Psicológica (2013), 34, 383-406.

Measuring anxiety in visually-impaired people: A comparison between the linear and the nonlinear IRT approaches

Pere J. Ferrando[*1], Rafael Pallero[2], Cristina Anguiano-Carrasco[1]

[1]Universidad 'Rovira i Virgili' (Spain); [2]ONCE (Spain)

The present study has two main interests. First, some pending issues about the psychometric properties of the CTAC (an anxiety questionnaire for blind and visually-impaired people) are assessed using item response theory (IRT). Second, the linear model is compared to the graded response model (GRM) in terms of measurement precision, sensitivity to change, and person fit, and the results are also used to illustrate the functioning and advantages of IRT models. The participants were 670 blind or visually-impaired people from different Spanish cities. The results showed that the CTAC scores are accurate enough for practical purposes, and that respondents are quite consistent in their responses. Model-data fit was acceptable in both cases, and both models lead to similar results regarding the trait estimates, with the exception of extreme respondents who were better assessed with the linear model. The GRM assessed measurement precision better, and both models showed high sensitivity to change around cut-off values. Person-fit results were also similar in both models.

Visual impairment is expected to have a substantial impact on individuals, because of the shortcomings and restrictions it entails in their everyday life activities (WHO, 2001). Their self-assessment of the situation is usually made in a context of severe anxiety, which can give rise to a self-perception of inefficacy (e.g. I am unable to do the things I used to do). This perception may, in turn, be associated with a series of anxiety responses that are both physiological (e.g. sweaty hands) and cognitive (negative worries, recurring thoughts; see e.g. Lazarus, 2000). Anxiety responses of the

[*] The research was partially supported by grants from the Spanish National Organization of the Blind (ONCE) and the Spanish Ministry of Economy and Competitiveness (PSI2011-22683). Correspondence: Pere Joan Ferrando. Universidad 'Rovira i Virgili'. Facultad de Psicologia. Carretera Valls s/n . 43007 Tarragona (Spain). E-mail:perejoan.ferrando@urv.net

types just described are likely to appear in everyday situations, even in people who have already acquired coping resources (Welsh, 1997). Psychological assessment of the anxiety responses discussed above would make it possible to design intervention programs aimed at reducing the high anxiety levels. This reduction is expected to have three main effects: first, quality of life would be improved; second, the process of learning the adaptive skills needed to cope with visual impairment would be facilitated; and third, the skills acquired would be easier to maintain and generalize.

At present there are very few instruments available to measure psychological variables related to visual loss. In particular, anxiety scales specifically intended for blind and visually impaired people are very scarce. In English (a) some existing measures intended for the general population have been adapted to create more specific anxiety scales (Hardy, 1968), and (b) some questionnaires that measure general adjustment have included a set of anxiety items (Bauman, 1963; Dodds, 1991; Fitting 1954).

The state of affairs outlined above prompted our research group to develop a Spanish instrument specifically intended to measure anxiety for the blind and visually-impaired. The instrument was called the CTAC (the Spanish acronym for 'Tarragona Anxiety Questionnaire for the Blind'), and was designed to measure specific anxiety related to visual impairment in a range of everyday situations of the type discussed above. More specifically, and as mentioned above, the CTAC aimed to measure two related (physiological and cognitive) components of anxiety, so it was conceived as a bi-dimensional instrument (Pallero, Ferrando, & Lorenzo-Seva, 2006; Ferrando, Lorenzo-Seva & Pallero, 2009). Since its initial conception we have assessed the dimensionality of the CTAC scores in a series of studies and, so far, the results can be summarized as follows. First, the unidimensional model already fits the data reasonably well. Second, fitting the bi-dimensional model slightly improves the model-data fit and leads to a clearly identifiable solution with two highly correlated factors. So, we believe that both the use of the test as an essentially unidimensional instrument or as a bi-dimensional instrument (as e.g. in Ferrando et al., 2009) is justifiable. And, in fact, the scoring procedure of the CTAC allows both a double scoring and a single general scoring to be computed. Although it will not be discussed any further in this article, a plausible approach for integrating both views is to fit a bifactor solution (e.g. Reise, 2012).

The target population for which the CTAC is intended is relatively reduced, and all the studies that have been made on how it performs (including the dimensionality studies discussed above) have been based on

small samples. This is the main reason why, so far, relatively simple approaches have been used to assess its psychometric properties: classical test theory (CTT) and exploratory factor analysis (FA). The results obtained so far are positive: The CTAC scores show acceptable reliability levels, and are useful for assessment and decision purposes (Pallero et al., 2006). The usefulness of the test has prompted it to be used more and a relatively large sample is now available. This situation can be exploited to assess some issues that could not be satisfactorily addressed with the simple methodology used so far. The main issues that need to be assessed are: (a) the amount of individual precision of the trait estimates, particularly around the potential cut-off points that are used to decide the need for psychological treatment, (b) the sensitivity of the trait estimates for detecting changes (mainly treatment-induced), and (c) the assessment of individual consistency when responding to the questionnaire. Given the aims of the test the relevance of the first and second issues is clear. The CTAC scores are mainly intended to be used for flagging respondents with high anxiety levels, and also in follow up studies to assess improvement due to psychological treatment. As for the third issue, given the importance of the decisions derived from the interpretation of the CTAC scores, it is critical to assess whether the participant is responding consistently to the questionnaire. If he/she is not, the score obtained must be considered as uninterpretable.

For the three issues we aim to study, test length is critical in the appropriate assessment of the corresponding properties. The measurement precision, sensitivity for detecting change and the power for detecting person misfit all improve as the number of items increases. For this reason in the present study we shall treat the CTAC scores as essentially unidimensional, so they will be taken from the complete item set.

Methodological Basis and Purposes of the Study

Methodologically, the three issues discussed above are better addressed within an item response theory (IRT) framework. Issues (a) and (b) are better assessed by using information curves and (potentially) optimal trait level estimates (instead of raw scores). Issue (c) can be addressed by using IRT-based person-fit assessment.

