same for all the respondents.

2. PRACTICAL REQUIREMENTS FOR FITTING THE MODELS

For the two models available in InDisc, accurate measurement of the trait level has the same requirements as the corresponding standard model in which only this person parameter is estimated. So, even in the less informative case of binary items, acceptable trait estimates for all the respondents are expected to be achieved with relatively short tests of acceptable quality. Our position, however, goes in the same line as that of Emons et al. (2007): that even in well-designed tests, 20 is a safe minimal number of items for achieving accurate individual trait estimation.

Obtaining accurate person discrimination estimates is far more demanding than obtaining accurate trait estimates. The amount of measurement error for the PDD estimates depends mostly on (a) the number of items, (b) the distance between the item and the person locations on the trait continuum, and (c) the quality (i.e. discrimination) of the items. Thus, for a given respondent with a certain trait level, individual discrimination can only be estimated accurately if there are enough items with good discrimination that are sufficiently distant from his/her trait location. At the overall level then, it follows that reliable estimation of the person discriminations for most of the respondents requires a test that is long enough and made up of items that have a wide dispersion of locations across the trait continuum and a relatively low amount of IDDs.

Apart from the general determinants above, theoretical and empirical results, clearly suggests that the minimal requirements vary largely according to the model that is fitted to the data, and, in the case of the model for ordered-categorical responses, to the number of response categories. (see Ferrando, 2009). In favourable conditions: item locations that are widely spread and evenly distributed around the mean trait level (zero in the InDisc scaling), and that have acceptable discriminations, reasonable PDD estimates for all the sample respondents can be obtained from 15 items both from the continuous model and from the graded-response model with 5 or more response categories. In the case of binary responses, the 20-item recommendation is a safer minimum (Ferrando, 2004). We emphasize again that we are assuming items that vary widely in location when providing these minimal requirements. Accurate individual discrimination is plainly unfeasible in tests in which all of the items have similar locations, no matter how long is the test.

Until more extensive simulation is carried out, our recommendations at present are the following. First, inspect the item means and item-total correlations using conventional analysis, and check whether they met the InDisc requirements. Second, check that (a) the chosen InDisc model fits the data better than the corresponding conventional model, and (b), the estimated reliability of the person discrimination estimates is acceptable both at the individual and at the marginal level. Section 4 explains how to assess these two important points.

We shall now discuss practical requirements for choosing between the DTCRM and the DTGRM. The simpler DTCRM assume the item scores to be continuous and unbounded, which is never the case with psychometric item responses (at most they are bounded and approximately continuous). The bounded nature of the item scores imply that (a) the item-trait relations are nonlinear rather than linear, and (b) the conditional distributions become more asymmetrical and with decreased variance toward the ends of the scale, whereas in equation (4) the conditional variance is assumed not to depend on the trait level. So, for ordered-categorical scores, the DTGRM is theoretically more appropriate than the DTCRM, and the latter must be always viewed as an approximation.

However, the DTCRM has also the advantages of its simplicity and robustness. A more in depth discussion is provided in Ferrando (2002), but we shall note here that in most practical applications in the personality domain based on items with 5 or more categories, the DTCRM works generally well. So, it is a recommended option when (a) the number of categories is 5 or more, and (b) the number of items is relatively short, thus leading to potentially unstable person estimates if the DTGRM is chosen.

3. ESTIMATION PROCEDURES

For both the DTCRM and the DTGRM, InDisc uses a conventional two-stage conditioned estimation procedure (e.g. McDonald, 1982). In the first stage, (item calibration) the structural parameters of the chosen model are estimated. These parameters are: item locations, IDDs, and the average PDD in the population. In the second stage (scoring) estimates of the central level or person location (θ_i) and of the PDD (σ_i^2) are obtained for each individual. Model-data fit or model appropriateness is also assessed at both, the calibration and the scoring stages.

The estimation procedures in InDisc have been chosen mainly for practical reasons. At the calibration stage, we wanted a simple and robust procedure that allow stable estimates to be obtained even in large item sets and not too large sample sizes. At the scoring stage, we wanted a procedure able to provide finite and plausible estimates for all the respondents in the sample under analysis.

