

Using Graphical Methods in Assessing Measurement Invariance in Inventory Data

Albert Maydeu-Olivares and Osvaldo Morera
University of Illinois at Urbana-Champaign

Thomas J. D’Zurilla
State University of New York at Stony Brook

Most measurements of psychological constructs are performed using inventories or tests in which it is assumed that the observed interdependencies among the item responses are accounted for by a set of unobserved variables representing the psychological constructs being measured. Using test scores, researchers often attempt to detect and describe differences among groups to draw inferences about actual differences in the psychological constructs measured by their instruments. However, direct comparisons using tests scores are not meaningful unless these are commensurable or invariant across populations. Most researchers simply assume that measurement 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 discusses the difficulties faced to evaluate measurement invariance when, as is most common, a psychological construct is assessed by means of a test or inventory composed of categorical items. In this context, graphical methods (fitplots) can be a useful auxiliary tool that we believe has been overlooked. An example is provided to illustrate the use of fitplots in assessing measurement invariance in inventory data.

Most measurements of psychological constructs are performed using multi-item inventories or tests 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 test scores, researchers often attempt to detect and describe differences among groups (e.g., males vs. females, Japanese vs. Americans, etc.) to draw inferences about actual differences in the psychological constructs measured by these tests or inventories. However, direct comparisons

The participation of the first author in this research was supported by a Post-doctoral Scholarship from the Ministry of Education and Science of Spain. Requests for reprints should be sent to Albert Maydeu-Olivares. Faculty of Psychology. University of Barcelona. P. Vall d’Hebron 171. 08035 Barcelona (Spain). E-mail: amaydeu@psi.ub.es

A. Maydeu-Olivares, O. Morera and T. D'Zurilla

using tests scores are not meaningful unless these test scores 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, comparisons are meaningless because either a different construct is being measured in each of the populations or alternatively, 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 test 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.

This article is intended as a tutorial for those researchers interested in assessing measurement invariance in inventory or test data within the context of latent trait models. The focus of the article, however, is on assessing measurement invariance within unidimensional parametric item response models. In this context, graphical methods (fitplots) can be a useful auxiliary tool that we believe has been overlooked and an example is provided to illustrate the use of fitplots in assessing measurement invariance in inventory data.

Measurement Invariance, Factorial Invariance, Test Bias, Item Bias, and Differential Item Functioning (DIF)

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

$$(1) \quad \text{Prob}(\mathbf{y}|\boldsymbol{\theta} = t, X = x) = \text{Prob}(\mathbf{y}|\boldsymbol{\theta} = t), \text{ for all values of } x \text{ and } t$$

that is, if the probability of observing a set of measurements \mathbf{y} (a set of dependent variables) for a fixed level of the predictors $\boldsymbol{\theta} = t$, is independent of group membership. In other words, a set of measurements \mathbf{y} is invariant with respect to x if the relationship between \mathbf{y} and $\boldsymbol{\theta}$, given by $\text{Prob}(\mathbf{y}|\boldsymbol{\theta} = t)$ 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) \mathbf{y} and the independent variables $\boldsymbol{\theta}$ can be uni or multidimensional, continuous or categorical, and their relationship given by $\text{Prob}(\mathbf{y}|\boldsymbol{\theta} = t)$ can be linear or nonlinear. For instance, if \mathbf{y} and $\boldsymbol{\theta}$ are single observable continuous variables and $\text{Prob}(\mathbf{y}|\boldsymbol{\theta} = t)$ 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.

This article, however, discusses the case in which (a) the measurements (dependent variables) \mathbf{y} are inventory or test items, (b) the independent variables $\boldsymbol{\theta}$ representing the psychological constructs being measured are unobserved continuous variables, and (c) the relationship between the dependent and independent variables is given by a parametric non-linear latent trait model.

There are some terms that are closely related to the concept of measurement invariance. For instance, within the context of the common factor model (i.e., a *linear* latent trait model) the term *factorial invariance* is commonly used in place of measurement invariance. Also, when the data to be fitted are the items of a test 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 test 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 test for which measurement invariance holds and some items for which it does not hold. The items for which measurement invariance does not hold are said to be “biased” or to show *differential item functioning* (DIF).

Linear vs. Non Linear Latent Trait Modeling of Inventory Data

Latent trait models for the relationship between a p -dimensional set of observed dependent variables \mathbf{y} and a q -dimensional set of unobserved continuous variables $\boldsymbol{\theta}$ are characterized by the conditional or local independence assumption

$$(2) \quad \text{Prob}(\mathbf{y}|\boldsymbol{\theta} = t) = \prod_i \text{Prob}(\mathbf{y}_i|\boldsymbol{\theta} = t) \quad i = 1, \dots, p$$

for some $q < p$. Elementary probability theory yields from Equation 2 the usual expression for latent trait models (e.g., Bartholomew, 1987)

$$(3) \quad \text{Prob}(\mathbf{y}) = \int \prod_i \text{Prob}(y_i | \theta = t) f(t) dt$$

where $f(t)$ denotes the density of the unobserved variables θ which, in the context of latent trait models, are denoted as latent traits. This is a wide class of models which includes among many other models the common factor model, and most item response models (IRMs).

