ASSESSING MEASUREMENT INVARIANCE IN QUESTIONNAIRES WITHIN LATENT TRAIT MODELS USING ITEM RESPONSE THEORY

Albert Maydeu-Olivares,
Thomas J. D'Zurilla and
Osvaldo Morera

96-41

WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 96-41
Statistics and Econometrics Series 12
June 1996

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
C/ Madrid, 126-128
28903 Getafe (España)
Fax: (341)-624-98-49

ASSESSING MEASUREMENT INVARIANCE IN QUESTIONNAIRES WITHIN LATENT TRAIT MODELS USING ITEM RESPONSE THEORY

Albert Maydeu-Olivares *
Thomas J. D'Zurilla **
Osvaldo Morera ***

Abstract

Using questionnaires or inventories, researchers often perform mean comparisons between different populations (e.g., males vs. females) in order to draw inferences about actual differences in the constructs being measured. However, such comparisons are not meaningful unless the assessments obtained in each of the populations are conmesurable or invariant across populations. Most researchers simply assume that measuremnt invariance holds. However, the extent to which this assumption is a reasonable one for specific measures and specific populations should be tested empirically. Using item response theory, the present study shows how gender measurement invariance can be determined when, as is most common, a psychological construct is assessed by means of a questionnaire or inventory composed of categorical items. To illustrate our method, the Positive Problem Orientation scale of the Social Problem-Solving Inventory-Revised (D'Zurilla, Nezu & Maydeu-Olivares, 1996) was assessed and found to be reasonably gender invariant, whereas the Negative Problem Orientation scale was not.

Key Words:

Goodness-of-fit, fit plots, categorical data, factor analysis

* Universidad Carlos III de Madrid. Dto. Estadística y Econometría. Universidad Carlos III de Madrid. This research was supported by a Post-doctoral Scholarship from the Ministry of Education and Science of Spain.
** Dept. of Psychology. State University of New York at Stony Brook.
** Dept. of Psychology. University of Illinois at Urbana-Champaign.

ASSESSING MEASUREMENT INVARIANCE IN QUESTIONNAIRES WITHIN LATENT TRAIT MODELS USING ITEM RESPONSE THEORY

Albert Maydeu-Olivares
Dept. Statistics and Econometrics
Universidad Carlos III de Madrid
C/ Madrid 126-128
28903 Getafe (Spain)
E-mail: amaydeu@est-econ.uc3m.es

Thomas J. D'Zurilla
Dept. of Psychology
State University of New York at Stony Brook

Osvaldo Morera
Dept. of Psychology
University of Illinois at Urbana-Champaign

Abstract

Using questionnaires or inventories, researchers often perform mean comparisons between different populations (e.g., males vs. females) in order to draw inferences about actual differences in the constructs being measured. However, such comparisons are not meaningful unless the assessments obtained in each of the populations are conmesurable or invariant across populations. Most researchers simply assume that measuremnt invariance holds. However, the extent to which this assumption is a reasonable one for specific measures and specific populations should be tested empirically. Using item response theory, the present study shows how gender measurement invariance can be determined when, as is most common, a psychological construct is assessed by means of a questionnaire or inventory composed of categorical items. To illustrate our method, the Positive Problem Orientation scale of the Social Problem-Solving Inventory-Revised (D'Zurilla, Nezu & Maydeu-Olivares, 1996) was assessed and found to be reasonably gender invariant, whereas the Negative Problem Orientation scale was not.

Key Words: Goodness-of-fit, fit plots, categorical data, factor analysis

1 Introduction

Most measurements of psychological constructs are performed using multi-item inventories or questionnaires in which it is assumed that the observed interdependencies among the item responses are accounted for by a set of unobserved variables (denoted by common factors or latent traits) representing the psychological constructs being measured. Using questionnaire scores, researchers often perform mean comparisons between different populations (e.g., males vs. females, Japanese vs. Americans, etc.) to draw inferences about actual differences in the psychological constructs measured by these questionnaires or inventories. However, such comparisons are not meaningful unless the assessments

obtained in each of the groups are commensurable or invariant across populations. Within the context of latent trait models, non-comparable measurement exists when the relations between the observed variables and the latent traits differ across populations. When non-commensurability of measurements occurs, observed mean scale differences across populations are meaningless because a different construct is being measured in each of the populations or, in other words, the same construct is measured differently across groups. For example, suppose it is found in a hypothetical depression questionnaire that items reflecting negative cognitive appraisals are more strongly related to the depression construct measured by the questionnaire in women than in men, whereas items reflecting behavioral maladjustments show a stronger relationship with the construct for men than for women. Obviously, men's scores and women's scores could not be compared using this hypothetical questionnaire, since for each gender a different type of depression is being measured.

In this paper, we will first formally define measurement invariance. Then, we will describe the relationship between this concept and the related concepts of test or questionnaire bias and relational equivalence. Next, we will discuss a procedure to assess measurement invariance when a psychological construct is assessed by means of a multi-item inventory or questionnaire. Finally, we will present a practical example of how gender measurement invariance can be assessed using this procedure, focusing on the constructs of positive and negative problem orientation (D'Zurilla, Nezu & Maydeu-Olivares, 1996; Maydeu-Olivares & D'Zurilla, 1995).

2 Measurement invariance and factorial invariance

A formal definition of measurement invariance can be given as follows: Suppose a set of $\underline{n}$ measurements $\underline{y}$, has been obtained on a random sample of subjects. Suppose further that these measurements are a statistical function of another set of $\underline{p}$ random variables $\underline{\theta}$. Now consider a variable indicating group (or population) membership, denoted by $\underline{x}$. We will say that our set of measurements $\underline{y}$ is invariant with respect to $\underline{x}$ if

$$\text{Prob}\ (\mathbf{y} \mid \theta = \mathbf{t}, X = x) = \text{Prob}\ (\mathbf{y} \mid \theta = \mathbf{t}), \qquad \text{for all values of } \underline{x} \qquad (1)$$

that is, if the probability of observing a set of measurements $\underline{y}$ (a set of dependent variables) for a fixed level of the predictors $\theta = \mathbf{t}$, is independent of group membership. In other words, a set of measurements $\underline{y}$ is invariant with respect to $\underline{x}$ if the relationship between $\underline{y}$ and θ, given by Prob $(\mathbf{y} \mid \theta = \mathbf{t})$ is the same regardless of group membership. This is a definition of measurement invariance that has gained widespread consensus (see Meredith, 1993; Millsap & Everson, 1993).

It is important to note that the definition given in Equation 1 is very general. The measurements (dependent variables) $\underline{y}$ and the independent variables θ can be uni or multidimensional, continuous or categorical, and their relationship given by Prob $(\mathbf{y} \mid \theta = \mathbf{t})$ can be linear or nonlinear. For instance, if $\underline{y}$ and $\underline{\theta}$ are single observable continuous variables and Prob $(\mathbf{y} \mid \theta = \mathbf{t})$ is a linear function, then testing the effects of population membership with moderator variables in regression analysis is just a special case of testing for measurement invariance as defined in Equation 1.

In this paper, however, we will concern ourselves solely with the case in which the independent variables θ are unobserved (latent) and the relationship between the dependent

and independent variables is a (linear or non linear) latent trait model. When the model is a linear trait model (i.e., the common factor model) then the term factorial invariance is commonly used in place of measurement invariance.

3 Measurement invariance, test bias, item bias and relational equivalence

When the data to be fitted are the items of a questionnaire or inventory, and it is postulated that a latent trait model underlies the observed responses, the terms test bias or measurement bias are commonly used instead of lack of measurement invariance. That is, a questionnaire is said to be biased when it fails to show measurement invariance across populations. When an instrument is shown to be biased, it may be possible to identify some items in that questionnaire for which measurement invariance holds and some items for which it does not hold. Then, a measurement invariant questionnaire can be obtained by simply removing the items for which measurement invariance does not hold. The items for which measurement invariance does not hold are said to be "biased" or to show differential item functioning (DIF). There is a large literature on identifying single biased items (see, for instance, Thissen, Steinberg & Gerrard, 1986; Thissen, Steinberg & Wainer, 1988, 1993).

When dealing with questionnaire data for which a latent trait model is postulated, it is necessary to investigate two types of measurement invariance (Drasgow, 1984, 1987). The first type of measurement invariance to be assessed consists in examining whether the relationship between the latent trait and the questionnaire items is measurement invariant. This type of measurement invariance is the focus of this paper. The second type of measurement invariance to be investigated consists of examining the relationships between the latent trait (usually estimated using questionnaire scores) and external variables that we wish to predict. To avoid terminological confusions and following the literature (e.g., Reise, Widaman & Pugh, 1993; Drasgow, 1987), we will reserve the term measurement invariance to refer to the former, while we will denote the latter by differential prediction, although both are special cases of measurement invariance as defined by Equation 1.

Differential prediction is generally assessed by regression analysis with group-membership as a moderator variable. Measurement invariance should be assessed before assessing the existence of differential prediction across groups (see Drasgow, 1982, 1984, Drasgow & Kang, 1985) because if a questionnaire is not measurement invariant, it simply cannot be used across populations. On the other hand, if a questionnaire is measurement equivalent but it yields differential predictions, decisions on the criterion variable can still be made provided that predicted criterion scores are used instead of direct questionnaire scores (Drasgow, 1984). As Drasgow (1984) has pointed out: "Regardless of whether a test (or questionnaire) has equivalent relations with a criterion across subpopulations, it seems prudent to require equivalent measurement. Equivalent relations with a criterion despite nonequivalent measurement of the latent trait would lead to the suspicion of nonequivalent measurement of the criterion across subpopulations." (p. 134)

4 Linear vs. non linear latent trait modeling of inventory data

When fitting questionnaire data the most commonly used latent trait model is the common factor model. Assessing measurement invariance in the context of the common factor model (i.e. assessing factorial invariance) requires simultaneous modeling of the means and covariances of the observed variables using multiple group factor analysis. This

simultaneous modeling of means and covariances is necessary because for a measurement to be strictly or strongly factorially invariant across populations (Meredith, 1993: pp. 532-536): a) the matrix of factor loadings must be equal across populations, and b) the mean differences in the observed variables must "all be conveyed through mean differences in the common factors between populations" (Meredith, 1993: p. 535). Descriptions of multiple group linear factor analysis can be found in Bollen (1989), or in Jöreskog and Sörbom (1989).

Unfortunately, if the questionnaire is composed of categorical items (e.g., yes-no questions, Likert items, etc.), multiple group linear factor analysis should not be used in principle to assess measurement invariance because the relationship between a continuous factor and a categorical variable cannot be linear (see McDonald, 1985: pp. 198-223), and therefore, the common factor model is known a priori to be false. In general, if multiple group linear factor analysis fitted by maximum likelihood is used to assess measurement invariance in this common situation, the chi-square goodness-of-fit test will be distorted and the assessment of measurement invariance will be invalid (see Bollen, 1989, pp. 433-439). However, as the number of categories in the items increases (thus resembling a continuous variable), and as the item histograms become more similar to a normal distribution, the assumptions underlying the use of multiple group linear factor analysis might be reasonably met. Conversely, as the number of categories decreases and with increased levels of skewness and kurtosis this method becomes grossly inappropriate [1].