3.2 Item Calibration

Both the DTCRM and the DTGRM can be calibrated by fitting a unidimensional FA model, with additional identification restrictions, to the appropriate correlation matrix: Product-moment (DTCRM) or polychoric (DTGRM). In both cases, the standard FA output would provide the standardized loadings α and the corresponding standardized residual variances. Now, from the general assumptions above, the following result is obtained.

$$\frac{1-\alpha^{2}{}_{j}}{\alpha^{2}{}_{j}} = E(\sigma_i^2) + \sigma_{\varepsilon j}^2 . \tag{8}$$

Result (8) means that the standardized loadings do not contain sufficient information to separately identify the average PDD and the IDDs. To achieve this identification, InDisc uses a scaling restriction based on a 'marker' item. The item with the largest loading (i.e. with the smallest IDD) is taken as a marker, and so, treated as if its IDD was zero. Then, relative to this scaling, the average PDD is estimated as

$$\frac{1-\hat{\alpha}^{2}{}_{(\max)}}{\hat{\alpha}^{2}{}_{\max}} = \hat{E}(\sigma_i^2) . \tag{9}$$

where $\hat{\alpha}_{(\max)}$ is the largest estimated standardized loading. The remaining IDDs are obtained from equation (8).

As a summary, the calibration procedures as implemented in InDisc are as follows. For the DTCRM, first, the unidimensional FA model is fitted to the inter-item product-moment correlation matrix, and estimated loadings and residual variances are obtained. Next, the item locations in (3) are obtained from the marginal means (see Ferrando, 2019 for details). Finally the IDDs and the average PDD are obtained by using the marker identification restriction in equation (9).

In the DTGRM case, the unidimensional FA model is fitted to the inter-item polychoric correlation matrix, and estimated loadings and residual variances are obtained. Next, the FA-based threshold estimates are reparameterized to location estimates (the δ's in equation 7). Finally the IDDs and the average PDD are obtained by using the same procedure as in the DTCRM.

The FA calibration procedure is the same for both models but based on the corresponding inter-item correlation matrix. The unidimensional FA model is fitted to this matrix using the minimum-residual unweighted least squares (ULS) criterion as implemented in the “psych” R package (Revelle, 2018; see the “psych” guide in https://cran.r-project.org/web/packages/psych/psych.pdf for more details).

3.2 Individual scoring

The search for a scoring procedure that produces finite and plausible estimates for all respondents has led us to choose Bayes expected a posteriori (EAP, Bock & Mislevy, 1982) score estimation for both models. EAP has also an additional advantage here: that the regression towards the mean or shrinkage phenomenon characteristic of Bayes estimation is towards an appropriate central value: the average PDD estimated as a structural parameter at the calibration stage. For both, the DTCRM and the DTGRM, details on EAP estimation of θ and σ^2_i are provided in Ferrando (2019).

The most relevant information to be provided in this section is that concerning the prior distributions. In InDisc, the prior for θ is set as standard normal, and for σ^2_i is set as a scaled inverse χ^2 distribution, a theoretically appropriate prior given that, in the InDisc modelling, the PDDs are variances (Novick & Jackson, 1974). With regards to this last prior, the most relevant practical problem is to determine the amount of weakness or diffuseness of the prior. A too weak or diffuse (i.e. uninformative) prior might lead to implausible or out of bound PDD estimates for some individuals, especially in the case of items with few categories and/or short tests. A too tight prior, on the other hand, would provide PDD estimates with very little variability and concentrated around the average PDD estimated at the calibration stage. The ‘compromise’ choice taken in InDisc is to set the scale parameter and the degrees of freedom of the inverse chi-square distribution so that (a) the prior mean coincides with the average PDD obtained at the calibration stage, and (b) the prior variance is 0.5. Extensive program checks suggest that this choice produces plausible estimates in most of the conditions expected to be found in practice. If, on the other hand, implausible results for some respondents are obtained with this prior, a tighter alternative prior will be used, with a scaling parameter of 4 and 6 degrees of freedom. This tighter prior will only be applied for the respondents from which plausible estimates cannot be obtained. If, once this tighter prior have been used, there still are some implausible results, an even tighter prior will be used, with a scaling parameter of 8 and 10 degrees of freedom for the affected respondents. In closing this section we note that for both, θ and $\sigma^2{}_i$ priors, InDisc uses as default rectangular quadrature in 30 equally spaced points (see Mislevy, 1986). However, the number of quadrature points can be modified at the user's request.