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).

There are two general approaches to estimating item response models (see Mislevy, 1986; Takane & de Leeuw, 1987). The first one consists in estimating the nonlinear relation between the items and the latent traits using all the information contained in the pattern of item responses by maximum likelihood estimation. The second approach consists in estimating the nonlinear relation between the items and the latent trait minimizing a weighted distance based on measures of pairwise association between item responses (e.g., tetrachoric or polychoric correlations), therefore, employing only partial information about the subjects' responses. Full information estimation of item response models is implemented for example in BILOG (Mislevy & Bock, 1990), MULTILOG (Thissen, 1991) or PARSCALE (Muraki & Bock, 1993), whereas limited information estimation of item response models is implemented for instance in LISCOMP (Muthén, 1987), LISREL (Jöreskog & Sörbom, 1998), or EQS (Bentler & Wu, 1995).

As examples of unidimensional item response model consider the two-parameter logistic item response model (Birnbaum, 1968) and the two-parameter normal ogive model (Lord, 1952) which may be appropriate models for inventories whose items only have two options (for instance: yes-no, agree-disagree). According to these model, the probability that a subject with standing t on the latent trait θ endorses item i can be expressed as

$$(4) \quad \text{Prob}(y_i = 1 | \theta = t) = \frac{1}{1 + \exp[-a_i(t - b_i)]}$$

in the two-parameter logistic model and as

$$(5) \quad \text{Prob}(y_i = 1 | \theta = t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-a_i(t-b_i)} \exp\left(-\frac{z_i^2}{2}\right) dz$$

in the two-parameter normal ogive model. In both cases we assume that item i is coded $y_i = 1$ for endorsement and $y_i = 0$ for non-endorsement. The probability that the same subject does not endorse the item is then given by $\text{Prob}(y_i = 0|\theta = t) = 1 - \text{Prob}(y_i = 1|\theta = t)$. In these models, 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).

Often, tests or inventories consist of Likert-type items. In such cases, a model such as Samejima's (1969) graded model can be used instead of the two-parameter models described above. 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 (coded, say, from 0 to 4) is given by

$$\begin{aligned}
 & \text{Prob}(y_i = 0|\theta = t) = 1 - \text{Prob}(y_i \geq 1|\theta = t) \\
 & \text{Prob}(y_i = 1|\theta = t) = \text{Prob}(y_i \geq 1|\theta = t) - \text{Prob}(y_i \geq 2|\theta = t) \\
 (6) \quad & \text{Prob}(y_i = 2|\theta = t) = \text{Prob}(y_i \geq 2|\theta = t) - \text{Prob}(y_i \geq 3|\theta = t) \\
 & \text{Prob}(y_i = 3|\theta = t) = \text{Prob}(y_i \geq 3|\theta = t) - \text{Prob}(y_i \geq 4|\theta = t) \\
 & \text{Prob}(y_i = 4|\theta = t) = \text{Prob}(y_i \geq 4|\theta = t)
 \end{aligned}$$

where each of the probabilities appearing on the right hand side of Equation 6 is modeled by either a two-parameter logistic or normal ogive model. For instance, in the case of logistic functions

$$(7) \quad \text{Prob}(y_i \geq k|\theta = t) = \frac{1}{1 + \exp[-a_i(t - b_{ik})]} \quad i = 1, \dots, n; k = 1, \dots, m-1$$

That is, in Samejima's model, each item has one a_i parameter and $m - 1$ threshold parameters b_{ik} , 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 model.

In the item response literature the functions $\text{Prob}(y_i = k|\theta = t)$ are called the *option response functions* (ORFs) of the model.¹ 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 models and the graded model just described are useful models for fitting binary and

¹ In the literature, these functions are also referred to as "category response functions" or "item category response functions".

Likert items, respectively. Note, however, but there is a vast array of other IRMs that can also be used. Thissen and Steinberg (1986) provide a useful taxonomy of unidimensional parametric IRMs (see also van der Linden & Hambleton, 1997).

Assessing Measurement Invariance Within Item Response Models

The assessment of measurement invariance across populations in item response models may proceed as follows:

1. Select an IRM that may be appropriate to the data and fit it separately to each of the populations.

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) the latent trait means and variances are freely estimated except for some minimal constraints to identify the model. If this model satisfactorily fits the data, then the inventory (under the selected model) is *measurement invariant* across populations with respect to the latent trait. If it fails to be measurement invariant we proceed to the next step.