So far, IRT applications to the assessment of personality variables in visually-impaired populations have been relatively scarce, and most of the reported studies are conventional calibrations of existing binary-item instruments using a standard model (Ferrando, Pallero, Anguiano-Carrasco & Montorio, 2010; Lamoureux et al., 2007; Gothwal, Wright, Lamoureux,

& Pesudovs, 2009; Cochrane, Marella, Keeffe, & Lamoureux, 2011). Unlike these reported studies, however, the CTAC items use a 5-point graded response format. Not only is five a reasonable number of response points for fitting a non-linear IRT graded response model, but evidence also suggests that responses on 5-point response scales are, in most cases, well fitted by linear models (Hofstee, ten Berge & Hendricks, 1998). More specifically, both theoretical (Lord, 1952, 1953) and empirical (Muthén & Kaplan, 1985; Olsson, 1979) evidence suggests that the linear model works well with this type of item when (a) the discriminatory power of the items is moderate or low, and (b) the items have no extreme locations. This is because, in these conditions, the item-trait regressions are essentially linear and homoscedastic for the range of trait values that contains most of the respondents (Ferrando, 2002). Previous analysis based on CTT obtained moderate discriminations and rather symmetrical distributions for most of the CTAC items (Pallero et al. 2006).

The discussion above provides two starting points for this study: (a) both linear and non-linear IRT models are expected to be appropriate for assessing the three critical issues above, and (b) the comparison between the results provided by both approaches is of both substantive and theoretical interest. As for point (b), comparisons between the linear and the nonlinear approaches have already been made in the literature (e.g. Ferrando, 1999; McDonald, 1982, 1999). However, in most cases these comparisons are purely theoretical or have focused on issues such as item estimates or effects on external validity. In contrast, our study aims to make two new contributions. First, when possible, we shall make theoretical predictions that we shall contrast with the empirical data and discuss. Second, our study will focus on the three points discussed above. As far as linear/nonlinear comparisons in terms of conditional precision, sensitivity to change, and person fit is concerned, the present study appears to be new.

In the rest of this section we shall briefly discuss the two models to be compared in the study, and derive the predictions to be contrasted with the empirical results. More specific information will be provided in the Method section.

The linear model used in our study is Spearman's factor analysis (FA) model, usually known as the congeneric model in the psychometric literature (Jöreskog, 1971). In this paper, we shall use the terms linear and congeneric indistinctly. As for the non-linear model we shall consider Samejima's (1969) normal ogive version of the graded response model (GRM). This version (or its virtually indistinguishable logistic counterpart) is the one that is most used in practical applications (Baker, 1992;

Samejima, 1969, 1997). Although the initial conceptualization of the GRM is clearly different from FA modeling, both models can be related by using a general FA formulation based on an underlying variable approach (see Ferrando, 1999, 2002). Essentially, in the linear modeling it is assumed that the congeneric model holds directly for the observed item scores. In the GRM it is assumed that the congeneric model holds for the response variables that underlie the observed scores.

Because the responses to the CTAC items are discrete and bounded, the linear model cannot be strictly true and must be taken as an approximation (Mellenbergh, 1994). On the other hand, the GRM is theoretically more plausible because it correctly treats the item scores as discrete and bounded variables.

As discussed above, given the properties of the CTAC items, the linear model is expected to provide a good approximation in our study. However, even if we accept this point, why should we not use only the theoretically superior GRM? There are reasons not to discard the linear model from the outset. The GRM is a complex model that makes strong assumptions that might not be met. Furthermore, its complexity makes both the calibration and the scoring processes prone to instability.

Overall, in the conditions that we assume 'a priori' for the CTAC case, our starting prediction is that both the linear model and the GRM will lead to very similar results and fit the data equally well. Furthermore, because there is a sizeable number of response points and the samples are not too large, the estimates provided by the simple linear model are likely to be more stable.

We turn now to more specific predictions, and we shall start with those concerning the scoring of the individuals. In this study the chosen scores are the maximum likelihood (ML) individual trait level estimates. In the linear case they can be obtained in the closed form and are the well known Bartlett's factor scores (e.g. McDonald, 1982; Mellenbergh, 1994). In the case of GRM, no closed-form estimator exists, so trait estimates must be obtained iteratively.

Our first prediction regarding individual scores is that the regression of the linear estimates on the GRM estimates will be S-shaped but nonlinearity will only be apparent at the ends of the curve. This prediction is based on the following results. First, the congeneric ML estimate (i.e. Bartlett's score) is a linear combination of the raw item scores. Second, in the GRM the relation between the ML trait estimates and the 'true' trait levels is linear with unit slope. And, if the test is reasonably long, the estimates are related to the 'true' trait levels according to the assumptions of

an error-in-variables model (e.g. Samejima, 1977). So, the regression of the linear estimates on the GRM estimates is expected to have essentially the same shape (i.e. S-shaped) as the regression of the test scores on the true θ levels (possibly with a slight attenuation due to the measurement error). Furthermore, given that the CTAC items are expected to have moderate discrimination and non-extreme locations, the regression is expected to be essentially linear in the trait range that contains most of the respondents.

Our second prediction is that, at both trait ends, the linear estimates will be closer to zero than the GRM estimates and that the latter will have a greater dispersion. The basis for this prediction is as follows. First, finite ML estimates based on the GRM do not exist for totally extreme patterns. Furthermore, the estimates may take very extreme values for near-extreme patterns, particularly when the spread of item locations is relatively small and the item discriminations are high (Kim & Nicewander, 1993). Although these conditions are not expected in CTAC, some appreciable instability at the extremes is still expected. In the linear model, however, because the estimate is a weighted composite of the raw scores, finite estimates exist even for the totally extreme patterns. Furthermore, the changes that occur in the trait estimate as the pattern becomes extreme are gradual: no instability exists for near-extreme patterns.

We shall now discuss the predictions regarding measurement precision and sensitivity to change. As far as the latter point is concerned, the situation we consider here is a repeated-measures design in which the individual is administered the CTAC on two occasions with a retest interval that is long enough to avoid retest effects.

The basic measure to derive predictions in both cases (precision and sensitivity) is the test information, understood as a measure of conditional precision (Mellenbergh, 1996). In both the linear model and the GRM, the amount of information is related to the precision of the ML estimate of the trait level. So, it assesses both the accuracy of our chosen ML scores as estimates of the 'true' trait levels and the sensitivity of these scores for detecting change.