4. ASSESSING MODEL APPROPRIATENESS

Item calibration by ULS FA provides conventional measures of goodness of model-data fit at the structural correlation level. In particular, of the ULS-derived goodness-of-fit indicators provided in the “psych” output, the InDisc output provides four: the chi-square statistic and corresponding degrees of freedom; the root mean square of the residuals (RMSR; a descriptive absolute index), The Tucker-Lewis or Non-Normed fit index (a comparative index relative to the null independence model), and the root mean squared error of approximation (RMSEA; an index of relative fit per degree of freedom).

If the structural fit of the model is considered acceptable according to the measures above, it can be assumed that the unidimensional FA model fits well the inter-item correlation matrix, which is a necessary, but not sufficient condition for considering that the corresponding DTM is appropriate. In effect, the standard model and the DT model are indistinguishable at the implied-inter-item correlation level. So, if fit is acceptable at the calibration stage, what is needed for considering the DTM as appropriate is to further assess if there is non-negligible variation in the PDD estimates over respondents. Otherwise, the simpler standard model in which all respondents are assumed to be the same amount of PDD would have to be chosen according to the parsimony principle.

The approach implemented in InDisc is based on a likelihood ratio (LR) statistic. For a single respondent *i*, let $L_i^0(\hat{\theta}_i, \hat{\sigma}^2)$ be the value of the likelihood function evaluated by using the person location estimate that is obtained under the restriction that all the PDDs have a constant value (the estimate of this constant value is the mean of the PDDs obtained at the calibration stage according to equation 9). Now, let $L_i^1(\hat{\theta}_i, \hat{\sigma}_i{}^2)$ be the corresponding likelihood function value using both the person location and the PDD estimate. The LR statistic and the transformed value provided by InDisc are

$$\Lambda_i = \frac{L_i^0(\hat{\theta}_i, \hat{\sigma}^2)}{L_i^1(\hat{\theta}_i, \hat{\sigma}_i^{\,2})} \quad ; \quad s_i = -2\ln(\Lambda_i)\,. \tag{10}$$

The LR statistic Λ_i is a descriptive normed index with values in the range 0-1. Values close to 0 indicate that the DTM provides a substantially better fit than the corresponding standard model. As for s_i, under very restrictive conditions it could be considered to be a value randomly drawn from a χ^2 distribution with one degree of freedom. And, by further assuming experimental independence between respondents, the sum $Q=\Sigma s_i$ asymptotically approaches a χ^2 distribution with N degrees of freedom. However, we do not propose Q as a rigorous and strict inferential statistic, but only as a useful reference for assessing whether the DTM fits better than its standard counterpart. In this spirit, Q values two or more times greater than N would clearly suggest that the DTM is more appropriate than the standard model.

That there is non-negligible variation in the PDDs is of little practical interest if this characteristic cannot be estimate with reasonable accuracy for most of the individuals in the population of interest. For each individual, the InDisc output provides the point estimates of the trait level and the PDD as well as the corresponding posterior standard deviations (PSDs) which serve as standard errors (e.g. Bock & Mislevy, 1982) and the individual reliability estimates based on these standard errors. So, the magnitude of the PSD and corresponding reliability allows the practitioner to judge the accuracy of the estimates (both trait level and PDD) for the individual under scrutiny. Furthermore, a marginal reliability estimate can be obtained by averaging the squared PSDs in the sample of *N* individuals (Brown & Croudace, 2015). In InDisc, this marginal reliability estimate is obtained as:

$$Trait\ Levels : \rho(\hat{\theta}) = \frac{Var(\theta)}{Var(\theta) + \frac{\sum_{i}^{N}\left[PSD(\hat{\theta}_i)\right]^2}{N}}$$

$$PDDs : \rho(\hat{\sigma}^2) = \frac{Var(\sigma^2)}{Var(\sigma^2) + \frac{\sum_{i}^{N}\left[PSD(\hat{\sigma}^2{}_i)\right]^2}{N}} \qquad (11)$$

Expressions (11) differ from those provided in Ferrando (2019). They were chosen here because, of all the possible estimates of the marginal reliability, those obtained with (11) were the closest to the empirical reliabilities directly estimated by using a split-half schema. As for interpretation, acceptable marginal reliability would indicate that the characteristic of interest (trait level or PDD) can be measured with reasonable accuracy for most of the respondents belonging to the population of interest.