3. Identify which items are causing the misfit. In other words, identify which items are biased or show DIF under the selected IRM. This can be done as follows: First, we fit a measurement invariant model in which the parameters of an item are *not* constrained to be equal across populations. Then, 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, we should perform n separate analyses.

Once the DIF items have been located, a decision about whether to delete or retain the DIF items should be made based on judgment about the source of DIF. If the DIF items are retained, then our inventory or test is only *partially* measurement invariance (Byrne, Shavelson & Muthén, 1989). If, on the other hand, the DIF items are removed from the inventory, then we may want to investigate whether the revised inventory is measurement invariant by repeating Step 2.

However, the assessment of measurement invariance by fitting an item response model faces a major problem: how does one assess the goodness-

of-fit of the model to the data? Indeed, assessing the goodness-of-fit of IRMs is considerably more difficult than in linear factor models (see McDonald & Mok, 1995). The next section is devoted to this important issue.

Assessing the Goodness-of-Fit in Item Response 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 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

$$(8) \quad G^2 = 2 \sum_{\text{all cells}} \hat{c}_A \log \frac{\hat{c}_A}{\hat{c}_B}$$

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, if the larger model (e.g. 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 (e.g. an unrestricted multinomial model), provided that the contingency table has few empty cells. 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.²

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

² 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 G^2 statistic could be used with moderately large sample sizes.

$$(9) \quad \chi^2 = \sum_{\text{all cells}} \frac{(c - \hat{c})^2}{\hat{c}}$$

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, X^2 the statistic follows in large samples a chi-square distribution provided that the contingency table has few empty cells.

Software programs that estimate the parameters of IRMs by full information maximum likelihood (e.g., BILOG, MULTILOG) routinely provide G^2 statistics. On the other hand, software programs for structural equation modeling that can estimate the normal ogive model using limited information estimation (e.g., LISCOMP, LISREL or EQS) do not provide estimates of the G^2 or X^2 statistics.³

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 parameters of the IRM 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, *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, solely relying on this nested test to determine measurement invariance is inappropriate.

In Step 3 of the procedure described earlier we can also use nested G^2 statistics to assess on a one item at a time basis whether the items in the

³ Although programs for limited information estimation of the normal ogive model provide also a chi-squared goodness of fit test of the model, this test does not directly assess the goodness of fit of the model against the data. Instead, it assesses whether the restrictions imposed by the normal ogive model on the tetrachoric/polychoric correlations are plausible. See Muthén (1993) for a thorough discussion of this issue.

test 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 for 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 may be used in Steps 2 and 3 of the procedure presented above to assess measurement invariance within IRMs. Since the G^2 statistic is provided in the output of standard software programs for full information maximum likelihood estimation with facilities to handle multiple populations such as MULTILOG (Thissen, 1991), Steps 2 and 3 can be performed readily.⁴ However, how shall we perform Step 1?

Using Graphical Methods (Fitplots) to Explore Measurement Invariance in Item Response Modeling

There is not yet a fully satisfactory solution to the problem of assessing the goodness-of-fit of item response models when m^n is large relative to the sample size. Therefore, there is not a definitive answer as to what is the best way to perform Step 1. In this article, we would like to point out a technique, the use of fitplots, that has been available for some time in item response modeling⁵ and that we believe can be rather helpful in assessing measurement invariance. By using fitplots, we mean to examine the plots of the option response functions (ORFs) as well as confidence intervals around some points in the latent trait continuum for each of the ORFs.

Specifically, if all ORFs fall within the estimated confidence intervals, this would suggest that the model fits the data. Otherwise, if we observe that the ORFs fall outside the confidence intervals in one or more items, then these particular items are not well fitted by the model. It is important to realize, however, that the fitplots are more useful in assessing the misfit rather than the fit of the model. If the fitplots show no misfit, this is an indication that the model fits appropriately the first order marginals of the overall contingency table. Yet, it may very well be the case that a model satisfactorily fits the first order marginals and but does not satisfactorily fit the overall contingency table. In other words, a good fit of the model as

⁴ For detailed instructions, see Example 17 in Thissen (1991), and Thissen, Steinberg and Wainer (1993).

⁵ They are for instance available in BILOG (Mislevy & Bock, 1990).

A. Maydeu-Olivares, O. Morera and T. D'Zurilla

assessed by fitplots may be interpreted, *at best*, 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.

Fitplots are useful to help us assessing measurement invariance within parametric unidimensional IRMs in two ways: In Step 1, by obtaining fitplots for the non-measurement invariant model, we have an indication of whether the selected IRM fits the data, at least approximately. In Step 3, by obtaining fitplots for the measurement invariance model, we can *see* the magnitude of DIF for any given item. Hence, fitplots provide some assessment of the relevance of DIF for any given item. Most importantly, since fitplots provide us with an assessment of fit for varying levels of the latent trait and at every option of the item, they help us better describe DIF.