A better alternative for assessing measurement invariance with questionnaire data consists of fitting a nonlinear latent trait model, thus effectively treating the items as categorical variables instead of as approximately continuous variables. In the testing literature, these models are generally referred to as item response models. An item response model is simply a nonlinear factor model in which the relationship between the item and the factor is not assumed to be linear, as in common factor analysis, but follows instead a non-linear shape such as a logistic or a normal ogive curve. Introductory accounts of these models can be found in Thissen and Steinberg (1988), or Hulin, Drasgow and Parsons (1983). See also Reise, Widaman, and Pugh (1993) for a comparison of the linear factor analysis and the item response theory approaches to assessing measurement invariance.

There are two general approaches to estimating item response models (see Mislevy, 1986; Maydeu-Olivares, 1996 a) which can be shown to be equivalent (Takane & de Leeuw, 1987). The first one consists in estimating the nonlinear relation between the items and the latent trait using all the information contained in the pattern of item responses by maximum likelihood estimation. The program MULTILOG (Thissen, 1991) uses this approach to fit the multiple group item response analysis required to test measurement equivalence. The second approach consists in estimating the nonlinear relation between the items and the latent trait using weighted least squares estimation based on measures of pairwise association between item responses (that is, using tetrachoric or polychoric correlations) . The programs LISCOMP (Muthén, 1987), LISREL/PRELIS (Jöreskog. & Sörbom, 1993 a, 1993b), and EQS (Bentler & Wu, 1993) use this approach to fit multiple group item response analysis.

Serious problems arise with both approaches when trying to test measurement invariance because assessing the goodness of fit of item response models is considerably more complicated than assessing the goodness of fit of the common factor model. In this paper, we will discuss only the problems associated with testing measurement invariance by fitting item response models by maximum likelihood. The interested reader may wish to

consult Muthén (1993) for an excellent review of the problems associated with fitting and testing item response models by weighted least squares.

5 Using item response modeling to assess measurement invariance

Very few applications of item response models to psychological research have been reported. In one of them, Waller and Reise (1990) fitted a two-parameter logistic model to the Absorption Scale (Tellegen, 1982). According to the two-parameter logistic model, the probability that a subject with standing t on the latent trait θ endorses item i can be expressed as

$$\text{Prob}\,(u_i = 1 \mid \theta = t) = \frac{1}{1 + \exp[-a_i(t - b_i)]} \qquad (2)$$

where item i is coded $u_i = 1$ for endorsement and $u_i = 0$ for non-endorsement. The probability that the same subject does not endorse the item is then given by

$$\text{Prob}\,(u_i = 0 \mid \theta = t) = 1 - \text{Prob}\,(u_i = 1 \mid \theta = t) \qquad (3)$$

The a_i item discrimination parameter plays a role similar to that of the factor loadings in linear factor analysis, and b_i is a threshold parameter indexing item extremity (see Hulin et al., 1983).

The two-parameter logistic model (Birnbaum, 1968) is an appropriate model for inventories (like the Absorption Scale) whose items only have two options (for instance: yes-no, agree-disagree). When, as is common, an inventory consists of Likert-type items, a model such as Samejima's (1969) graded model can be used instead of the two-parameter logistic model. According to Samejima's graded model, the probability that a subject t would endorse each of the categories of a 5-point Likert-type item i is given by

$$\begin{aligned}
&\text{Prob}\,(u_i = 0 \mid \theta = t) = 1 - \text{Prob}\,(u_i \geq 1 \mid \theta = t) \\
&\text{Prob}\,(u_i = 1 \mid \theta = t) = \text{Prob}\,(u_i \geq 1 \mid \theta = t) - \text{Prob}\,(u_i \geq 2 \mid \theta = t) \\
&\text{Prob}\,(u_i = 2 \mid \theta = t) = \text{Prob}\,(u_i \geq 2 \mid \theta = t) - \text{Prob}\,(u_i \geq 3 \mid \theta = t) \\
&\text{Prob}\,(u_i = 3 \mid \theta = t) = \text{Prob}\,(u_i \geq 3 \mid \theta = t) - \text{Prob}\,(u_i \geq 4 \mid \theta = t) \\
&\text{Prob}\,(u_i = 4 \mid \theta = t) = \text{Prob}\,(u_i \geq 4 \mid \theta = t)
\end{aligned} \qquad (4)$$

where each of the probabilities appearing on the right hand side of Equation 4 is of the type given by Equation 2. Therefore, in Samejima's model each item has one a_i parameter and m-1 threshold parameters b_i, where m is the number of options. Note that if an item only has two categories, then Samejima's graded response model reduces to the two-parameter logistic model described in Equations 2 and 3. As in linear factor analysis, the two-parameter logistic model and Samejima's graded model can be fitted by maximum likelihood estimation procedures assuming a normal distribution of the latent trait being measured.

The functions $\text{Prob}\,(u_i = k \mid \theta = t)$ are called the option response functions (ORFs) of the model. Under the assumption that there is no guessing or similar psychological

phenomenon underlying the subjects' responses that would require option response functions with nonzero lower asymptotes, the two-parameter logistic model and the graded model just described are useful models for fitting binary and Likert items, respectively. Note, however, but there is a vast array of other IRMs that can be used instead. Thissen and Steinberg (1986) provide a useful taxonomy of unidimensional IRMs, that is, IRMs that assume that only one latent trait underlies the observed responses.

6 Assessing measurement invariance in item response models

Within the context of IRMs, the assessment of measurement invariance across populations (for example, across gender) proceeds as follows:

(1) Select an IRM that may be appropriate to the data and fit it separately to each of the populations. If we find that no model fits all the populations closely enough, but instead, we need to use different models to fit different populations, then we have a case of gross measurement invariance.

(2) If we find a model that fits the data in all populations, we shall assess whether the model is measurement invariant across populations. Loosely speaking, this amounts to determining whether the model fits the data if (a) we force the parameters of the model to be equal across populations so that the relationship between each of the items and the latent trait is the same for all populations, and (b) all latent traits have the same variance although they are allowed to have different means, thus capturing all the differences between the observed variables by the difference in latent trait means.

If this model fits satisfactorily the data then we say that the latent trait measured by this inventory is measurement invariant across gender. As mentioned before, the computer program MULTILOG (Thissen, 1991) can be used to assess measurement invariance for unidimensional item response models [2]. This program will fit both the two-parameter logistic and Samejima's graded model to multiple groups.

(3) If the above model does not fit satisfactorily, then we should identify which items are causing the misfit of the measurement invariant model. In other words we should identify which items are biased (or show DIF). This can be done as follows: We fit a measurement invariant model in which the parameters of an item are not constrained to be equal across populations. The difference between the fit of the measurement invariant model and the fit of this model (in which all items are measurement invariant, but the tested item) will give us an indication of the contribution of a single item to the misfit of the measurement invariant model. Hence, for a n item inventory, Step 3 requires performing n separate analyses.

Once the DIF items have been located, they should be removed from the inventory. We then should test whether the inventory formed by the remaining items is measurement invariant repeating Step 2 (and Step 3 if necessary).

7 Assessing the goodness-of-fit in item response models

The major problem faced when assessing measurement invariance of a psychological construct by fitting an item response model is determining the goodness-of-fit of the model to the data. This is important, because results based on poorly fitting models are uninterpretable. Indeed, assessing the goodness-of-fit of IRMs is considerably more difficult than in linear factor models. Since IRMs are models for categorical data, the G^2

and X^2 statistics (see Agresti, 1990) can be used to assess their goodness-of-fit. The MULTILOG program, for instance, provides estimates of the G^2 statistic.

The G^2 statistic compares two nested models (say model A nested in Model B) by taking the ratio of the likelihood of the data under each model. Its general form is given by

$$G^2 = 2 \sum_{\text{all cells}} \hat{c}_A \log \frac{\hat{c}_A}{\hat{c}_B} \qquad (5)$$

where $\hat{c}_A$ is the expected cell frequency in the contingency table under Model A, $\hat{c}_B$ is the expected cell frequency in the contingency table under Model B, and we sum over all cells of the contingency table. In large samples, and if the larger model (in this case model B) is correct, this likelihood ratio statistic is distributed as a chi-square distribution with degrees of freedom equal to the difference of degrees of freedom between the two models.

The G^2 statistic can be used to assess the goodness of fit of an item response model to the data at hand by comparing the fit of the item response model against a more general model, such as a general multinomial model, provided that the contingency table has few empty cells [3]. Item response models are fitted to a contingency table of size m^n, where m = number of options per item, and n = number of items. Thus, a 10-item scale consisting of 5-point Likert-type items contains 5^{10} = 9,765,625 cells. Clearly, these statistics are useless in most psychological applications because we can not collect enough data to fill most cells in such contingency tables [4].

The X^2 statistic present similar problems in these situations. The general form of this statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(c - \hat{c})^2}{\hat{c}} \qquad (6)$$

where $\hat{c}$ is the expected cell frequency in the contingency table under the model, c is the observed cell frequency in the m^n contingency table and we sum over all cells of the contingency table. Like the G^2 statistic, the X^2 statistic follows in large samples a chi-square distribution provided that the contingency table has few empty cells. This is because when m^n is large relative to the sample size, the observed cell frequencies will be very poorly estimated (most of them will be empty), and therefore the chi-square approximation to the distribution of the X^2 or G^2 statistics will not be accurate.

Interestingly, work by Haberman (1977) suggests that the G^2 statistic could be used to assess the fit of the measurement invariant model relative to a non-measurement invariant model (in Step 2). A non measurement invariant model would be fitted as the measurement invariant model except that the a_i and b_i parameters would not be constrained across groups. Clearly, the measurement invariant model is a special case of (it is nested within) the non measurement invariant model. Reise, Widaman and Pugh (1993: p. 559) have suggested using a nested G^2 statistic to assess the relative merits of both models. This nested G^2 statistic is obtained by subtracting the G^2 statistic of the measurement invariant model from the G^2 statistic of the non-measurement invariant model. The resulting statistic is asymptotically distributed as a chi-square with degrees of freedom equal to the difference of degrees of freedom between the two models [5], but only if the non-measurement invariant model fits the data. In other words, the results of this nested G^2 statistic will be correct only if the chosen IRM (for instance Samejima's graded model) without equality

constraints across groups fits the data. Since we have seen that the G^2 statistic can not be used in most instances to check a model against the data, reliance on this nested test as the sole method to determine measurement invariance appears risky.

In Step 3 of the procedure described above we can also use nested G^2 statistics to assess on a one item at a time basis whether the items in the questionnaire or inventory show DIF. In this case, the corresponding nested G^2 statistic is obtained by subtracting the G^2 statistic of the measurement invariant model from the G^2 statistic of a model in which the parameters of all items are constrained to be equal across populations, except the parameters of the item being tested for DIF. The resulting statistic is also distributed in large samples as a chi-square with degrees of freedom equal to the difference of degrees of freedom between the two models, that is, the number of parameters in that item. Again, this test will be correct only if the larger model is an appropriate model for the data.

In summary, nested G^2 statistics can be used in Steps 2 and 3 of the procedure presented above to assess measurement invariance. Since the G^2 statistic is provided in the output of standard software programs such as MULTILOG (Thissen, 1991), Steps 2 and 3 can be performed readily. However, Steps 2 and 3 are meaningless unless we have a way to determine whether the chosen item response model fits the data. In other words, we must seek some procedure to be used in Step 1.