5. PROGRAM USAGE

5.1 InDisc R package: Techical details

InDisc was developed in R 3.6.1, and is distributed as an R package. The current version, which is available through CRAN, contains one main function that implements all the procedures described above. It runs with R versions more recent than 3.5.0, and it is operational in any operational system that supports R (Windows, Linux and Mac OS).

5.2 Program availability and installation

The package is available through the CRAN website (https://CRAN.R-project.org/package=InDisc), and can be downloaded directly from the CRAN website, and installed manually. Otherwise, the package can be installed as any other R package through the command line using:

```
> install.packages("InDisc")
```

5.3 Usage and input arguments

The function usage is the following:

```
> InDisc(SCO, nquad = 30, model = "linear", approp = FALSE,
display = TRUE)
```

Where the description of the input arguments are:

SCO	The data matrix containing the item scores.
nquad	The number of quadrature points for EAP estimation (default is 30).
model	The model to be used: “graded” (DTGRM) or “linear” (DTCRM).
approp	Determines if the appropriateness indices will be computed and printed in the console (logical variable, FALSE by default).
display	Determines if the output will be displayed in the console (logical variable, TRUE by default).

5.4 Output

The output will be printed in the console if the argument display was TRUE, and it will look as follows:

```
$INDIES
             theta       PDD PSD (theta)  PSD (PDD)   th reli  PDD reli
FAC -0.3536449769 0.5276088   0.1982579 0.25969195 0.9606938 0.8604710
     0.8388829281 0.2751096   0.1764851 0.11729701 0.9688530 0.9679779
    -0.5665837603 0.5309085   0.2022677 0.23957794 0.9590878 0.8787283
    -0.5126518601 0.2693279   0.1733151 0.12141829 0.9699619 0.9657665
     1.3975053292 0.3266656   0.1808566 0.15912717 0.9672909 0.9426106
     1.4531437169 0.2554437   0.1736983 0.11075976 0.9698289 0.9713483
     1.1486078686 0.3086771   0.1781839 0.15151937 0.9682505 0.9476867
    -1.2574619814 0.4221989   0.1875729 0.22506741 0.9648164 0.8914269
    -1.8535233446 0.2297871   0.1667344 0.10339718 0.9721996 0.9749386
     1.2631254196 0.2726965   0.1713418 0.12756690 0.9706420 0.9623454

[ reached getOption("max.print") -- omitted 90 rows ]

$degrees_of_freedom
[1] 560

$Model_Chi_square
[1] 844.8372

$RMSR
[1] 0.08384021

$TLI
[1] 0.7335781

$RMSEA
```

```
[1] 0.09945507

$EVARI
[1] 0.5346372

$reli_theta
[1] 0.9589096

$aver_r_theta
[1] 0.9589096

$reli_PDD
[1] 0.8217937

$aver_r_PDD
[1] 0.8491271

$LR_stat
[1] 0.2079331

$Q_Chi_square
[1] 318.5465
```

All the data will be stored in a data frame if the user introduces an output variable name before the function. On this data frame, the user can find all the indices as the following arguments:

INDIES	Matrix including the theta scores, the PDDs, the PSDs (theta), the PSDs (PDD) and the reliabilities for the theta scores and PDDs for each participant.
degrees_of_freedom	Degrees of freedom for the fitted model.
Model_Chi_square	Chi Square statistic for the fitted model.
RMSR	Root Mean Square of the Residuals.
TLI	Tucker Lewis Index of factorial reliability.
RMSEA	Root Mean Squared Error of Approximation.
EVARI	Average of the PDDs.
reli_theta	Marginal reliability of the trait estimates.

aver_r_theta	Average of the individual reliabilities (trait level).
reli_PDD	Marginal reliability of the PDD estimates.
aver_r_PDD	Average of the individual reliabilities (PDD).
LR_stat	Likelihood ratio statistic.
Q_Chi_square	Approximate Chi Square statistic with N degrees of freedom.