In the linear model the test information does not depend on the trait level. So, the plot of the amount of information, which we shall term the test information curve (TIC), is flat, with constant information throughout the trait range. The amount of constant information depends only on (a) the number of items, and (b) their discriminating power. On the other hand, in the GRM the amount of information is a complex function of the trait level which depends on (a) the number of items, (b) the number of response categories (five in our case), (c) the items' discriminating power, and (d) the

distances between the item locations. The amount of information increases with the number of items, the number of categories, and the discriminating power (Samejima, 1969, chapter 6). However, the impact of determinant (d) is not so clear (Baker, 1992). Therefore, it is very difficult to predict the relations between the TIC provided by the GRM and the constant amount of information provided by the linear model. In both cases, the information increases with the number of items and their discriminating power. However, it is difficult to go any further.

As discussed above, the constant information predicted by the linear model cannot be a correct result, so the information is expected to be approximately constant only for the range of trait values in which the item response function is essentially linear. Therefore, only in this range the estimated precision and measurement of change are expected to be approximately correct As for the GRM, from previous results we can assume that the CTAC item locations are generally well spread and centered around the population mean of θ and that the items' discriminating power is only moderate. If this is so, it follows that the GRM-based TIC should be relatively flat, centered around zero and provide a reasonable amount of information over a wide range of trait values. From this result, two predictions can be made. First, measurement precision and sensitivity to change are expected to be maximum around the zero trait mean. Second, precision and sensitivity to change are expected to be acceptable over a wide range of trait values. Finally we are unable to predict the relation between the amount of information provided by both models, and what we propose is to empirically assess this issue.

Finally we turn to person-fit assessment. Of the various types of parametric person-fit procedures (see e.g. Meijer & Sijtsma 1995, 2001 for reviews) this study focuses on global scalar-valued indices, which assess the extent to which a response pattern is consistent given the chosen model (the linear model or the GRM in our case) and the estimated trait value of the respondent. More specifically, we shall use global indices based on the likelihood function. Like all person-fit indices developed so far, likelihood-based indices have both theoretical (there are approximations) and practical shortcomings (e.g. Magis, Raîche & Béland, 2012). However, they are simple and practical, and perform reasonably well when used as first-step devices for flagging potentially inconsistent respondents (Ferrando, 2007; Meijer & Sijtsma 1995, 2001).

The specific indices we shall use in this study are (a) the polytomous extension of Levine and Rubin's (1979) index (l_{ZGRM}; Drasgow, Levine & Williams, 1985) for the GRM-based analyses, and the *lco* index proposed by

Ferrando (2007) for the congeneric model. The results of both indices are expected to be comparable for three reasons. First, both indices are likelihood-based. Second, they are independent of the trait level, and therefore expected to detect misfitting patterns equally well at all trait levels. Finally, they both refer to a theoretical distribution (l_{ZGRM} standard normal and *lco* chi-square).

In spite of this comparability, however, it is hard to make predictions about the relation between l_{ZGRM} and *lco* due to the approximate nature of both indices. In a substantive study such as the present one, the indices are mainly intended to be used as screening devices for flagging potentially inconsistent respondents. So, in addition to assessing the degree of relation between the indices, we shall also assess whether they both flag mostly the same respondents as inconsistent.

METHOD

Participants. The participants were 670 visually impaired or blind people (39.7% men and 60.3% women; mean age 73.32 years and standard deviation 6.88; ranging from 59 to 92 years). They were all members of ONCE, and met the conditions under which the CTAC is intended to be used: a residual vision of 0.1 or lower on the Weker scale and/or a visual field of 10 degrees or lower. They had no other pathologies. Participants came from different Spanish cities (18.5% Tarragona, 23.6% Barcelona, 12.2% Sevilla, 13.7% Valencia, 13.4% Madrid, 14% other and 4.5% missing data). None of the participants were living in assisted centers. For all the participants, one psychologist per city read them the items and wrote down the answer on a paper and pencil questionnaire.

It is perhaps relevant to note that the CTAC sample is regularly updated, and that the first 350 of this sample of 670 had been used in the previous studies referred to in this paper.

Instruments. The CTAC (Pallero et al., 2006) is made up of 35 items, with a 5-point response format, and, as discussed above, aims to measure the physiological and emotional behaviors that reflect anxiety. Each item has two parts. In the first part the respondent is asked to imagine him/herself in a situation related to their visual deficiency that the researcher explains. In the second part the respondent has to answer the degree of anxiety the imagined situation may evoke nowadays, using a suggested adjective that may refer to emotional or cognitive anxiety. An item example could be:

-"Imagine that you are home alone, you drop a spoon and you can't find it. To what extent do you feel helpless?"

In previous studies on the dimensionality of the CTAC, a pair of items (9 and 27) that differ in the evoked degree of anxiety but which are very similar in both form and content were flagged as problematic. This pair is likely to behave as a locally dependent doublet, thus giving rise to problems of biased estimates and distorted goodness-of-fit results. For this reason item 9 (the least discriminating) was omitted in the present study and all the analyses that follow were based on the remaining 34-item set.

Procedures

Model Estimation and Scoring

Both the congeneric model and the GRM were fitted using an FA approach. The congeneric model was fitted by using a standard FA based on the mean vector and the inter-item covariance matrix. The GRM was fitted by using a factor analytic limited-information estimation procedure based on the bivariate polychoric tables between pairs of item scores. To make the results as comparable as possible, both models were fitted using a robust estimation procedure with mean and variance-corrected goodness-of-fit statistics. In the congeneric model we used robust maximum likelihood estimation. In the GRM we used robust weighted least squares estimation. In both cases the models were estimated using the program Mplus 6.11 (B. Muthén & L.K. Muthén, 2010). Once the models had been fitted and their appropriateness had been assessed (item calibration), the item parameters were taken as fixed and known, and used to obtain ML estimates of the trait level for each individual (individual scoring).

Assessment of Measurement Precision and Sensitivity to Change

Measurement precision was assessed by computing the amount of test information as a function of the trait level. The general expression we used to obtain the expected information, which is applicable to both models (e.g. Kendall & Stuart, 1977) is

$$I(\theta) = -E\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] \quad (1)$$

where *log L* is the log-likelihood for the corresponding response vector according to the model. As mentioned above, the amount of information is

related to the precision of the ML estimate of the trait level. More in detail, as the number of items increases without limit, the standard error of the ML estimate is

$$s.e.(\hat{\theta} \mid \theta) = \left(\frac{1}{I(\theta)} \right)^{1/2} \tag{2}$$

For both models, the information values obtained were next used to plot the TICs and check the predictions discussed above. Finally, the relation between the amount of information provided by both models was assessed by using the concept of relative efficiency (Lord, 1974), which, in our case, is simply the ratio of the amount of information provided by both models, and obtained as a function of θ.