Given the serious difficulties associated with using G^2 or X^2 statistics to assess the goodness-of-fit of item response models when m^n is large relative to the sample size, Drasgow, Levine, Williams, Tsien and Mead (1995) have proposed checking whether the model fits the lower order marginals of the contingency table. At the bare minimum, we should check whether the option response functions match the observed data, that is, whether the model fits the first order marginals of the data. This can be performed, for instance, by graphical methods. The graphical method proposed by Drasgow et al. (1995) to assess the match of the predicted ORFs to the observed data consists in plotting each of the option response functions (ORFs) with 95% confidence intervals around 25 equally spaced points in the latent trait continuum for each of the ORFs. If all ORFs fall within the estimated confidence intervals, that would suggest that the model fits the data. Otherwise, if we observe that in one or more items, the ORFs fall outside the confidence intervals, that indicates that these particular items are not well fitted by the model. The procedure used to draw the fitplots is presented in Appendix A.

The fitplots provide us with a pointwise assessment of the fit of the model. For every option of every item of the questionnaire or inventory model fit is assessed at a set of points in the latent trait continuum, providing us with information about at what levels of the latent trait continuum the misfit is taken place. It is important to realize, however, that the fitplots are more useful in assessing the misfit rather than the fit of the model. This is because if the fitplots show no misfit, this is an indication of the model fitting appropriately the first order marginals of the overall contingency table. However, it may very well be that a model fits satisfactorily the first order marginals and yet does not fit satisfactorily the overall contingency table. In other words, at best, a good fit of the model as assessed by fitplots may be interpreted as an indication of an approximate fit of the model to the data. In this sense, we may refer to the fitplots as being a practical goodness of fit index.

The usefulness and limitations of using a graphical method in helping us to assess measurement invariance in inventory data will be illustrated now by an example.

8 An Application to Gender Studies: Assessing Gender Measurement Invariance in Problem Orientation

The term Problem orientation (D'Zurilla & Nezu, 1982, 1990) is a set of metacognitive processes that reflect a person's general awareness and appraisals of everyday problems, as well as his or her own problem-solving ability (for example, generalized cognitive appraisals, causal attributions, self-efficacy expectancies, outcome expectancies). These generalized beliefs and expectancies are assumed to influence the specific perceptions and appraisals of new problematic situations, as well as the likelihood and efficiency of problem-solving performance in these situations. Recently, Maydeu-Olivares & D'Zurilla (1995, 1996) showed that problem orientation is not a unidimensional construct, but instead, represents two different, albeit related constructs, i.e., positive problem orientation and negative problem orientation. Consequently, we will assess measurement invariance using the procedures described above on separate measures of these two constructs. Positive problem orientation may be described as an adaptive, facilitative, problem-solving cognitive "set," which includes positive problem appraisal (i.e., viewing a problem as a challenge), commitment to a problem-solving coping strategy, problem-solving self-efficacy expectancies, and positive outcome expectancies. In contrast, negative problem orientation consists of maladaptive or disruptive cognitive processes and emotional states, such as negative problem appraisal (i.e., viewing a problem as a threat), self-inefficacy expectancies, negative outcome expectancies, and negative affect (e.g., anxiety, anger, depression). Negative problem orientation is related to psychological distress, such as depression and anxiety, whereas positive problem orientation is related to measures of positive psychological resources or 'wellness' such as optimism, self-esteem, and satisfaction with life (D'Zurilla, Nezu & Maydeu-Olivares, 1996).

In this study we will use item responses from 1043 college students. Of these, 492 were males and 551 females. Two 5-point Likert-type scales were checked for measurement invariance. These scales are the Positive Problem Orientation (PPO: 5 items) and Negative Problem Orientation (NPO: 10 items) scales of the Social Problem Solving Inventory-Revised (SPSI-R: D'Zurilla, Nezu & Maydeu-Olivares, 1996). In this inventory, subjects are asked how they typically think, feel, and behave when faced with problems in everyday living using the following scale {0 = Not at all true of me, 1 = Slightly true of me, 2 = Moderately true of me, 3 = Very true of me, 4 = Extremely true of me}.

We chose the PPO and NPO scales of the SPSI-R to illustrate the assessment of gender measurement invariance because 1) each scale was carefully constructed to be unidimensional and therefore unidimensional item response models are readily applicable to its scales, and 2) we have found gender mean differences in problem orientation on both scales across different samples and age groups (Maydeu-Olivares, D'Zurilla & Kant, 1994). The items composing each of these two scales are provided in Appendix B.

The means and standard deviation on the PPO scale were $\bar{x}$ = 12.42, std = 3.80 for men, and $\bar{x}$ =11.38, std = 3.95 for women. The means and standard deviation on the NPO scale were $\bar{x}$ = 14.60, std = 8.85 for men, and $\bar{x}$ =16.14, std = 9.24 for women. ANOVA analyses revealed significant gender mean differences in both positive and negative problem orientation: $F(1,1041) = 29.582$, $p < .001$ for PPO, and $F(1,1041) = 18.323$, $p < .001$ for NPO. However, do these observed differences in problem orientation reflect real differences between genders or are they merely measurement artifacts caused by differential item functioning across genders?

To answer this question, the parameters of the PPO and NPO items were estimated by maximum likelihood by the MULTILOG (Thissen, 1991) computer program using Samejima's graded model with and without equality constraints across genders. The estimated item parameters are presented in Table 1. In the measurement invariant model,

---

Insert Table 1 about here

---

the item parameters were forced to be equal across gender, the latent traits' variances of both genders were fixed at one, the latent trait means for women were fixed at zero and the latent trait means for men were estimated as -.39 in NPO and .31 in PPO. In the non-measurement invariant model, no constraints on the item parameters were imposed, but the same constraints as above were applied to the latent trait variances and latent trait means. In this case, the latent trait means for men were estimated as -.83 in NPO and .33 in PPO. The standard error of the estimated latent trait means was in all cases .06. Since the latent trait distributions are not equal, the item parameters reported in Table 1 are not directly comparable. We can obtain comparable parameters by performing a suitable transformation on the model parameters and latent trait distribution of one of the populations (see Hulin, Drasgow & Parsons, 1983: p. 26). In our setup, if we apply the following transformation

$$\begin{aligned} \theta^* &= \theta - \mu_\theta \\ b^* &= b - \mu_\theta \end{aligned} \qquad (7)$$

to the sample whose distribution is not standard normal (in this case, the male sample), the distribution of the transformed latent trait, θ^*, will be standard normal, and the transformed thresholds, $\underline{b}^*$, will be comparable to those of the reference distribution. The slope parameters, $\underline{a}$, are directly comparable and need not be transformed. To see that this transformation does not change the model, note that the ORFs for the models considered here depend on the term $\underline{a}\,(\underline{t} - \underline{b})$ in Equation 2. If we apply the transformation given in Equation 7 to this term, we obtain

$$a(t^* - b^*) = a\{(t - \mu_\theta) - (b - \mu_\theta)\} = a(t - b) \qquad (8)$$

and hence the ORFs expressed as functions of the original parameters and the ORFs expressed as functions of the transformed parameters are equivalent.

Uniformly lower thresholds imply higher probability of higher item scores, and thus a higher average scale score. This can be seen in Table 1. For instance, after we transform men's NPO thresholds to make them comparable to women's using Equation 7, that is using $\underline{b}^* = \underline{b} + (.83)$ in the non-measurement invariant model, the women's NPO thresholds are uniformly lower and thus their average scale score will be higher than men's [6].

8.1 Assessment of measurement invariance

The procedure described above was used to assess measurement invariance.

In Step 1, we inspected the fit plots for all PPO and NPO items to assess the practical goodness-of-fit of the non-measurement invariant model [7]. In other words, we assessed whether Samejima's graded model fitted appropriately the men and women's samples separately. The inspection of the fit plots revealed that all the ORFs corresponding

to the PPO were within their estimated 95% confidence intervals. Furthermore, the fitplots indicated that the model fitted somewhat better the male than the female sample. This was also true in the NPO items. However, in this case, the fitplots showed some degree of significant misfit in items 3 and 10 in the female sample. To illustrate the use of fitplots in assessing model fit, we show in Figures 1 and 2 the fitplots corresponding to NPO's item 3. This is the worst fitting item as assessed by the fitplots. The fitplots for men are presented in Figure 1, and the fitplots for women are presented in Figure 2. In both figures, there are five plots for each item, corresponding to each of the five categories of

---

Insert Figures 1 and 2 about here

---

the item. In each of the plots, the horizontal axis is the Negative Problem Orientation latent trait and the vertical axis is the probability of endorsing that particular option given the subject's level on NPO.

As it can be observed in these plots, the probability of endorsing Option 1 decreases as the NPO level increases, whereas the probability of endorsing Option 5 increases at higher levels of NPO. Finally, for Options 2, 3, and 4, the probability of endorsing these options increases up to a point on the NPO scale and then decreases. In these figures, 25 equally spaced 95% confidence intervals have been estimated for each of the options. Note that not all 25 confidence intervals have been drawn for each of the options. In certain instances only the midpoints of the confidence interval (represented by $*$) have been drawn. This is because less than five subjects on that NPO interval chose that particular option. Hence, the confidence intervals would be very poorly estimated and therefore it is safer not to estimate them. Whether the confidence intervals have or have not been drawn help us in interpreting the fitplots. For example, as can be seen in Figure 1, very few confidence intervals have been drawn for options 4 and 5. This indicates that very few men chose these options in NPO's item 3. Furthermore, note that option 4 has been chosen mostly by men with a level on NPO's latent trait between $+1$ and $+2$.

That Samejima's graded model yields a better fit to men's data than to women's data as assessed by the fitplots can be readily seen comparing Figures 1 and 2. In men (Figure 1), the model slightly overpredicts the probability of endorsing option 1 at medium-high levels of NPO and slightly overpredicts the probability of endorsing option 2 at medium levels of NPO. Since in both cases the predicted ORFs are within the estimated confidence intervals, these misfits could be considered of minor importance. In women (Figure 2), the model underpredicts the probability of endorsing option 1 up to a level of about -1 in the NPO latent trait and then it overpredicts it. At very low levels of NPO (below -2) this overprediction lies outside the confidence intervals and therefore it may be considered significant. The model also underpredicts significantly the probability of endorsing option 2 at these low levels of NPO. Finally, the model overpredicts the probability of endorsing option 4 at high levels of NPO, although in this case the misfit is not too gross (it lies within the confidence intervals).

From our inspection of the fitplots corresponding to PPO and NPO we would conclude that Samejima's graded model fits reasonably well PPO's data in both genders (at least its first order marginals), and NPO's data for men, but not for women. In this latter sample, Samejima's graded model may be an appropriate model for all NPO items but items 3 and 10. Hence, some remedial measure should be adopted. For instance, we may (a) remove these items from the inventory, or (b) seek an alternative IRM. Here, for illustrative purposes we shall keep these two items and proceed with Steps 2 and 3.

In Step 2, we computed a nested G^2 statistic to determine whether the measurement invariant model fits significantly worse than the non-measurement invariant model. The values of the G^2 statistic for PPO and NPO under the non-measurement invariant model are 12350.7 and 2275.3, respectively, and under the non-measurement invariant model are 12443.6 and 2306.1, respectively. Therefore G^2_{dif} = 30.8 on 25 d.f., p = .196, for PPO, and G^2_{dif} = 92.9 on 50 d.f., p < .001, for NPO. Given these results, we conclude that measurement invariance holds for PPO but not for NPO. The analysis of PPO is finished, since using the fitplots we have determined that the model fits approximately the data, and using a nested G^2 statistic that measurement invariance holds. For NPO, we should perform Step 3 and try to determine which items show differential item functioning, that is, which items are most responsible for the lack of measurement invariance in NPO. For completeness, we shall also perform Step 3 for the PPO items.