6. Empirical example

The following example uses the dataset included in the InDisc package: CTAC35, which is a dataset containing 758 observations and 35 items, corresponding to the CTAC questionnaire (Pallero, Ferrando, & Lorenzo-Seva, 1998). The CTAC questionnaire measures anxiety in situations related to visual deficit, is intended to be used in the general adult population with severe visual impairment, and uses a 5-point Likert format. CTAC was designed to assess two subscales: cognitive anxiety and physiological anxiety, but since they are highly correlated, they can be considered related subscales from an overall anxiety scale.

The usage of the InDisc package should be the following:

```
> InDisc(SCO = CTAC35, nquad = 30, model = “graded”, approp =
TRUE, display = TRUE)
```

The output looks like the following:

```
$INDIES
           theta       PDD PSD (theta)  PSD (PDD)   th reli  PDD reli
FAC -0.065374703 0.2582460   0.1497897 0.12049315 0.9775631 0.9662687
     0.022498430 0.7283670   0.2061971 0.38846581 0.9574828 0.7337612
    -0.191278258 0.3555251   0.1643611 0.15737810 0.9729854 0.9437947
     0.643456856 0.5580985   0.1918494 0.29533356 0.9631938 0.8266386
     1.316853765 0.3571208   0.1791750 0.18185971 0.9678963 0.9263364
     0.675945448 1.7302302   0.2779443 0.81747116 0.9227470 0.3836153
    -1.655209645 0.7795517   0.2840174 0.50440865 0.9193341 0.6204424
    -2.277606086 0.3663463   0.2985836 0.22130346 0.9108478 0.8946487
    -1.183982073 0.3013486   0.1752420 0.15933348 0.9692902 0.9424702
     0.970947058 0.3051314   0.1571833 0.15264381 0.9752934 0.9469487
[ reached getOption("max.print") -- omitted 748 rows ]


$Degrees_of_freedom
[1] 560

$Model_Chi_square
[1] 3441.21

$RMSR
[1] 0.06176572

$TLI
[1] 0.7471719
```

```
$RMSEA
[1] 0.09051577

$EVARI
[1] 0.8168513

$reli_theta
[1] 0.9446712

$aver_r_theta
[1] 0.9446712

$reli_PDD
[1] 0.6344124

$aver_r_PDD
[1] 0.7152701

$LR_stat
[1] 0.656112

$Q_Chi_square
[1] 3605.168
```

We start first by assessing the appropriateness of the fitted model (bottom rows in the table above). The unidimensional UVA model chosen here fitted the correlation matrix acceptably well, mainly in terms of the residual indices. As for the appropriateness of the DTGRM compared to that of the standard model, the approximate chi square statistic suggests that there is non-negligible variation in the PDD values across individuals, which supports the use of the dual model. However, this variation neither is excessively large in terms of the LR statistic.

We turn now to the scoring results. The first rows above, show the theta point estimates, the PDD point estimates, and their corresponding PSDs and reliabilities for each participant (this example shows the first 10 ones). Thus, for the first respondent, the trait point estimate (i.e. anxiety level) is average (near the zero mean), and the discriminal dispersion is relatively low (compared to the average PDD estimate of 0.81). So, it can be inferred that this individual answered the CTAC items with high consistency. Both the reliabilities corresponding to the trait and PDD point estimates for this individual are rather high, which implies that the point estimates are accurate and can be trusted.

The marginal reliabilities are also quite acceptable. That of the trait estimates is similar to the conventional reliability of test scores that can be obtained with a good personality test (0.944 is quite good by all standards). The marginal reliability of the PDDs is smaller, but this is an expected result (see Ferrando, 2004). For the PDD standards it can be considered quite acceptable. Overall, the results suggest that: (a) the anxiety levels can be assessed with high accuracy for most of the individuals of the population for which the CTAC is intended, and (b) the PDDs can be assessed with reasonable accuracy in most individuals of this population. Overall, the results suggest that the choice of the DTGRM is appropriate and provide useful information beyond that can be obtained from the standard model.