It is particularly important to obtain some assessment of the goodness of fit of the non-measurement model. Nested G^2 statistics are not meaningful unless we have some indication that the chosen item response model fits the data. This can be obtained, for instance, by the use of fitplots. In fact, fitplots performed in Step 1 may reveal items that are not well fit by the model in one of the populations, thus providing an early detection of possible DIF.

An Example:

Assessing Gender Measurement Invariance in Problem Orientation

The term *Problem orientation* (D'Zurilla, 1986) refers to 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. Problem orientation is not a unidimensional construct, but instead, represents two different, albeit related constructs, that is, *positive* problem orientation and *negative* problem orientation (Maydeu-Olivares & D'Zurilla, 1995). Consequently, we will assess measurement invariance using the procedures described above on separate measures of these two constructs.

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, 1998). 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 have found gender mean differences in problem orientation on both scales across different samples and age groups (D'Zurilla, Maydeu-Olivares & Kant, 1998). In the present sample, 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?

We may want to answer this question within the framework of item response models since the SPSI-R was carefully constructed so that each of its scales be unidimensional (see Maydeu-Olivares & D'Zurilla, 1995).

Here we shall fit Samejima's logistic graded model to these data using full information maximum likelihood as implemented in MULTILOG (Thissen, 1991). The fitplots were drawn as in Drasgow, Levine, Williams, Tsien and Mead (1995) using IOCCDRAW (Williams, 1992). However, in Drasgow et al., model parameters were estimated on one sample and fitplots obtained on a cross-validation sample. This requires splitting in half the available data. Given our sample size, in this article parameter estimates and fitplots were obtained on the same set of data. It is worth emphasizing that fitplots can be obtained for any other unidimensional item response model. For instance, fitplots of the PPO scale under Samejima's graded normal ogive model and a nonparametric model can be found in Maydeu-Olivares (1994).

Assessment of Measurement Invariance

Step 1

We fitted Samejima's logistic graded model to each scale using the multiple group facilities of MULTILOG (Thissen, 1991) without imposing equality constraints across genders on the item parameters. The estimated item parameters of this non-measurement invariant model are presented in Table 1.

Table 1
Item Parameters Estimated by Maximum Likelihood Using Samejima's Graded Model

Negative Problem Orientation															
Non-measurement Invariant Model											Measurement Invariant Model				
Item	Men					Women					Men and Women				
	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				
Item	Men					Women					Men and Women				
	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	-0.25	-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	-0.84	-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	2.0

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 7.

In MULTILOG, a normal distribution of the latent trait is assumed. For identification purposes, the variance of the latent traits were set to one in both genders, and the latent trait means for one of the genders (women) were set to zero. 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 $.06$ in both cases.

To assess the practical goodness-of-fit of the non-measurement invariant model we inspected the fit plots for all PPO and NPO items. In other words, we assessed whether Samejima's graded model fitted appropriately the men and women's samples separately. 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 of both PPO and NPO 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 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. In these figures, the solid line represents the option response function (ORF) under Samejima's logistic graded model, given by Equations 6 and 7. As can be observed in these plots, according to this model 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, empirical proportions (see Drasgow et al., 1995: p. 147) represented by a * for 25 equally spaced points on the latent trait continuum are drawn along with their estimated 95% confidence intervals. Whenever a small number of people in an interval of the latent trait continuum chose a particular option, the confidence interval around the empirical proportions was not drawn to indicate that that particular empirical proportion (and its confidence interval) may be very poorly estimated. Hence, 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 $+2$.