In Step 3, in order to determine whether a particular item showed DIF we fitted n models in which the item parameters were constrained to be equal for all items, except for the parameters of the item being tested for DIF, which were allowed to be different across gender. For instance, we fitted Samejima's graded model to the NPO items, forcing the parameters of all items to be equal across gender, except for the parameters of item 1. This model yielded a G^2 of 12440.1. Subtracting this from the value of the G^2 statistic for the measurement invariant model we can test whether item 1 shows DIF, G^2_{dif} = 12443.6 - 12440.1 = 3.5 on 5 d.f., p = .623. Since allowing the parameters of item 1 to be different across gender does not significantly improve the fit of the measurement invariant model, we conclude that this item does not show DIF. In Table 2 we present the G^2_{dif} statistics of all NPO and PPO items. In order to obtain an overall Type I error of α = .05 in assessing

---

Insert Table 2 about here

---

whether the items of an inventory show DIF, we may use the Bonferroni inequality and divide the overall α by the number of items in the inventory. We shall therefore use α = .05 / 10 = .005 and α = .05 / 5 = .01 for NPO and PPO, respectively, to assess the significance of the G^2_{dif} statistics presented in Table 2. As expected, none of the PPO items show evidence of DIF. As for NPO, the three items that show largest evidence of DIF are items 6, 10, and 3. Of these, only the G^2_{dif} statistic of item 6 is significant at its corresponding alpha level and therefore we conclude that measurement invariance does not hold in this item or in other words, that NPO's item 6 shows DIF. It would have been surprising not to find some degree of DIF in NPO items 3 and 6 since we have seen in Step 1 that these items behave differently in men and women.

We shall now inspect the fitplots of some NPO items to help us understand the strengths and limitations of fitplots in assessing model fit in the context of multiple group item response modeling. These plots will also provide us with a graphical illustration of what is meant by measurement invariance. In Figures 3 and 4 we present the fitplots corresponding to NPO's item 3 under the measurement invariant model.

Insert Figures 3 and 4 about here

These figures illustrate perfectly was is meant by differential item functioning or lack of measurement invariance. The measurement invariant model predicts that both genders will

respond similarly to this item (the ORFs in both figures are identical). However, the empirical proportions with their estimated confidence intervals clearly show that men and women respond very differently to this item, and hence that this item shows DIF. Since we have obtained fitplots for this item under the non-measurement invariant model we know that the DIF revealed by the nested G^2_{dif} statistic for this item is due to the lack of fit of Samejima's graded model to the female sample rather than to imposing constraints across populations. In fact, comparing Figure 3 with Figure 1, and Figure 4 with Figure 2 we see that the measurement invariant model does not fit this item substantially worse than the non-measurement invariant model.

We can use the fitplots corresponding to an NPO item with a high p-value on the G^2_{dif} statistic, for example item 2, as a graphical illustration of measurement invariance. The fitplots corresponding to this item under the measurement invariant model are presented in Figures 5 and 6. These figures clearly show how the measurement invariant model fits very well the empirical proportions in both genders. Furthermore, we can see that the empirical proportions are very similar in both men and women.

Insert Figures 5 and 6 about here

Finally, in Figures 7 and 8 we present the fitplots corresponding to the NPO item with highest DIF, item 6. In these Figures, the ORFs are all within the estimated

---

Insert Figures 7 and 8 about here

---

confidence intervals for the empirical proportions. This could be interpreted as implying that the misfit of the measurement invariant model is not too large. However, notice that although the ORFs do not depart significantly from the empirical proportions, they reveal a type of misfit different from that appearing in Figures 3 and 4. Here there is clearly a consistent bias throughout the latent continuum. For instance, the model overestimates the probability of endorsing option 2 in the male sample up to a level of about -.75 in the NPO latent trait and then underestimates it, or underestimates the probability of endorsing option 3 in the female sample up to a level of -.5 of NPO's latent trait and then underestimates it. Furthermore, the empirical proportions are very different at low levels of NPO, thus suggesting that men and women respond very differently to this item. This is the kind of bias that the G^2_{dif} statistic is set to detect, and the fitplots help us understand the source of differential functioning of this item across gender.

In summary, we have found that Samejima's graded model fits satisfactorily the PPO data (at least its first order marginals) and that measurement invariance across gender holds in this scale. In NPO, two of the items (items 3 and 10) are not well fitted by Samejima's graded model in the female sample, although the misfit is not too large as assessed by the fitplots. Furthermore, NPO is not measurement invariant under this model because men and women respond differently to item 6. When this item is removed from the inventory, NPO is measurement invariant under this model at an alpha level of 1%, since $G^2_{dif} = 65.1$ on 45 d.f., $p = .027$. If we were to remove from the inventory not only item 6 but also items 3 and 10 the resulting scale would be measurement invariant at an alpha level of 5% since in that case we obtain $G^2_{dif} = 49.0$ on 35 d.f., $p = .058$.

According to the measurement invariant model, the mean of men's PPO latent trait is .31 standard deviations higher than women's, whereas the mean of men's NPO latent

trait is .39 standard deviations lower than women's (with a standard error of .06 in both cases). Clearly, these mean differences are significant. Since PPO but not NPO can be shown to be measurement invariant, this difference corresponds to actual differences in level of PPO but not in NPO. However, we have seen that by removing items 3, 6 and 10 we can construct a shortened NPO scale that can be shown to be measurement invariant under Samejima's graded model. The mean of men's NPO latent trait in this shortened scale is .26 standard deviations lower than women's (with a standard error of .06). This mean difference is substantially lower than the one estimated using the full NPO scale, but it is still significant and hence we conclude that there are also actual differences in NPO across gender.

8.2 Discussion

We have been able to show that the mean differences found in positive and negative problem orientation are "real" in the sense that we have ascertained that exactly the same construct is being measured across gender. The finding that women present a more negative orientation than men towards solving their everyday problems deserves very close attention. No gender differences have been found in abstract problem solving (Maccoby & Jacklin, 1974; Kesler, Denney & Whitely, 1976), creativity (Alpaugh & Birren, 1975; Kogan, 1974; Maccoby & Jacklin, 1974), nor in real-life problem-solving skills (Maydeu-Olivares, D'Zurilla & Kant, 1994). Yet, we have seen in this paper that women are more likely than men to: a) view a problem as a threat rather than as a challenge, b) show less positive outcome expectancies and more negative outcome expectancies, c) show less problem-solving self-efficacy, and d) present maladaptive or disruptive cognitive processes and emotional states, such as negative causal attributions (e.g., blaming oneself for problems), and negative affect (e.g., anxiety, anger, depression). We believe that finding an appropriate explanation for this phenomenon will help us understand the cognitive and emotional processes that might be partially responsible for the higher incidence of depression and other forms of psychological distress in women.

9 Concluding remarks

Measurement invariance should be investigated whenever differential item functioning across populations is suspected, and not only in those instances where mean differences across populations are found. In this respect, Thissen, Steinberg and Gerrard (1986) provide a hypothetical example where measurement invariance does not hold despite the absence of mean group differences. Since most psychological constructs are measured by questionnaires composed of categorical items, the assessment of measurement invariance is likely to require the fit of multiple group item response models. This can be accomplished by the use of commercially available software. Measurement invariance can then be assessed by performing nested tests comparing the fit of measurement invariant vs. non-measurement invariant items. Before performing nested tests it is necessary to test whether the selected model fits the data. However, existing statistics to assess the goodness-of-fit of item response models models require samples much larger than those found in most psychological research. Here we have proposed using a practical goodness-of-fit index, namely, the inspection of confidence intervals constructed for each of the option response functions under consideration, in helping us assessing measurement invariance. It is crucial to use some measure of model fit to the data because a nested test

may fail to indicate lack of measurement invariance if most of the differences between option response functions across populations are omitted from the nested test because the option response functions do not capture the data in one of the populations in the first place!.

It is important to note that measurement invariance is model dependent. For instance, in this paper we have shown how one scale (PPO) is measurement invariant under Samejima's graded model, and how some items need to be removed from another scale (NPO) for this to be measurement invariant under the same model. However, it is possible that none of these scales is measurement invariant under an alternative item response model for Likert-type data, say Masters' (1982) partial credit model. Conversely, it is possible that under another item response model, say Bock's (1972) nominal model, both scales are measurement invariant (meaning that no items need to be removed from NPO). Furthermore, we could also use a mixed model in which some items are fitted using one IRM and other items using an alternative IRM. For instance, we could fit all items of NPO but items 3,6 and 10 using Samejima's graded model and these three items using Bock's nominal model, and check whether NPO is measurement invariant under this combined model using the procedures described in this paper. Here we use Samejima's model to fit these data because elsewhere we have shown (Maydeu-Olivares, 1996 a) that these data sets are better fitted by Samejima's graded model than by any other unidimensional parametric IRM including Bock's and Masters', and that only non-parametric IRMs outperform Samejima's model (for a discussion of parametric vs. non-parametric item response models see Maydeu-Olivares, 1994).

References

Agresti, A. (1990). Categorical data analysis. New York: Wiley.

Alpaugh, P.K., & Birren, J.E. (1975). Are there sex differences in creativity across the adult life span?. Human Development, 18, 461-465.

Beck, A.T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. Journal of Consulting and Clinical Psychology, 42, 861-865.

Bentler, P.M. & Wu, E.J.C. (1993). EQS/Windows User's Guide, Version 4. Los Angeles, CA: BMDP Software, Inc.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick [Eds.] Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37 (1), 29-51.

Bollen, K.A. (1989). Structural equations with latent variables. New York: Wiley.

Chang, E.C., D'Zurilla, T.J., & Maydeu-Olivares, A. (1994). Assessing the dimensionality of optimism and pessimism using a multimeasure approach. Cognitive Therapy and Research, 18, 143-160.

Drasgow, F. (1982). Biased test items and differential validity. Psychological Bulletin, 92, 526-531.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement invariance and equivalent relations with external variables are central issues. Psychological Bulletin, 95, 134-135.

Drasgow, F. (1987). Study of the measurement bias of two standarized psychological tests. Journal of Applied Psychology, 72, 19-29.

Drasgow, F., & Kang, T. (1985). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. Journal of Applied Psychology, 69, 498-508.

Drasgow, F., Levine, M.V., Tsien, S., Williams, B., & Mead, A.D. (1995). Fitting polychotomous item response theory models to multiple-choice tests. Applied Psychological Measurement, 19, 143-165.

D'Zurilla, T.J. (1986). Problem-solving therapy: A social competence approach to clinical intervention. New York: Springer.

D'Zurilla, T.J., & Nezu, A.M. (1990). Development and preliminary evaluation of the Social Problem-Solving Inventory (SPSI). Psychological Assessment. A Journal of Consulting and Clinical Psychology, 2, 156-163.

D'Zurilla, T.J., Nezu, A.M. & Maydeu-Olivares, A. (1996). Manual of the Social Problem-Solving Inventory-Revised. Dept. of Psychology. State University of New York at Stony Brook.