We turn now to the assessment of change. Two procedures were used to consider change as statistically significant. The first one (Speer, 1992; Reise & Haviland, 2005) is approximate but very simple. It consists of (a) setting a confidence band around the test score obtained at Time 1, and (b) considering change as significant if the score at Time 2 is beyond this band. Let $\hat{\theta}_1$ be the ML trait estimate for the individual obtained at Time 1. A 90% confidence band is then computed as

$$\hat{\theta}_1 \pm 1.65 \; s.e.(\hat{\theta}_1 \mid \theta_1) \tag{3}$$

where *s.e.* is the standard error of estimate in (2).

The second, more complete procedure, takes into account that both the Time 1 and Time 2 estimates contain measurement error (Finkelman, Weiss, & Kim-Kang, 2010). Using the same critical value as above, the minimum difference in ML values that is required to consider change as significant is

$$D = 1.65\sqrt{s.e^2.(\hat{\theta}_1 \mid \theta_1) + s.e^2.(\hat{\theta}_2 \mid \theta_2)} \tag{4}$$

Person-Fit Assessment

For each respondent, the l_{ZGRM} and *lco* indices were computed by using the ML trait estimates described above. If we denote by l_{0GRM} the log-

likelihood of a response pattern for which the GRM holds, the l_{ZGRM} index is given by

$$l_{ZGRM} = \frac{l_{0GRM} - E(l_{0GRM})}{\sqrt{Var(l_{0GRM})}} \tag{5}$$

(see Drasgow, Levine & Williams, 1985 for details). If the 'true' trait level were known, the distribution of l_{ZGRM} would be expected to approach the standard normal as the test gets longer (Drasgow, et al., 1985). Because the ML estimate is used instead of the unknown trait level, the reference distribution is only approximate, and more so if the test is short (e.g. Magis et al., 2012).

The *lco* index is given by

$$lco(i) = -2(\log(\mathbf{x}_i \mid \hat{\theta}_i) - \sum_{j}^{n}(\log\frac{1}{\sigma_{\varepsilon j}\sqrt{2\pi}})) \tag{6}$$

where $\hat{\theta}_i$ is the maximum likelihood estimate of respondent's *i* trait level, $\sigma_{\varepsilon j}$ is the residual standard deviation of item *j* and the first term on the right hand side of the equal sign is the log-likelihood for the corresponding response vector ($\mathbf{x}_i$) according to the model. If the congeneric model were correct, and the item parameters were known, the distribution of *lco* would be chi-squared with *n*-1 degrees of freedom (Ferrando, 2007). Because neither of the two conditions is met, the index must also be considered as an approximation.

The relation between both indices will be assessed by plotting and inspecting their joint distribution. In this assessment it must be taken into account that both indices are interpreted in the opposite sense. In the case of l_{ZGRM} small values (i.e. large negative values) are indicators of misfit. In the *lco* case large positive values are indicators of misfit.

RESULTS

Preliminary Analyses

As a first step for judging the adequacy of the models we assessed the discriminating power and the marginal distribution of the item scores. The CTT-based item discriminations ranged from 0.40 to 0.69, with a mean of

0.59. As for the distributions, most of them were unimodal and fairly symmetrical. The item means (in the 1-5 raw scoring) ranged from 2.02 to 3.65 with an average of 2.89, and in none of the distributions was the skewness coefficient larger than one in absolute value (see Muthén & Kaplan, 1985). The kurtosis was negative in all cases (i.e. platykurtic distributions) and ranged from -1.48 to -0.29. Overall, as expected, the CTAC items are characterized by moderate discriminations and distributions that are not too extreme. So, in principle, we consider the data to be amenable for both the linear and the GRM-based analyses. As for the kurtosis, although it is not excessive in any case (see Muthén & Kaplan, 1985), the values obtained justify the use of the robust estimation procedures discussed above.

Item Calibration and Scoring

For both models, the goodness-of-fit results are in table 1. Apart from the chi-square statistic, we used other indices of fit: the RMSEA point estimate and its 90% confidence interval (Browne, & Cudeck, 1993), the comparative fit index (CFI, Bentler, & Bonett, 1980), the gamma-Goodness-of-fit index (GFI, Tanaka & Huba, 1985) and the root mean square of the standardized residuals (RMSR-z).

Table 1. Goodness of fit assessment

Model	χ^2	df	RMSEA	90% C.I.	CFI	GFI	RMSR-z
Linear	1179.41	525	0.043	(0.040;0.046)	0.87	0.95	0.049
GRM	1773.72	525	0.060	(0.057;0.063)	0.93	0.90	0.059

Overall the results in table 1 suggest that, as predicted, the fit is acceptable for both models, and more so if we take into account the size of the model. This statement, however, must be qualified. For the indices used here, reference cut-off values for considering model-data fit as good can be summarized as follows: RMSEA values of less than 0.06 (Hu & Bentler, 1999) or less than 0.08 (Browne & Cudeck, 1993); CFI and GFI values greater than 0.90 (Bentler & Bonnet, 1980; see also the review in Hu & Bentler, 1999) or greater than 0.95 (Hu & Bentler, 1999), and RMSR-z values less than 0.08 (Hu & Bentler, 1999). If we use these references, our

results can be interpreted as follows. First, the absolute fit, mainly based on the assessment of the magnitude of the residuals, can be regarded as acceptable, as indicated by the RMSR-z and RMSEA values. Second, the fit as measured by the amount of explained covariation (GFI), and the fit as defined with respect to the null model of no inter-item relations (CFI), will only be marginally acceptable in some cases. This second result is not due to particular model misspecifications but to the general moderate discrimination characteristic of the CTAC items (see e.g. McDonald, 1999). Because of the moderate inter-item consistency, the amount of explained covariation is not too high (GFI), and there is no dramatic improvement in fit when the prescribed model is used instead of the null model (CFI).

Even when the estimation methods were chosen to make the linear and GRM results as comparable as possible, they are still different. So, comparisons are necessarily descriptive. However, even when this limitation is acknowledged, it seems clear that the linear model fits better in terms of the magnitude of the residuals and the amount of explained covariation, as indicated by the chi-square, RMSEA, RMSR-z and GFI indices. On the other hand, the GRM appears to fit better in relative terms as expressed by the CFI.