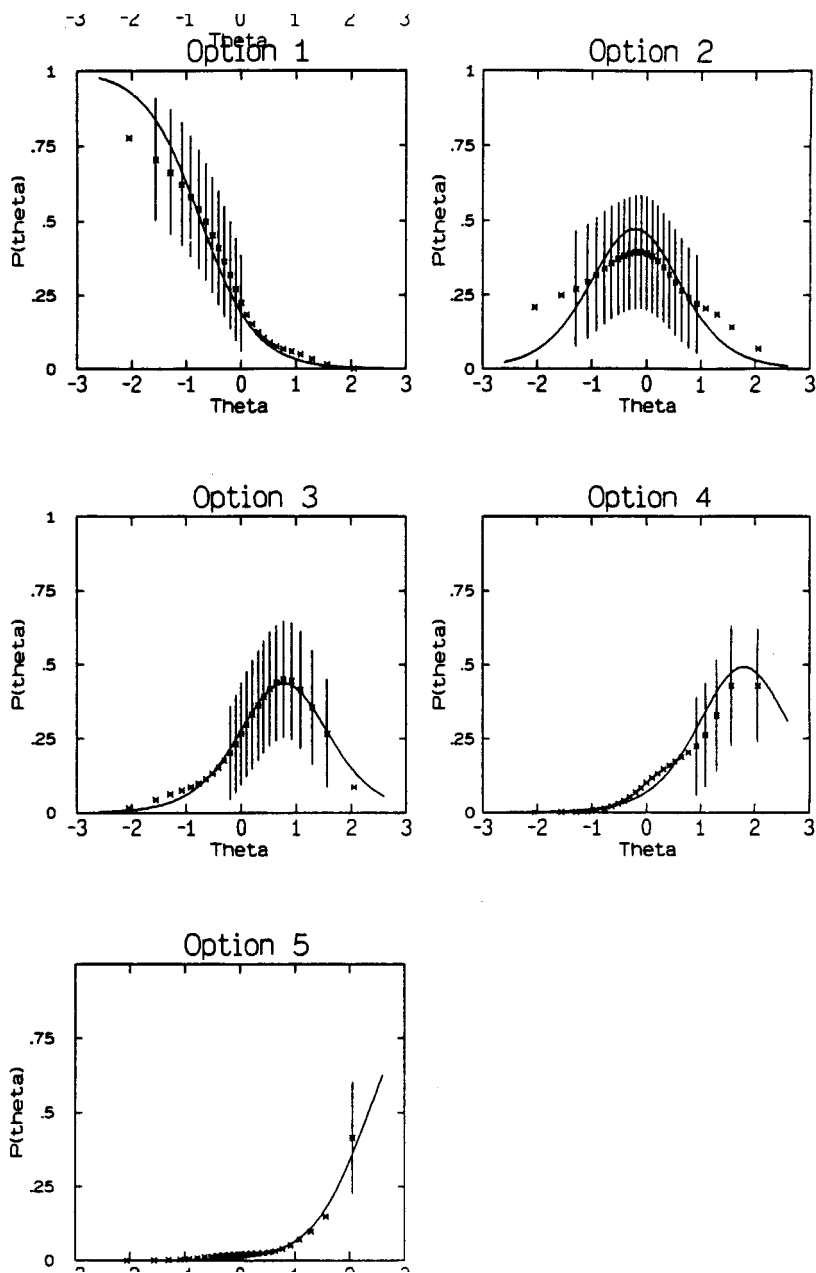


Figure 1

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

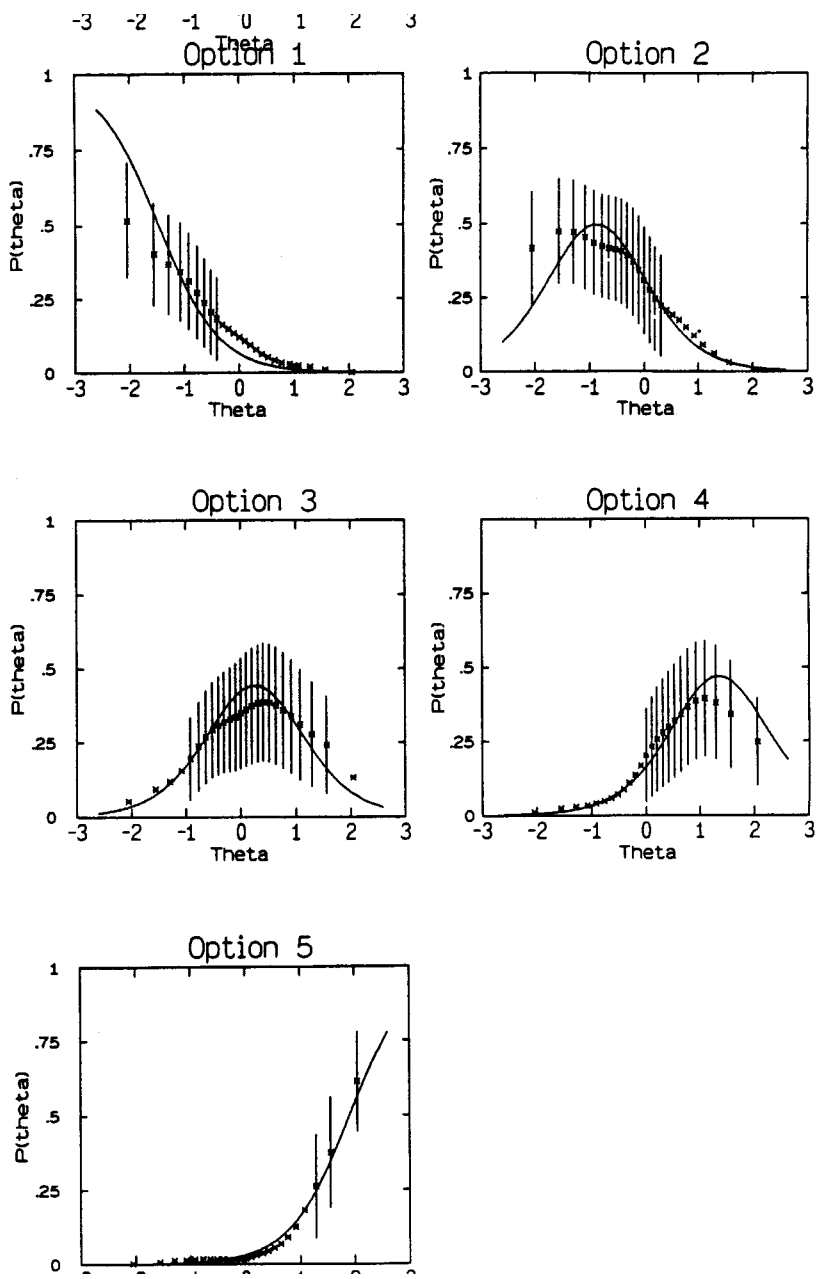


Figure 2
Fitplots of item 3 of the Negative Problem Orientation scale in the female sample according to the non-measurement invariant model.

As for each item model fit is evaluated at different points, a criterion is needed for deciding whether the fitplots evidence a poor fit. We shall reject the model for an item if any intervals fail to include the ORF for any response option. All PPO items met this condition. However, in the NPO scale, items 3 and 10 did not meet this rule of thumb in the female sample.

In fact, in both scales the fitplots indicated that the model fitted somewhat better the male than the female sample. This can be readily observed in Figures 1 and 2. For men (Figure 1), the model slightly overpredicts the probability of endorsing option 1 at low levels of NPO and slightly overpredicts endorsing option 2 at medium levels of NPO. Since the predicted ORFs are within the estimated confidence intervals, these misfits could be considered of minor importance. In women (Figure 2), the model also overpredicts the probability of endorsing option 1 up to a level of about -1 in the NPO latent trait and then it overpredicts it above this level. 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 it lies within the confidence intervals.

Thus, the fitplots suggests that NPO items 3 and 10 behave differently across populations, as these items are not well fit by the model only in the female sample. At this point, a decision about whether to delete or retain these items should be made. For illustrative purposes we shall for now retain these items.

Step 2

We shall estimate a measurement invariant model and compare its fit with that of a non-measurement invariant model. A measurement invariant model is obtained under Samejima's logistic graded model by forcing the item parameters to be equal across gender. These constraints suffice to meet the definition of measurement invariance given by Equation 1. After fixing the latent trait means for women at zero for identification purposes, men's latent trait means were estimated as -.39 in NPO and .31 in PPO. The parameters of this measurement invariant model are also provided in Table 1.