Haberman, J.S. (1977). Log-linear models and frequency tables with small expected cell counts. Annals of Statistics, 5, 1148-1169.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory: Applications to psychological measurement. Homewood: Dow Jones-Irwin.

Jöreskog, K.G., & Sörbom, D. (1989) LISREL 7: A guide to the program and applications. Chicago, IL: SPSS Inc.

Jöreskog, K.G. & Sörbom, D. (1993 a). PRELIS 2. User's reference guide. Chicago, IL: Scientific Software.

Jöreskog, K.G. & Sörbom, D. (1993 b). LISREL 8. User's reference guide. Chicago, IL: Scientific Software.

Kesler, M.S., Denney, N.W., & Whitely, S.E. (1976). Factors influencing problem solving in middle-aged and elderly adults. Human Development, 19, 310-320.

Kogan, N. (1974). Creativity and sex differences. Journal of Creative Behavior, 8, 1-14.

Maccoby, E.E., & Jacklin, C.N. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Maydeu-Olivares, A. (1994). Parametric vs. non-parametric approaches to individual differences scaling. Psicothema, 6, 297-310.

Maydeu-Olivares, A. (1996 a). Modelos multidimensionales de respuesta a los items [Multidimensional item response models]. In J. Muñiz (Ed.). Psicometría [Psychometrics]. Madrid: Universitas.

Maydeu-Olivares, A. (1996 b). Fitting unidimensional item response models to actual Likert-type data. Manuscript submitted for publication. Dept. of Statistics and Econometrics. Universidad Carlos III de Madrid.

Maydeu-Olivares, A., & D'Zurilla, T.J. (1995). A factor analysis of the Social Problem-Solving Inventory using polychoric correlations. European Journal of Psychological Assessment, 11, 98-107.

Maydeu-Olivares, A., & D'Zurilla, T.J. (1996). A factor analytic study of the Social Problem-Solving Inventory: An integration of theory and data. Cognitive Therapy and Research, 20, 115-133.

Maydeu-Olivares, A., D'Zurilla, T.J., & Kant, G.L. (1994, May 6). Gender and age differences in social problem solving in college students, middle-aged, and elderly adults. Paper presented at the 66th Annual Meeting of the Midwestern Psychological Association, Chicago (IL).

McDonald, R.P. (1985). Factor analysis and related methods. Hillsdale, NJ: Lawrence Erlbaum.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58, 525-543.

Millsap, R.E., & Everson, H.T. (1993). Statistical approaches for measuring test bias. Applied Psychological Measurement, 17, 297-334.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Muthén, B. (1987). LISCOMP: Analysis of linear structural equations using a comprehensive measurement model. Mooresville, IN: Scientific Software.

Muthén, B. (1993). Goodness of fit with categorical and other non normal variables. In K.A. Bollen & J.S. Long [Eds.] Testing structural equation models. Newbury Park, CA: Sage.

Nolen-Hoeksema, S. (1990). Sex differences in depression. Stanford, CA: Stanford University Press.

Reise, S., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114, 552-566.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, No. 17.

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52, 393-408.

Tellegen, A. (1982). Brief manual for the Multidimensional Psychological Questionnaire. Unpublished manuscript. University of Minnesota, Minneapolis.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using Item Response Theory (version 6). Mooresville, IN: Scientific Software.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 567-577.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. Psychological Bulletin, 104, 385-395.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun [Eds.] Test validity. Hillsdale, N.J: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer [Eds.] Differential item functioning. Hillsdale, N.J: Erlbaum.

Waller, N.G., & Reise, S.P. (1990). Computerized adaptive psychological assessment. Journal of Psychological and Social Psychology, 57, 1051-1058.

Williams, B. (1992). IOCCDRAW [Computer program]. Champaign, IL: Model Based Measurement Laboratory. Dept. of Educational Psychology. University of Illinois.

Footnotes

[1] For example, the dichotomous items of the Hopelessness Scale (Beck, Weissman, Lester, & Trexler, 1974) generally show a very skewed distribution in non-clinical populations since very few subjects in these populations present the hopelessnesss symptoms measured by this inventory. Chang, D'Zurilla and Maydeu-Olivares (1994), analyzed the responses of a sample of college students to the items of this inventory using linear factor analysis and an item response model and showed that the conclusions may be radically different depending on the method being used.

[2] For detailed instructions, see Example 17 in Thissen (1991), and Thissen, Steinberg and Wainer (1993).

[3] The number of degrees of freedom in a multiple group item response model is

d.f. = (# groups * # response patterns) - (# groups) - (# parameters estimated) =
= [# groups * (# categories per item)$^{\#\ items}$] - (# groups) - (sum of distinct $\underline{a}$ and $\underline{b}$ parameters) - (# of estimated group means)

[4] There are rare instances where these statistics can indeed be used for assessing the fit of the model to the data. For example, if a test consists of 5 items consisting each of two categories (e.g.: yes-no, agree-disagree), then the size of the contingency table is $2^5 = 32$ cells, and the $\underline{G}^2$ statistic could be used with moderately large sample sizes.

[5] The number of degrees of freedom will be equal to the number of items times the number of parameters per item. For instance, if we are fitting the two parameter logistic model to a ten item test, the number of degrees of freedom of this nested test will be twenty.

[6] The transformed parameters will be equal to those obtained if each of the samples had been estimated separately since in single group analyses MULTILOG fixes the latent trait mean to zero and the latent trait variance to unity.

[7] The fitplots were drawn using IOCCDRAW (Williams, 1992). All fitplots are drawn in reference to a standard normal distribution, that is, after transforming the item parameters using Equation 7.

Table 1
Item parameters estimated by maximum likelihood using Samejima's graded model

Negative Problem Orientation

	non-measurement invariant model										measurement invariant model				
	men					women					men and women				
item	a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4
1	1.43	-2.03	-0.76	0.44	1.44	1.43	-1.59	-0.29	0.84	2.11	1.39	-1.63	-0.31	0.88	2.06
2	1.55	-1.20	-0.05	0.93	1.89	1.55	-0.81	0.34	1.34	2.56	1.52	-0.80	0.37	1.38	2.51
3	1.92	-1.61	-0.56	0.41	1.52	1.80	-1.47	-0.26	0.79	1.92	1.86	-1.32	-0.19	0.83	1.95
4	1.69	-1.89	-0.70	0.23	1.33	1.51	-1.86	-0.39	0.62	1.88	1.58	-1.66	-0.32	0.66	1.85
5	1.55	-2.20	-0.73	0.37	1.51	1.54	-1.47	-0.12	0.86	2.13	1.51	-1.64	-0.20	0.85	2.09
6	1.82	-2.11	-0.93	-0.05	1.11	1.84	-1.99	-0.76	0.16	1.37	1.77	-1.85	-0.64	0.28	1.48
7	2.29	-1.43	-0.46	0.35	1.19	2.11	-0.96	0.18	1.02	1.98	2.08	-0.99	0.10	0.95	1.88
8	1.38	-1.84	-0.63	0.18	1.18	1.42	-1.39	-0.01	0.86	1.92	1.34	-1.43	-0.09	0.78	1.85
9	2.24	-1.20	-0.33	0.41	1.34	2.13	-0.69	0.21	0.97	1.95	2.15	-0.73	0.17	0.93	1.91
10	2.15	-1.68	-0.62	0.31	1.37	2.39	-1.21	0.06	0.79	1.79	2.28	-1.24	-0.06	0.78	1.81

Positive Problem Orientation

	non-measurement invariant model										measurement invariant model				
	men					women					men and women				
item	a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4	a	b_1	b_2	b_3	b_4
1	1.62	-2.40	-1.25	-0.13	1.44	1.76	-2.47	-1.28	-0.07	1.56	1.67	-2.42	-1.24	-0.04	1.57
2	1.62	-2.66	-1.84	-0.75	0.84	1.70	-2.78	-1.45	-0.51	0.93	1.68	-2.68	-1.55	-0.56	0.94
3	1.39	-1.95	-0.53	0.71	2.11	1.50	-1.81	-0.54	0.57	2.04	1.40	-1.86	-0.49	0.70	2.18
4	1.90	-2.45	-1.35	-0.29	1.19	1.77	-2.60	-1.17	-0.04	1.42	1.87	-2.47	-1.19	-0.10	1.35
5	1.21	-1.97	-0.59	0.72	2.11	1.65	-1.49	-0.26	0.73	1.86	1.43	-1.62	-0.33	0.79	2.03

Notes: The a's are slope parameters, the b's are threshold parameters. Every item has one a parameter and m-1 b parameters (m = # options per item). The probability of endorsing each option given the model is obtained by substituting these item parameters into Equation 3.

Table 2
Differential item functioning (DIF) assessed by G^2_{dif} statistics

Negative Problem Orientation			Positive Problem Orientation		
item	G^2_{dif}	p-value	item	G^2_{dif}	p-value
1	3.5	.623	1	6.3	.278
2	2.3	.806	2	7.3	.199
3	9.9	.078	3	6.3	.278
4	8.2	.146	4	5.4	.369
5	9.4	.094	5	7.0	.221
6	21.5	.001			
7	7.6	.180			
8	6.9	.228			
9	1.7	.889			
10	12.5	.029			

Appendix A

A description of the procedure used to draw fitplots (Drasgow et al., 1995)

Assume N observations have been collected on n polychotomous items each with m categories. Fitplots are constructed in reference to a standard normal distribution. In models in which the latent trait does not follow a standard normal distribution, the model parameters are suitably transformed so that the distribution of the transformed model is standard normal.

Then, the fitplots are obtained as follows:

(1) Divide the latent trait continuum in p intervals S_l of equal width with midpoints t_l, $l = 1, \ldots, p$. Here we have used 25 intervals with midpoints given by the 2nd, 6th, ..., 98th percentile points of the standard normal distribution.

(2) Prior to collecting the data, the models we have considered here assume that the latent trait θ is distributed in the population as a standard normal density. After the data is observed, the probability that a randomly drawn subject from this population with response pattern $\mathbf{u}^*$ has standing t_l in the latent trait is

$$\text{Prob}\,(\theta = t_l \,|\, \mathbf{u} = \mathbf{u}^*) = \frac{\text{Prob}\,(\mathbf{u} = \mathbf{u}^* \,|\, \theta = t_l)\ \text{Prob}\,(\theta \in S_l)}{\sum_{l=1}^{p} \text{Prob}\,(\mathbf{u} = \mathbf{u}^* \,|\, \theta = t_l)\ \text{Prob}\,(\theta \in S_l)} \qquad (9)$$

where $\text{Prob}\,(\theta \in S_l)$ is the area corresponding to S_l under a standard normal distribution,

$$\text{Prob}\,(\mathbf{u} = \mathbf{u}^* | \theta = t_l) = \prod_{i=1}^{n} \text{Prob}\,(u_i = k | \theta = t_l) \qquad (10)$$

and $\text{Prob}\,(u_i = k | \theta = t_l)$ is the ORF evaluated at $\theta = t_l$.

(3) Using the posterior distribution of each respondent evaluated at p intervals, given by Equation 9, compute the proportion of respondents allocated to each interval (the empirical proportions) in item i and option k, $\hat{P}_{ik}(t_l)$, by

$$\hat{P}_{ik}(t_l) = \frac{N_{ik}}{N} \frac{\sum_{j:u_i=k}^{N} \text{Prob}\,(\theta = t_l \,|\, \mathbf{u} = \mathbf{u}_j^*) / N_{ik}}{\sum_{j=1} \text{Prob}\,(\theta = t_l \,|\, \mathbf{u} = \mathbf{u}_j^*) / N} \qquad (11)$$

where N_{ik} is the number of respondents who chose option k in item i. The summation in the numerator is over the respondents who chose option k in item i, whereas the summation in the denominator is over all subjects in the sample.