Table 2 shows the item parameter estimates obtained from both models: standardized loadings for the linear model, and discriminations and thresholds for the GRM. In addition, the table also shows the GRM item discrimination values that are predicted from the linear approximation (*a'*; see e.g. Ferrando, 2002). If the GRM-based discriminations are compared to their linear predictions, it is clear that they are systematically a little higher, as expected. However, the relation is very high, and the product-moment correlation between *a'* and *a* is r=0.99, a result that reinforces the appropriateness of the linear approximation in this case. Substantively, the results in table 2 agree with those found in previous studies, and show that the CTAC items (a) have moderate discriminations and (b) are generally centered on the trait mean and spread over a wide range of the trait distribution.

We turn now to the individual estimates. The product-moment correlation between both sets of ML estimates was r=0.99. Furthermore, figure 1 shows the scatterplot of both sets of scores, which behaves essentially according to the predictions we made above. First, it is noted that the relation is non-linear and essentially S-shaped, as has also been found in previous empirical studies (Dumenci & Achenbach, 2008). However, the nonlinearity is only noticeable at the ends of the scale, where the floor and ceiling effects mean that the linear estimates are somewhat

squeezed up. The slope is relatively low, and the relation is essentially linear throughout the range of θ that contains most of the respondents.

Table 2. Item parameter estimates from the linear model and the GRM

Item	λ	$a'_{(linear)}$	a	b_1	b_2	b_3	b_4
1	0.53	0.62	0.70	-1.04	0.06	1.01	2.00
2	0.59	0.73	0.81	-0.95	-0.21	0.55	1.54
3	0.56	0.67	0.78	-0.76	0.08	0.85	1.60
4	0.62	0.80	0.90	-1.83	-0.88	-0.15	0.86
5	0.63	0.81	0.94	-1.56	-0.72	-0.04	0.84
6	0.57	0.70	0.79	-1.05	-0.21	0.77	1.60
7	0.54	0.64	0.72	-1.53	-0.73	-0.10	1.00
8	0.60	0.75	0.85	-0.76	0.03	0.78	1.54
10	0.69	0.96	1.06	-1.66	-0.67	0.02	0.99
11	0.50	0.58	0.64	-1.45	-0.37	0.54	1.69
12	0.60	0.75	0.87	-1.80	-1.03	-0.3	0.74
13	0.60	0.76	0.86	-1.46	-0.61	0.10	1.15
14	0.66	0.89	1.02	-0.70	-0.04	0.59	1.40
15	0.70	0.99	1.12	-1.07	-0.44	0.19	0.93
16	0.62	0.79	0.91	-0.73	0.11	0.68	1.51
17	0.59	0.73	0.84	-1.36	-0.49	0.36	1.42
18	0.62	0.79	0.86	-1.24	-0.23	0.58	1.44
19	0.57	0.70	0.80	-2.15	-1.20	-0.45	0.59
20	0.60	0.76	0.84	-0.92	0.02	0.67	1.59
21	0.53	0.63	0.75	-0.02	0.71	1.48	2.33
22	0.54	0.63	0.73	-1.17	-0.40	0.29	1.11
23	0.43	0.47	0.58	0.05	0.92	1.61	2.53
24	0.62	0.79	0.91	-1.58	-0.72	0.04	0.88
25	0.35	0.37	0.42	-1.34	-0.20	0.69	2.16
26	0.61	0.77	0.88	-0.63	0.06	0.60	1.45
27	0.60	0.76	0.84	-1.23	-0.29	0.35	1.20
28	0.47	0.53	0.63	0.02	0.77	1.49	2.36
29	0.54	0.64	0.76	-0.18	0.64	1.44	2.41
30	0.44	0.50	0.54	-2.94	-1.25	0.00	1.67
31	0.68	0.92	1.02	-1.08	-0.20	0.39	1.21
32	0.53	0.62	0.69	-2.18	-1.21	-0.37	0.85
33	0.67	0.89	1.03	-1.13	-0.46	0.08	0.86
34	0.55	0.65	0.80	0.04	0.85	1.48	2.38
35	0.66	0.89	1.02	-0.88	-0.04	0.71	1.46

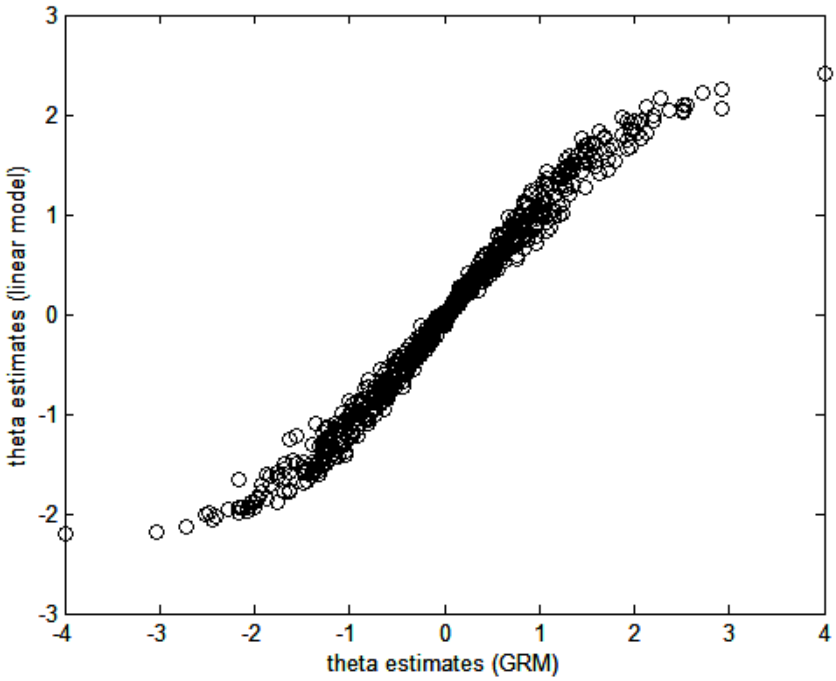

Figure 1. Scatterplot of the ML individual estimates based on the congeneric and the graded response model

The second feature of the graph, which also agrees with the predictions made above, is the slight dispersion of the ML GRM estimates at both ends of the scale which reflects the instability of these estimates for some near-extreme patterns. Although instability is very low, it is still noticeable even in this 34-item test.