The values of the G^2 statistic for PPO and NPO under the measurement invariant model are 2306.1 and 12443.6, respectively, and under the non-measurement invariant model are 2275.1 and 12350.7, respectively. Nested G^2 statistics were computed to determine whether the measurement invariant model fits significantly worse than the non-measurement invariant model, obtaining $G^2_{dif} = 31$ on 25 df , $p = .196$, for PPO, and $G^2_{dif} = 92.9$ on 50 df , $p < .001$, for

NPO.⁶ Given these results, we conclude that measurement invariance holds for PPO but not for NPO. The analysis of PPO is finished, since the use of the fitplots has determined that the model approximately fits the data, and the use of a nested G^2 statistic has shown that measurement invariance holds. For NPO, we will perform Step 3 and to determine which items account for the significant nested G^2 statistic obtained for this scale. For illustrative purposes, we shall also perform Step 3 for the PPO items.

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 df , $p = .623$. 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.

Using a Type I error of $\alpha = .05$ we found that as expected, none of the PPO items showed evidence of DIF. As for NPO, the three items that showed largest evidence of DIF were items 3, 6 and 10. Of these, only the G^2_{dif} statistic of items 6 and 10 is significant at this alpha level.

This indicates that the fit of these items is adversely affected when measurement invariance constraints are introduced in the model.

There is a difference between items 6 and 10, however. Item 6 was well fitted by the model across populations before measurement invariance constraints were introduced, whereas item 10 was not. On the other hand, the model did not provide a good fit to item 3 across populations to start with, but the fit was not significantly worsened by the introduction of constraints across populations. This can be readily observed by comparing the fitplots for this item under the measurement invariant model (Figures 3 and 4) and under the non-measurement invariant model (Figures 1 and 2).

⁶ When performing a nested G^2 statistic, there are five degrees of freedom for every item whose parameters are not constrained to be equal across groups. This is because each item has one a parameter and four b parameters and the restrictions imposed on the factor means and variances by MULTILOG are the same for the measurement and non-measurement invariant models.