(4) Estimate approximate 95% confidence intervals for each of the empirical proportions using

$$\hat{P}_{ik}(t_l) \pm 2\sqrt{\frac{\hat{P}_{ik}(t_l)\left[1 - \hat{P}_{ik}(t_l)\right]}{N_{ik}}} \qquad (12)$$

Whenever the sum of the posterior densities in item i and option k is less than five, the confidence interval for $\hat{P}_{ik}(t_l)$ is not drawn.

(5) Draw ORFs along with empirical proportions and confidence intervals for the empirical proportions.

Appendix B

Item content of the Positive and Negative Problem Orientation Scales of the Social Problem Solving Inventory-Revised (D'Zurilla, Nezu & Maydeu-Olivares, 1996)

Positive Problem Orientation

1. When my first efforts to solve a problem fail, I usually think that if I persist and do not give up too easily, I will be able to find a good solution eventually.
2. When I have a problem, I usually believe that there is a solution for it.
3. I usually confront my problems "head on," instead of trying to avoid them.
4. When I am faced with a difficult problem, I usually believe that I will be able to solve the problem on my own if I try hard enough.
5. When I have a problem, I usually try to see it as a challenge, or opportunity to benefit in some positive way from having the problem.

Negative Problem Orientation

1. I spend too much time worrying about my problems instead of trying to solve them.
2. I usually feel threatened and afraid when I have an important problem to solve.
3. I usually feel nervous and unsure of myself when I have an important decision to make.
4. When my first efforts to solve a problem fail, I get very angry and frustrated.
5. When I am faced with a difficult problem, I often doubt that I will be able to solve it on my own no matter how hard I try.
6. Difficult problems make me very upset.
7. When I am attempting to solve a problem, I often get so upset that I cannot think clearly.
8. I hate having to solve the problems that occur in my life.
9. I often become depressed and immobilized when I have an important problem to solve.
10. When my first efforts to solve a problem fail, I tend to get discouraged and depressed.

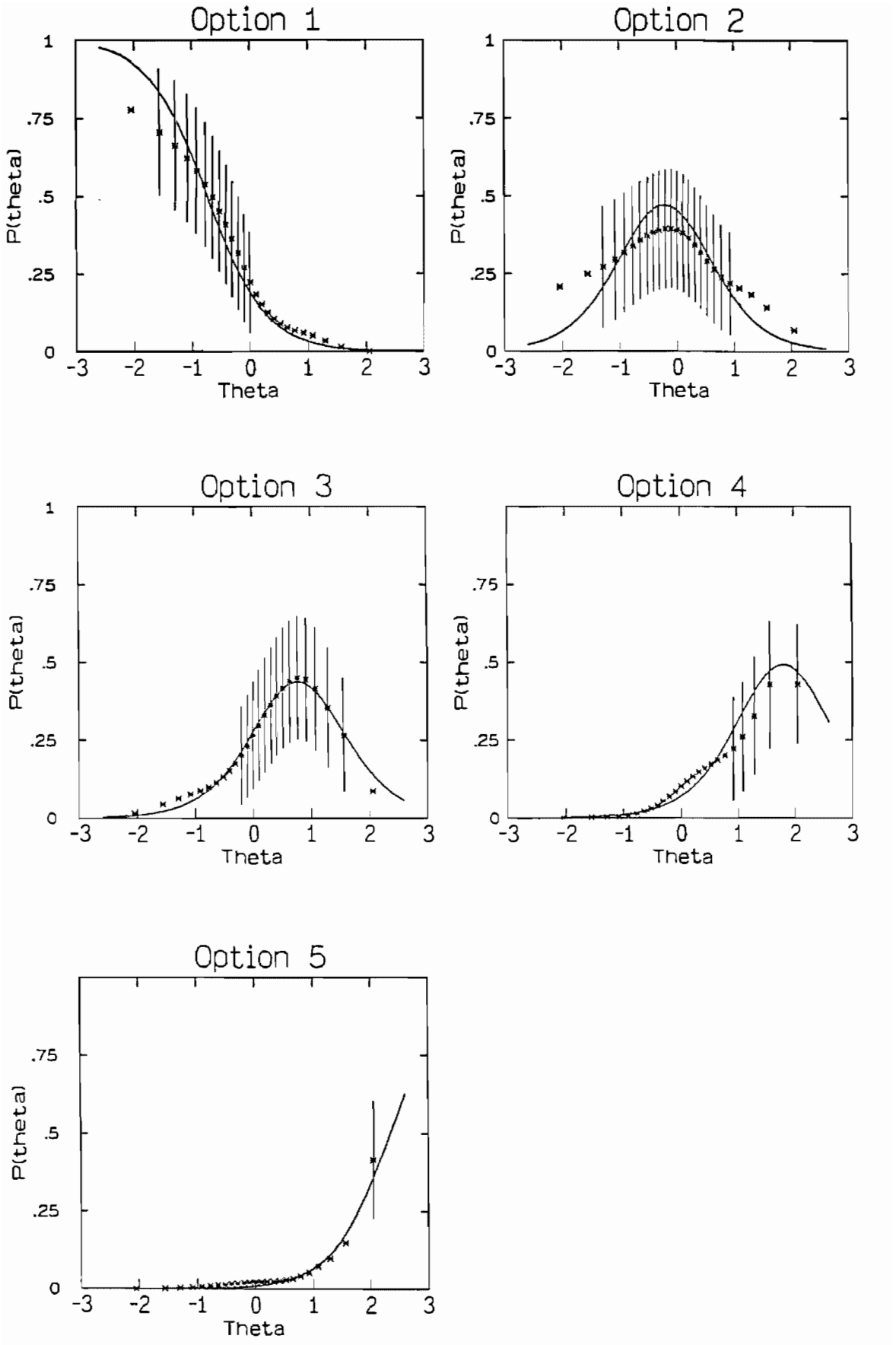

Figure 1. Fitplots of item 3 of the Negative Problem Orientation scale in the male sample according to the non-measurement invariant model.

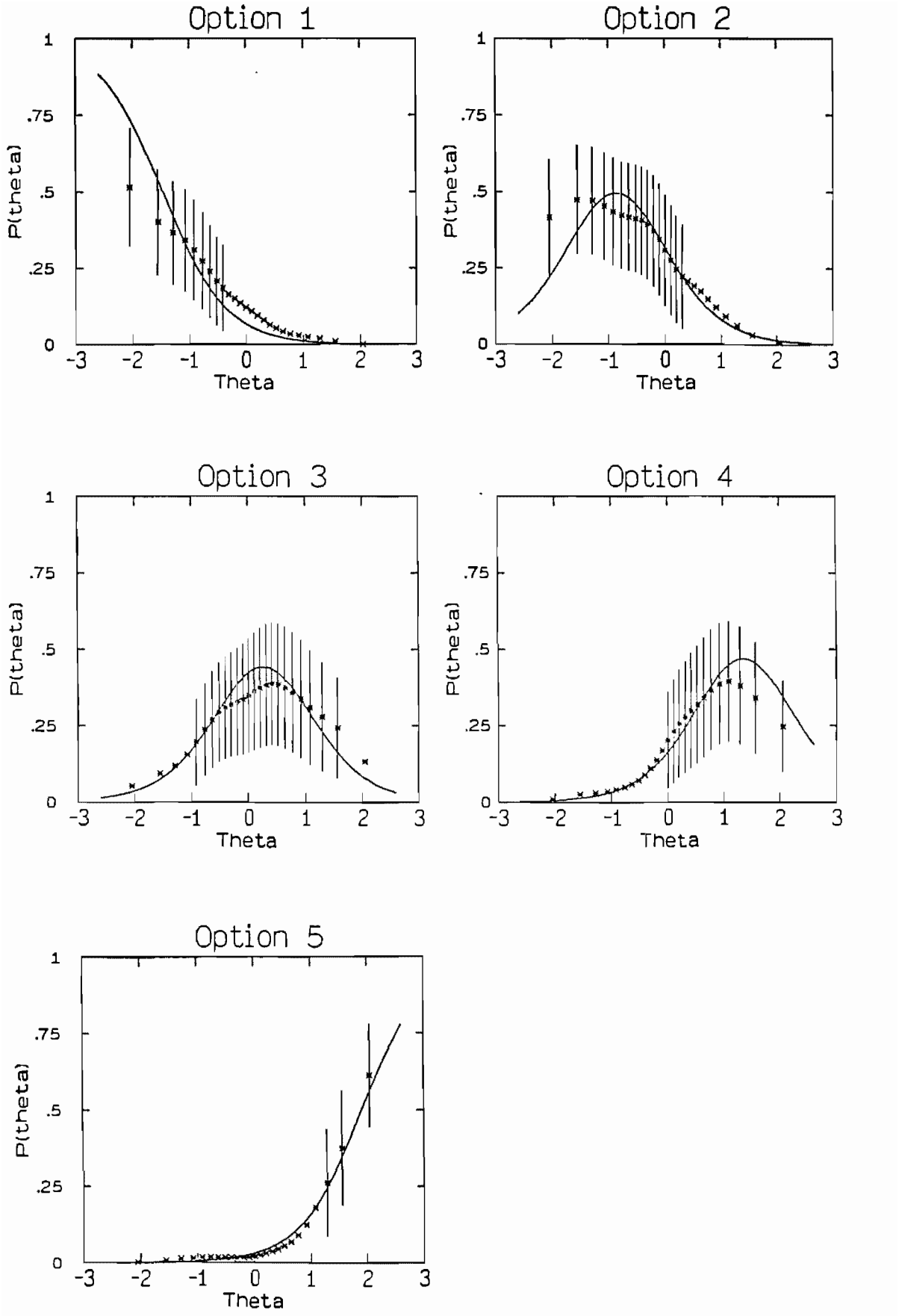

Figure 2. Fitplots of item 3 of the Negative Problem Orientation scale in the female sample according to the non-measurement invariant model.