Assessment of Measurement Precision and Sensitivity to Change

Figure 2 (a) displays the TIC based on the GRM together with the constant information line obtained from the congeneric model. As predicted, the CTAC scores provide the maximum amount of information around the population mean. Furthermore, the curve is not too peaked, so the scores provide a fair amount of information over a wide interval around the trait mean, as was also predicted. To see this point in more detail, we note that the constant amount of information obtained with the linear model

was 19, which corresponds to a reliability of 0.95 (see Ferrando, 2009). A comparison of both curves shows that, according to the GRM, the CTAC trait estimates are highly accurate (i.e. reliability above 0.95) over a trait interval of about θ=-1.4 to θ=1.4 which is a remarkable result for a test of this type.

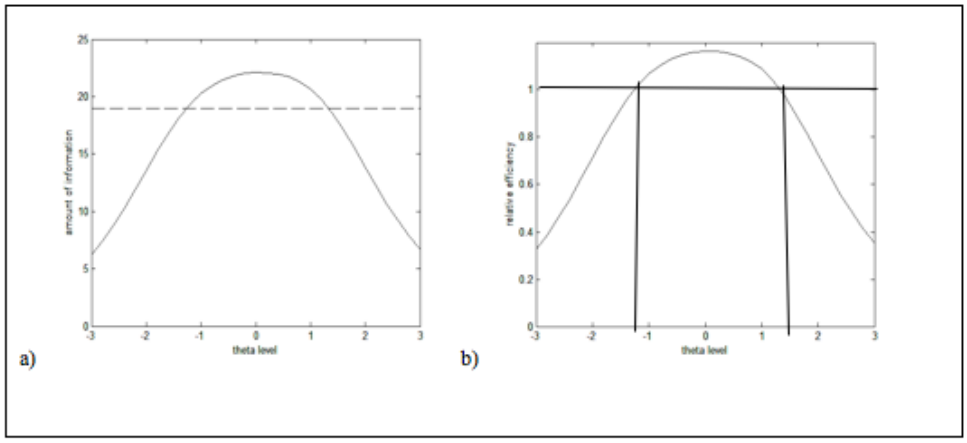

Figure 2. Conditional precision assessed from the linear model and the GRM. (a) TIC; (b) relative efficiency curve

Figure 2 (b) displays the relative efficiency of the GRM with respect to the linear model. For a central trait interval of about θ=-1.2 to θ=1.5, the relative efficiency is greater than one. Outside this interval, the precision falls below that predicted by the linear model. To interpret this result we must again consider that the constant information predicted by the congeneric model (i.e. 19) can only be approximately correct for the interval in which the regression of $\hat{\theta}$ on θ is linear. If we use figure 1 as an approximation, this interval is found to be about (-1.5; 1.5). Now, the average amount of information predicted by the GRM in this interval is 20 (assuming that θ is normally distributed). In our opinion this is a plausible interpretation: the constant information predicted by the congeneric model is interpreted as the average of the 'true' information over the interval in which the linear model is approximately correct.

The discussion above shows that the assessment of the conditional precision is, so far, the only aspect in which the GRM is clearly superior to the congeneric model. Besides, the CTAC proved to be precise over a wide region around the middle of the trait continuum. As mentioned above, this

profile is determined by the combination of (a) moderate discriminations, and (b) a wide spread of item locations which are centered around the trait mean, and would be desirable for a test intended to describe a general population. However, it is perhaps not so desirable for purposes of detection. In this respect, we note that the cut-off value used so far with the CTAC corresponds to a θ value of 0.40 (Pallero et al, 2006). At this point, the GRM-based amount of information is still 21.9, which leads to and *s.e.* of 0.21. Furthermore, in CTT terms this information corresponds to a reliability of 0.96. To sum up: the precision of the CTAC scores around the standard cut-off is still excellent.

We turn now to the issue of sensitivity to change. The constant amount of information estimated in the linear model, corresponds to an *s.e.* of 0.23. Now, by using the first approach in equation (3), it follows that the width of the 90% confidence band for detecting change is $2\times1.65\times0.23=0.76$ at any trait level. However, as discussed above, this result can be only trusted for the central range of θ. If we consider the more plausible GRM predictions, we find that, at the usual cut-off of 0.40, the width of the confidence interval is $2\times1.65\times0.21=0.69$, which is similar. Finally, if we use the more complete approach in (4) the minimum difference in the expected direction (i.e. one tail) required for the change to be considered as significant is 0.49. Substantively, the last two results imply considerable sensitivity. A reduction in the anxiety level in the order of 35% (first approach) or 49% (second approach) of the standard deviation would be detected as significant if it were to be obtained around the usual cut-off point. Indeed, for more extreme levels the sensitivity would be lower.

Person-Fit Assessment

The joint distribution of the l_{ZGRM} and *lco* indices is displayed in figure 3. To help interpret the figure, reference cut-off values of two standard deviations below the mean (l_{ZGRM}) and above the mean (*lco*) are included. These are the most usual cut-off values employed in applied research for flagging a respondent as potentially inconsistent (e.g. Meijer & Sijtsma 1995, 2001).

Several features in figure 3 are worth discussing. First, we note that the relation is negative (as predicted), essentially linear, and clearly heteroscedastic. Thus, the bottom right-hand quadrant, which is where there is most dispersion, contains most of the respondents, those who are regarded as being consistent with both indices. This information is substantively relevant, and suggests that most of the respondents answered the CTAC consistently. The dispersion in the top left-hand quadrant is

considerably lower, and it is here that the respondents that are flagged as inconsistent by both indices are concentrated. We note that the most inconsistent respondents are flagged with one index or the other. Finally, the non-diagonal cells suggest that both indices operate with a different degree of sensitivity. So, *lco* tends to flag many more respondents as inconsistent than l_{ZGRM}. This result is responsible for the relatively low values of agreement based on the resulting 2 ×2 contingency table. The values of the phi coefficient and Pearson's contingency coefficient were 0.48 and 0.43, respectively. It is clear that a higher degree of agreement would have been obtained by maintaining the cut-off value (-2) for l_{ZGRM}. and raising that of *lco*.