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			

Discussion

We shall now return to the question posed when introducing this example, namely, whether observed gender differences in problem orientation are due to actual differences in the problem orientation latent traits. 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.

Items 3, 6 and 10 could be removed from the NPO scale and repeat Steps 1 and 2. In so doing we found that this shortened NPO scale is measurement invariant for a Type I error of $\alpha = .05$, $G^2_{dif} = 49.0$ on 35 df , $p = .058$. The mean of men's NPO latent trait in this latter 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.

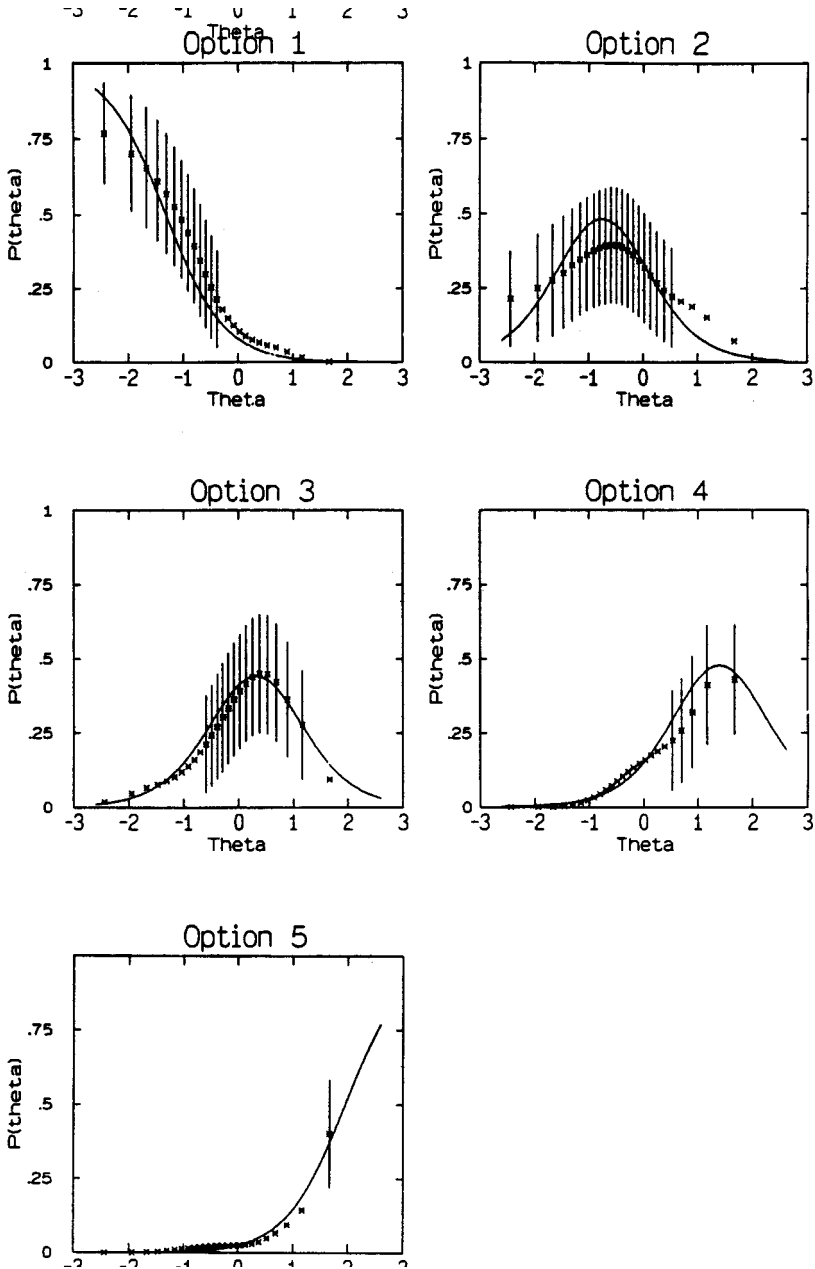


Figure 3

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

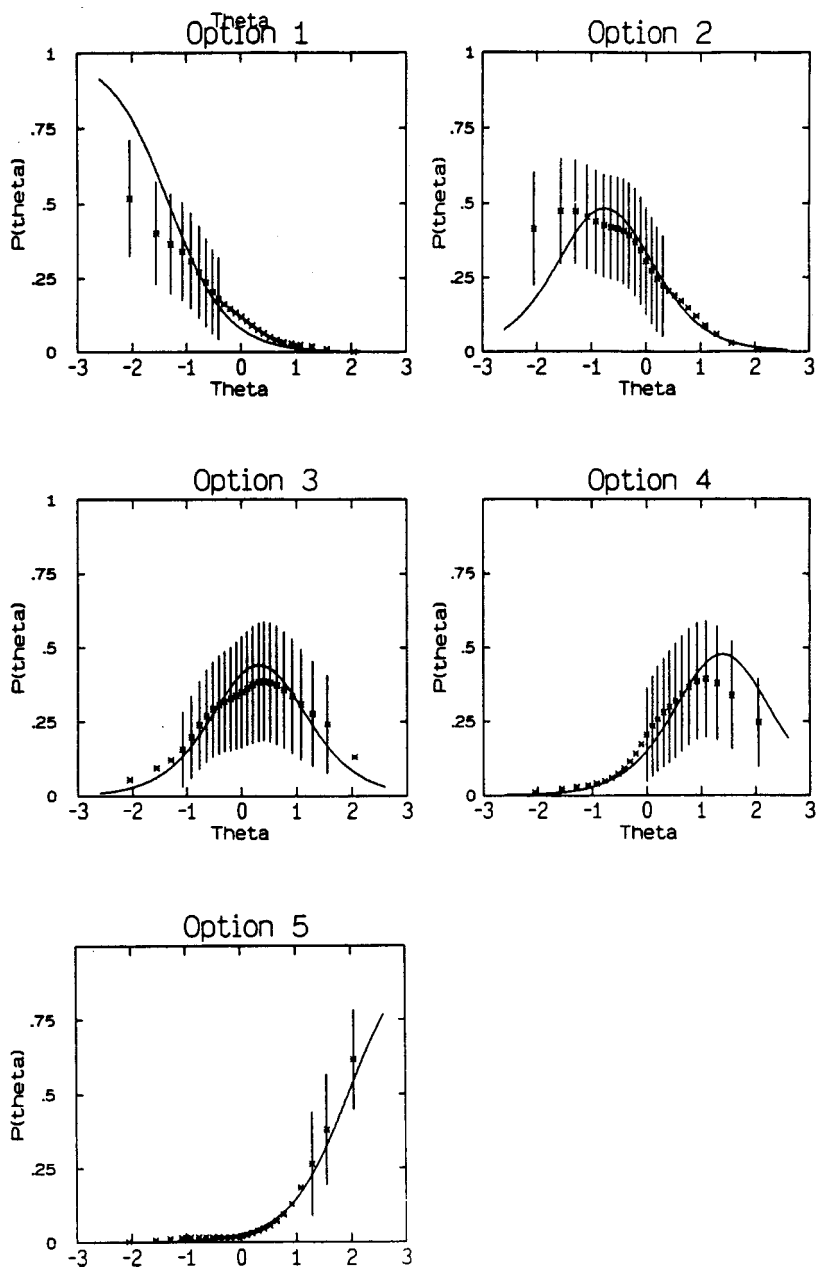


Figure 4

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

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 tests 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 assess whether the selected model fits the data. However, existing statistics to assess the goodness-of-fit of item response models require samples much larger than those found in most psychological research. It is crucial to use some measure of model fit to the data since a nested test may fail to indicate lack of measurement invariance if most of the differences between ORFs across populations are omitted from the nested test because the ORFs do not capture the data in one of the populations in the first place. In this article, we used for this purpose a practical goodness-of-fit index, namely, the inspection of confidence intervals constructed for each of the option response functions under consideration.

References

- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Bartholomew, D.J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.
- Bentler, P.M. & Wu, E.J.C. (1995). *EQS/Windows User's Guide, Version 5*. 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.