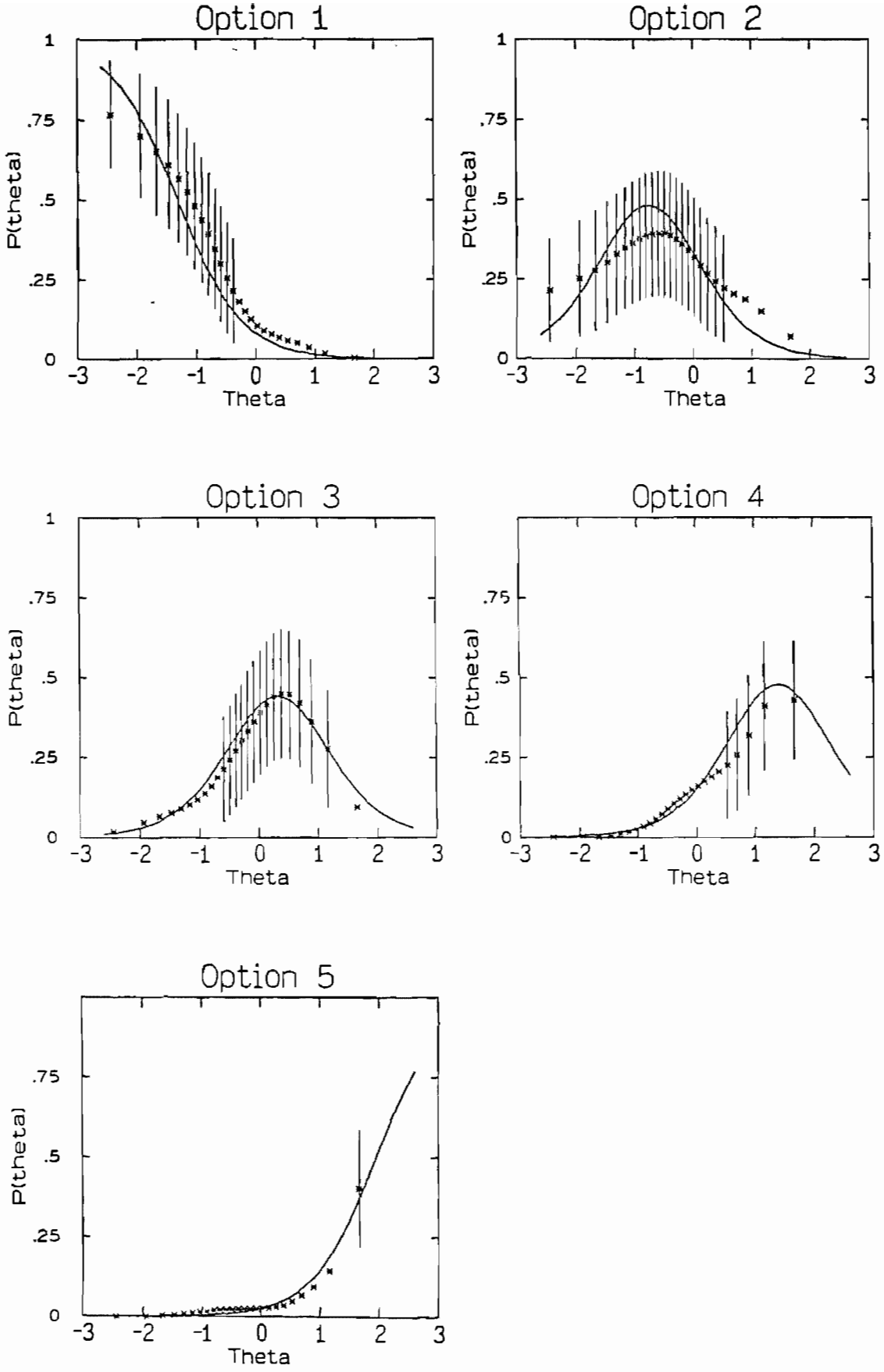

Figure 3. Fitplots of item 3 of the Negative Problem Orientation scale in the male sample according to the measurement invariant model.

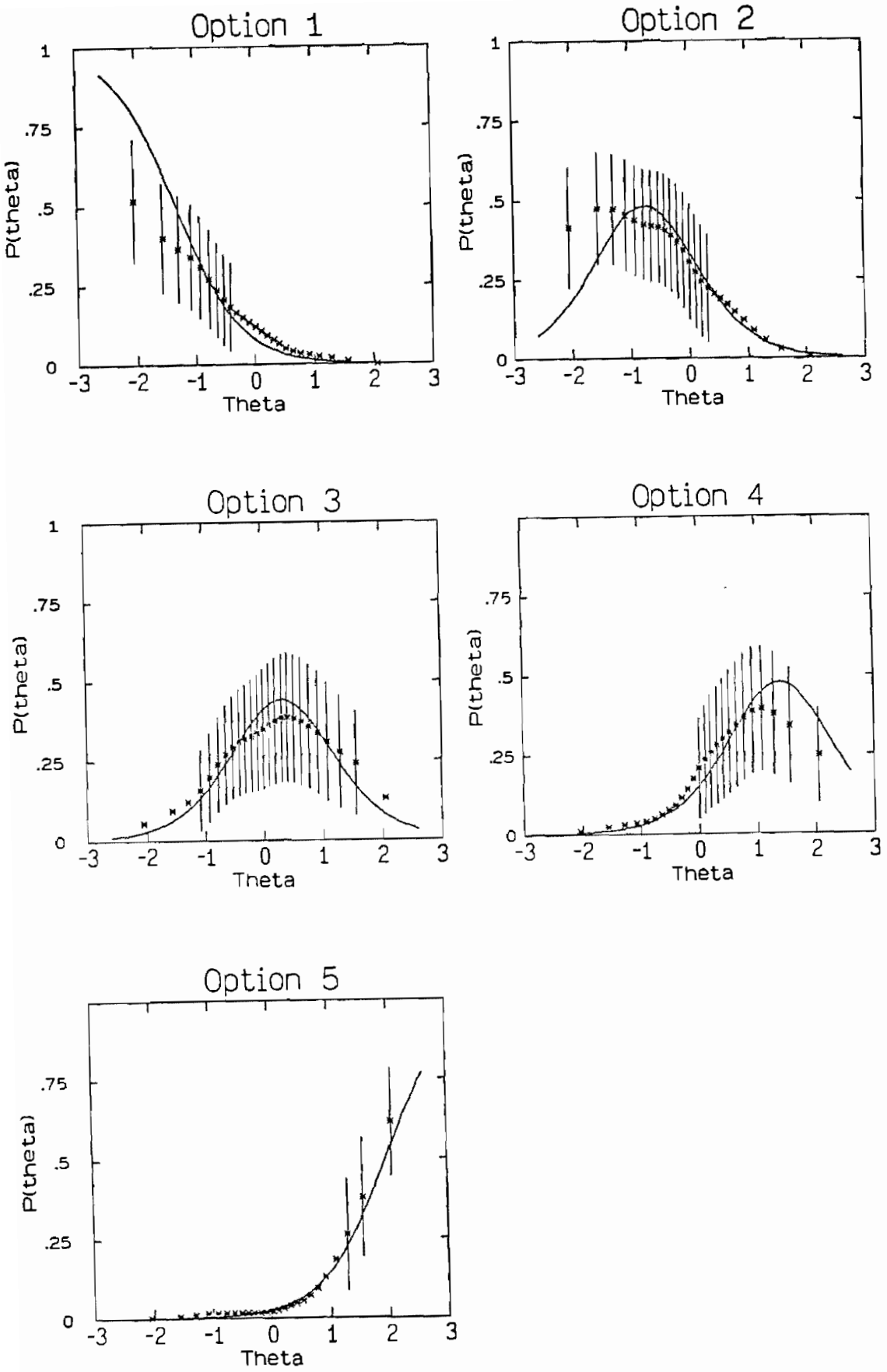

Figure 4. Fitplots of item 3 of the Negative Problem Orientation scale in the female sample according to the measurement invariant model.

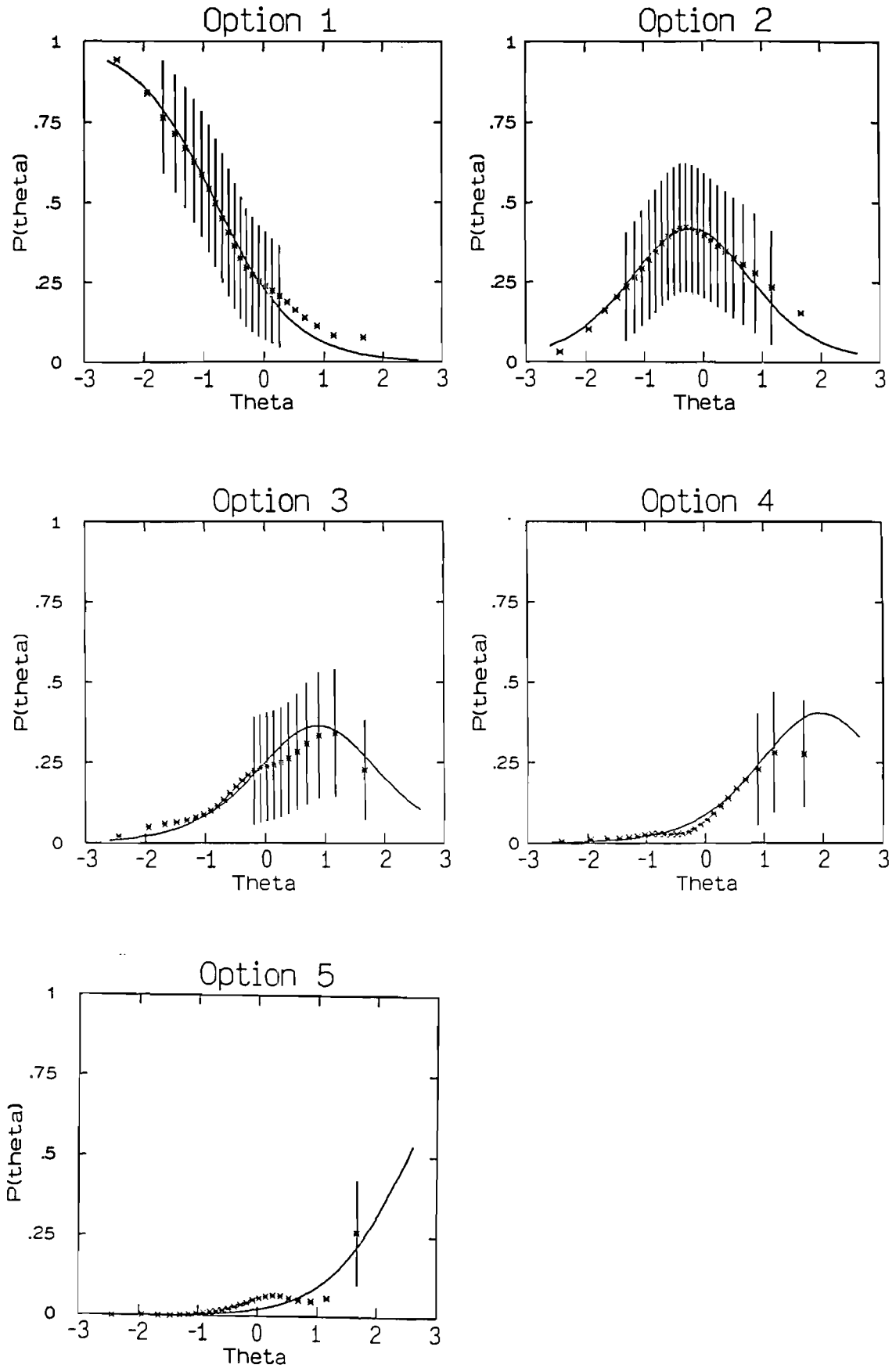

Figure 5. Fitplots of item 2 of the Negative Problem Orientation scale in the male sample according to the measurement invariant model.