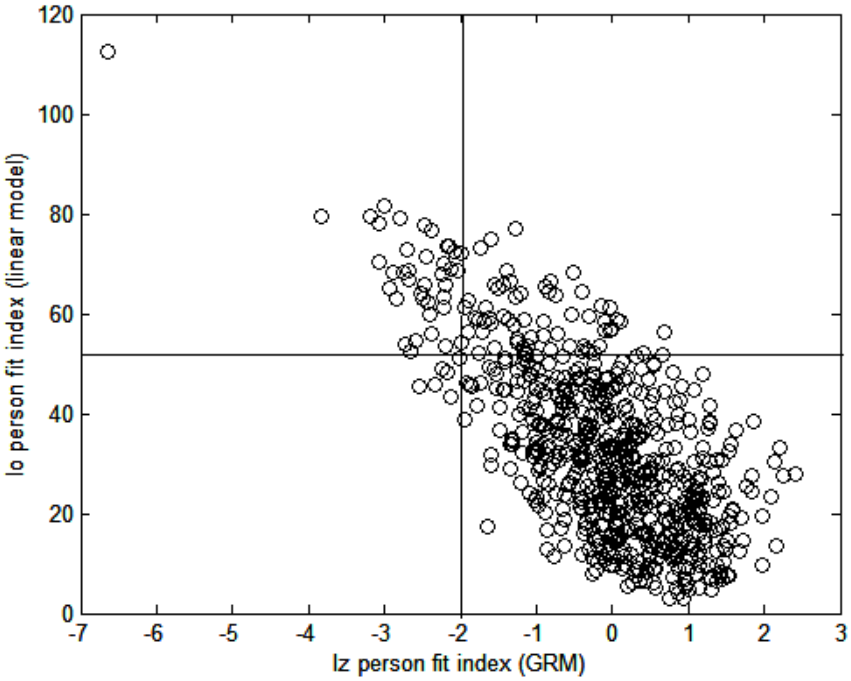

Figure 3. Scatterplot of the person fit indices based on the congeneric and the graded response models

DISCUSSION

We shall organize our discussion of the results into two stages: the calibration stage and the scoring stage.

As for calibration, and as predicted given the characteristics of the CTAC items, both models fitted the data reasonably well. The linear model seemed to provide a better fit in absolute terms whereas the GRM fitted better in relative terms (i.e. relative to the null model of independence). We also believe that the item estimates provided by the simple linear model were more stable in terms of re-sampling or cross-validation. However, we did not assess this issue. The estimates obtained agreed with previous assessments but provided a clearer picture of the test profile: (a) uniformly moderate item discriminations, and (b) item locations that spread over a wide range of the trait distribution and which are centered on the trait mean.

The scoring stage includes the main points of interest of the study. First, for the ML individual estimates, we obtained a near unit correlation between the congeneric and the GRM scores. This result clearly shows that, if the CTAC scores had been used in a validity study, both models would have led to the same results. A more detailed analysis shows that the GRM estimates are more spread at the ends, which is a potential theoretical advantage (greater discrimination). However, this spread is obtained at the expense of some instability, which might lead to over- or under-estimation for some individuals. Overall, we believe that the closed-form estimates provided by the linear model are higher here.

The GRM, however, is clearly better than the linear model when the main interest is to assess conditional precision. The GRM-based TIC is more realistic and shows the regions of θ in which the scores are more or less precise (the central region or extreme values, respectively). This information is useful for both decision purposes and for assessing sensitivity of change, so (a) extreme respondents are assessed with lower precision, and (b) at extreme levels, a larger difference is needed if the change is to be considered statistically significant. The linear model is unable to make these distinctions and the predicted constant precision can only be approximately correct for the trait range in which the test-trait regression is linear. Finally, as far as person-fit assessment is concerned, the relation between both indices is negative (as predicted), essentially linear, and heteroscedastic. And what is more important, the most inconsistent respondents are flagged in both models. Even though the indices appear to function with different degrees of sensitivity, we cannot say that one model works better than the other on this point.

The aim of the study was not to compare the congeneric model and the GRM in general terms, but only in the case of a specific instrument. However, if we try to generalize a little more, it seem reasonable to predict that both models will be appropriate for fitting non-extreme items with

moderate discriminations. If this is the case, and if the aim of the practitioner is to undertake a validity study or to obtain information about individual scores, then the linear model appears to be a good choice. It is simpler, it is likely to produce more stable item estimates, and it will provide more stable trait estimates especially at the extremes. On the other hand if the scores are to be used for making decisions or assessing individual change, the GRM appears to be a better choice. These conclusions might also be useful from an illustrative point of view. For a test such as the CTAC which is based on graded response items, theoretically nonlinear IRT modeling is clearly superior to linear modeling. Our empirical study illustrates what practical contribution these theoretical advantages make. Finally, to close this part of the discussion, it should be mentioned that the linear and non-linear comparisons just discussed are both model based. So, in both cases, results are only interpreted and compared if the model from which they have been obtained provides a good fit of the data. Therefore, when discussing the advantages of the linear model in some issues we are not advocating the use of descriptive linear approaches (mainly CTT) that are not model based and whose appropriateness cannot be assessed.

We turn now to more substantive contributions of the study. Overall, the results are positive. The CTAC is a precise instrument (according to personality standards) that is most precise in the middle of the trait continuum, and which provides substantial information over a wide region around this point. As mentioned above, this profile is more appropriate for a broad-bandwidth test intended for the general population than for a screening test aimed at detecting highly anxious respondents. To better fulfill this last aim, a more peaked TIC with the mode further towards the high end of the trait distribution would have been more appropriate. Even so, the fact that ample information is provided in a wide interval still makes the CTAC quite useful for detection purposes. In particular, the amount of information around the cut-off value used so far with the CTAC is still excellent, no matter which model was used to calculate it. Finally, as far as the person-fit results are concerned, it appears that most respondents answered the CTAC consistently, which is positive. Consistency is a basic requisite for interpreting scores, taking decisions or assessing change.