- Byrne, B.M., Shavelson, R.J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.
- 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., Maydeu-Olivares, A. & Kant, G. L. (1998). Age and gender differences in social problem solving in college students, middle age, and elderly adults. *Personality and Individual Differences*, 25, 241-252.
- D'Zurilla, T.J., Nezu, A.M. & Maydeu-Olivares, A. (1998). *Manual of the Social Problem-Solving Inventory-Revised*. North Tonawanda, N.Y.: Multi-Health Systems, Inc.
- 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. (1993). *LISREL 8. User's reference guide*. Chicago, IL: Scientific Software.
- Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs*, No. 7.
- Maydeu-Olivares, A. (1994). Parametric vs. non-parametric approaches to individual differences scaling. *Psicothema*, 6, 297-310.
- 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.
- McDonald, R. P. & Mok, M. M. (1995). Goodness of fit in Item Response Models. *Multivariate Behavioral Research*, 30, 23-40.
- 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.
- Mislevy, R.J. & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Chicago, IL: Scientific Software.
- Muraki, E. & Bock, R.D. (1993). *PARSCALE: Parametric Scaling of Rating Data*. Chicago, IL: Scientific Software.
- 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.
- 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.
- Thissen, D. (1991). *MULTILOG 6: Multiple, categorical item analysis and test scoring using Item Response Theory*. 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.

A. Maydeu-Olivares, O. Morera and T. D'Zurilla

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. (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.

Van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag

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

Accepted October, 1998.