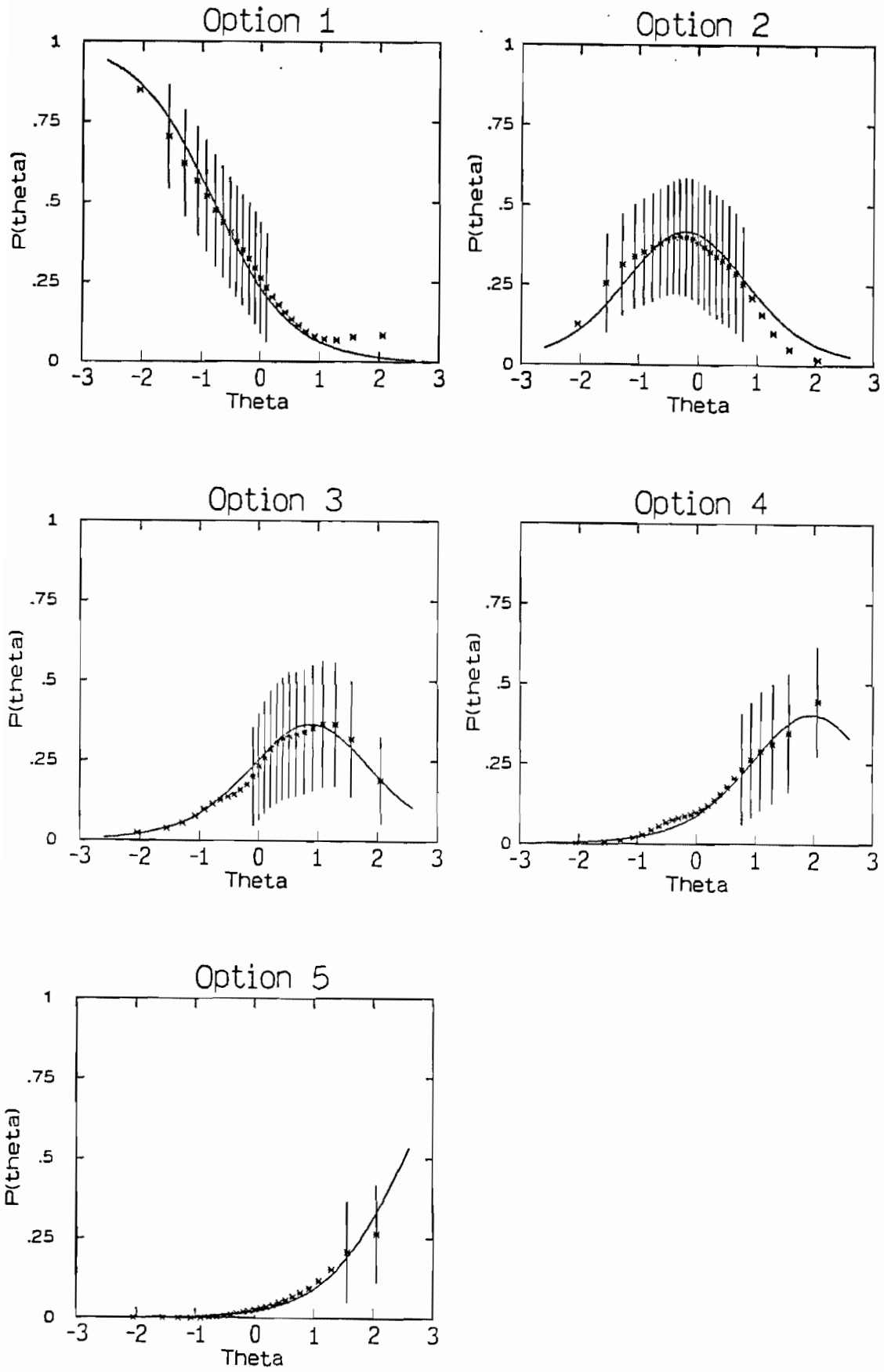

Figure 6. Fitplots of item 2 of the Negative Problem Orientation scale in the female sample according to the measurement invariant model.

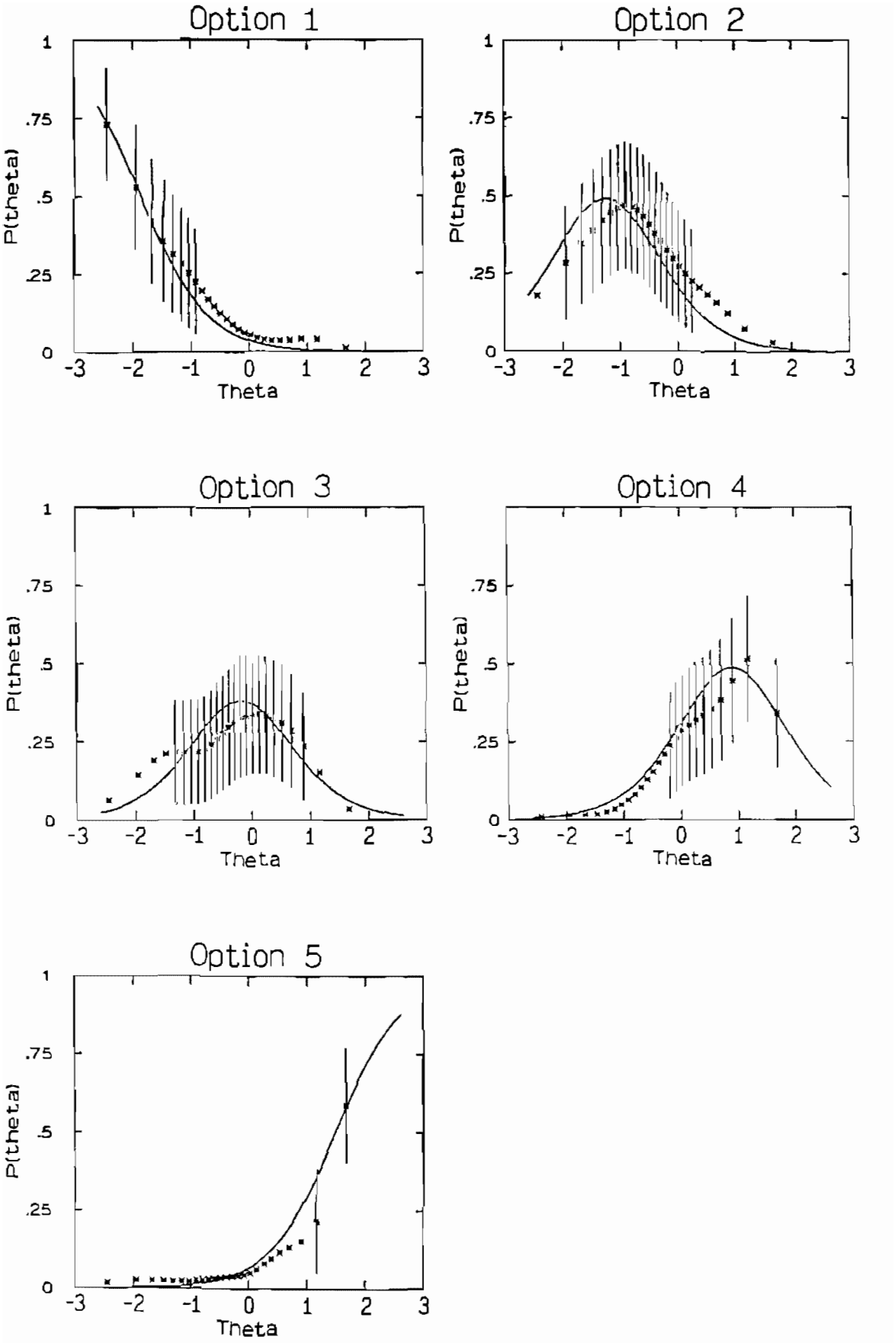

Figure 7. **Fitplots of item 6 of the Negative Problem Orientation scale in the male sample according to the measurement invariant model.**

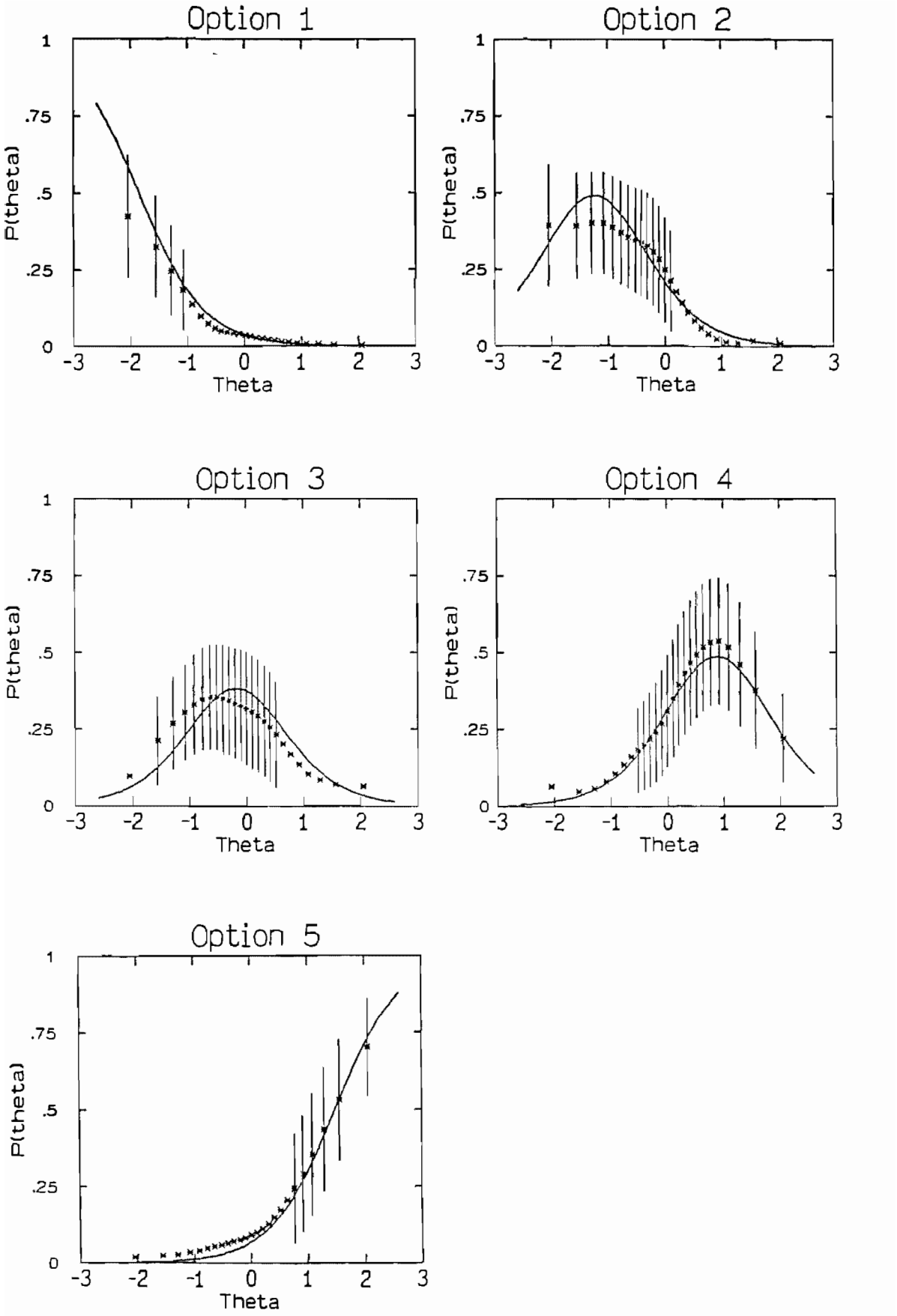

Figure 8. Fitplots of item 6 of the Negative Problem Orientation scale in the female sample according to the measurement invariant model.