RESUMEN

Evaluación de la ansiedad en deficientes visuales. Una comparación entre modelos lineales y no lineales de TRI. La presente investigación tiene dos intereses principales. En primer lugar, se evalúan algunas cuestiones pendientes acerca de las propiedades psicométricas del CTAC (cuestionario de ansiedad para ciegos y deficientes visuales) mediante la teoría de respuesta a los items (TRI). En segundo lugar, en las tres propiedades evaluadas: precisión de medida, sensibilidad al cambio e índices de ajuste de la persona, el modelo lineal se compara con el modelo de respuesta graduada (GRM) y los resultados obtenidos sirven además como ilustración de las posibilidades y ventajas de los modelos TRI. Los participantes son 670 ciegos o personas con deficiencia visual de diferentes ciudades españolas. Los resultados mostraron que las puntuaciones del CTAC son suficientemente precisas para el uso al que están destinadas, y que los participantes son generalmente consistentes en sus respuestas. Los datos eran apropiados para los dos modelos puestos a prueba llevando a resultados similares en lo referente a las estimaciones en los niveles en el rasgo, con la excepción de los participantes extremos que fueron mejor evaluados por el modelo lineal. El GRM mostró mayores ventajas en la evaluación de la precisión de la medida y ambos modelos mostraron alta sensibilidad al cambio alrededor del punto de corte. Los resultados concernientes al ajuste de la persona también fueron similares en ambos modelos.

REFERENCES

Baker, F.B. (1992). *Item response theory. Parameter estimation techniques*. New York: Marcel Dekker.

Bauman, M.K. (1963). *Characteristics of Blind and Visually Handicapped People in Profesional, Sales and Managerial Work*. Harrisburg, Pensylvania: Office for the Blind.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models* (pp. 136- 162). Newbury Park, CA: Sage.

Cochrane, G.M., Marella, M., Keeffe, J.E., & Lamoureux, E.L. (2011). The impact of vision impairment for children (IVI_C): Validation of a vision-specific pediatric quality-of-life questionnaire using Rash analysis. *Investigative Ophthalmology & Visual Science, 52*, 1632-1640.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391-404.

Dodds, A.G. (1991). *Psychological assessment and the rehabilitation process*. New Beacon, 75: 101-106.

Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Dumenci, L. & Achenbach, T.M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment, 20,* 55-62.

Ferrando, P.J. (1999). Likert scaling using continuous, censored and graded response models: effects on criterion related validity. *Applied Psychological Measurement, 23,* 161-175.

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

Ferrando, P.J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research , 42*, 481-507.

Ferrando, P.J. (2009). Difficulty, discrimination and information indices in the linear factor-analytic model for continuous responses. *Applied Psychological Measurement, 33*, 9-24.

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

Ferrando, P.J., Lorenzo-Seva, U. y Pallero, R. (2009). Implementación de procedimientos gráficos y analíticos para la construcción de formas paralelas. *Psicothema, 22*, 587-592.

Ferrando, P.J., Pallero, R., Anguiano-Carrasco, C. & Montorio, I. (2010). Evaluación de la sintomatología depresiva en población mayor con pérdida visual: un estudio de la Escala de Depresión Geriátrica. *Psicothema, 22,* 587-592.

Finkelman, Weiss.. & Kim-Kang (2010). Item selection and hypothesis testing for the adaptative measurement of change. *Applied Psychological Measurement, 34,* 238-254.

Fitting, E.A. (1954). *Evaluation of adjustment to blindness*. New York: American Foundation for the Blind.

Gothwal, V.K., Wright, T.A., Lamureux, E.L., & Pesudovs, K. (2009). Rasch analysis of the quality of life and vision function questionnaire. *Optometry and vision science, 86*, E836-E844.

Hardy, R. E. (1968-a). *The anxiety scale for the blind*. New York: American Foundation for the Blind.

Hofstee, W.K.B., Ten Berge, J.M.F., & Hendriks, A.A.J. (1998). How to score questionnaires. *Personality and Individual Differences, 25*, 897-909.

Hu, L.T., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Kendall, M.G. & Stuart, A. (1977). *The advanced theory of statistics* (Vol 2). London: Charles Griffin & Co.

Kim, J. K. & Nicewander, W.A. (1993). Ability estimation for conventional tests. *Psychometrika, 58,* 587-599.

Lamoureux, E.L., Pallant, J.F., Pesudovs, K., Rees, G., Hassell, J.B., & Keeffe, J.E. (2007). The impact of vision impairment questionnaire: An assessment of its domain structure using confirmatory factor analysis and Rasch analysis. *Investigative Ophthalmology & Visual Science, 48,* 1001-1006.

Lazarus, R. (2000). *Estrés y emoción: manejo e implicaciones en nuestra salud*. Bilbao: Desclée de Brouwer.

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

Lord, F.M. (1952). *A theory of test scores*. Psychometrika Monograph. No 7.

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-548.

Lord, F. M. (1974). Quick estimates of the relative efficiency of two tests as a function of ability level. *Journal of Educational Measurement, 11,* 247-254.

Lord, F.M. (1980). *Applications of Item Response Theory*. Hillsdale, New Jersey: LEA.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's lz* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*, 57-81.

McDonald, R.P. (1982). Linear vs. non linear models in Item Response Theory. *Applied Psychological Measurement, 6,* 379-396.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah (NJ): LEA.

Meijer, R.R., & Sijtsma, K. (2001). Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement, 25,* 107-135.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8,* 261–272.

Mellenbergh, G.J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29,* 223-237.

Mellenbergh, G.J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293-299.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171-189.

Muthén, L.K., & Muthén, B. (2010). *Mplus user's guide*. Sixth Edition. Los Angeles: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Pallero, R. Ferrando, P.J. & Lorenzo-Seva, U. (2006). *Cuestionario Tarragona de Ansiedad para Ciegos*. Madrid: ONCE.

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667-696.

Reise, S.P., & Haviland, M.G. (2005). Item Response Theory and the Measurement of Clinical Change. *Journal of Personality Assessment, 84*, 228-238

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

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika, 42*, 163-191.

Samejima, F. (1997). Graded response model. In W.J. van der Linden and R.K. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Speer, D.C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology, 60, 402*-408.

Tanaka, J.S., & Huba, G.J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 38,* 197-201.

Welsh, R.L. (1997). The psychosocial dimensions of orientation and mobility. In B.B. Blasch, W.R. Wiener & R.L. Welsh, (Eds.) *Foundations of orientation and mobility* (pp. 200–227). New York: A.F.B. Press.

World Health Organization (2001) *Clasificación Internacional del Funcionamiento, de la Discapacidad y de la Salud* (CIF), Madrid; IMSERSO.

(Manuscript received: 18 May 2012; accepted: 8 January 2013)