References

Ferrando, P. J. (2012). Assessing the discriminating power of item and test scores in the linear factor-analysis model. *Psicológica, 33*(1), 111-134.

Edwards, A. L., & Thurstone, L. L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17*(2), 169-180.

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

Brown, A., & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.). Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (pp. 307-333). New York: Routledge.

Conijn, J. M., Emons, W. H., van Assen, M. A., Pedersen, S. S., & Sijtsma, K. (2013). Explanatory, multilevel person-fit analysis of response consistency on the Spielberger State-Trait Anxiety Inventory. *Multivariate Behavioral Research, 48*(5), 692-718.

Coombs, C.H. (1948). A rationale for the measurement of traits in individuals. *Psychometrika, 13*, 59-68.

DeFleur, M.L. & Catton, W.R. (1957). The limits of determinacy in attitude measurement. *Social Forces, 35*, 295-300.

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological methods, 12*(1), 105.

Escorial, S., Navarro-González, D., Ferrando, P. J., & Vigil-Colet, A. (2019). Is individual reliability responsible for the differences in personality differentiation across ability levels? *Personality and Individual Differences, 139*, 331-336.

Ferrando, P.J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research, 37*, 521-542.

Ferrando, P.J. (2004). Person reliability in personality measurement: an item response theory analysis. *Applied Psychological Measurement, 28*, 126-140.

Ferrando, P. J. (2007). A Pearson-Type-VII item response model for assessing person fluctuation. *Psychometrika, 72*, 25.

Ferrando, P.J. (2009). A graded response model for measuring person reliability. British *Journal of Mathematical and Statistical Psychology, 62*, 641-662.

Ferrando, P. J. (2013). A general linear framework for modeling continuous responses with error in persons and items. *Methodology, 9*, 150-161.

Ferrando, P. J. (2014). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate Behavioral Research, 49*, 390-405.

Ferrando, P. J. (2016). An IRT modeling approach for assessing item and person discrimination in binary personality responses. *Applied Psychological Measurement, 40*, 218-232.

Ferrando, P. J. (2019). A Comprehensive IRT Approach for Modeling Binary, Graded, and Continuous Responses With Error in Persons and Items. *Applied Psychological Measurement 43*(5), 339-359.

Fiske, D.W. (1968). Items and persons: Formal duals and psychological differences. *Multivariate Behavioral Research, 3*, 393-401.

Jackson, D. N. (1986). The process of responding in personality assessment. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires (pp. 123-142). Berlin: SpringerVerlag.

LaHuis, D. M., Barnes, T., Hakoyama, S., Blackmore, C., & Hartman, M. J. (2017). Measuring traitedness with person reliabilities parameters. *Personality and Individual Differences, 109*, 111-116.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1*(4), 477-482.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31*, 19-26.

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement, 4*, 1-7.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*, 63-78.

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

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

Muthén, B. (1984) A general structural equation model with dichotomous, ordered, categorical and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Novick, M. R., & Jackson, P. H. (1974). Statistical methods for educational and psychological research. McGraw-Hill.

Pallero, R., Ferrando, P.J., & Lorenzo, U. (1998). Questionnaire Tarragona of anxiety for blind people. In E. Sifferman, M. Williams, & B.B. Blasch, (eds.): The 9th Internacional Mobility Conference Proceedings, (pp 250-253). Atlanta: Rehabilitation Research and Development Center.

Paunonen, S.V. (1988). Trait relevance and the differential predictability of behavior. *Journal of Personality, 56*, 599-619.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143-151.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (Psychometrika Monograph No. 17). Iowa City: Psychometric Society.

Taylor, J.B. (1977). Item homogeneity, scale reliability, and the self-concept hypothesis. *Educational and Psychological Measurement, 37,* 349-361.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56,* 622-663.

Torgerson, W. (1958). Theory and methods of scaling. New York: Wiley.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In New horizons in testing (pp. 83-108). Academic Press.