



Universidad Autónoma  
de Madrid

**Biblos-e Archivo**  
Repositorio Institucional UAM

**Repositorio Institucional de la Universidad Autónoma de Madrid**

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:

This is an **author produced version** of a paper published in:

Educational and Psychological Measurement 81.6 (2021): 1054 – 1088

**DOI:** <https://doi.org/10.1177/0013164421994321>

**Copyright:** © The Author(s) 2021

El acceso a la versión del editor puede requerir la suscripción del recurso

Access to the published version may require subscription

**Nominal Factor Analysis of Situational Judgment Tests: Evaluation of  
Latent Dimensionality and Factorial Invariance**

Javier Revuelta

Alicia Franco-Martínez

Carmen Ximénez

### Abstract

Situational judgment tests have gained popularity in educational and psychological measurement and are widely used in personnel assessment. A situational judgment item presents a hypothetical scenario and a list of actions, and the individuals are asked to select their most likely action for that scenario. Because actions have no explicit order, the item generates nominal responses consisting of the actions selected by the individuals. This paper shows how to factor-analyze the nominal responses originated from such a test, including the estimation of the number of latent factors and a factor invariance analysis in a multiple group design. The method consists of applying the MNCM, a multidimensional extension of the nominal categories model by Bock. The paper includes the results of two studies: 1) a simulation study about type-I error rate, statistical power and recovery of the parameters in a multi-group factorial invariance design, and 2) a real data example using responses to a situational judgment test measuring gender stereotypes to illustrate the approach. Results suggest the use of the AIC, BIC, and BICc indices to guide the selection of the number of factors with nominal responses. All the analyses are conducted using the computer program Mplus. The code is included as supplemental material for the readers so that they can adapt it to their own purposes.

**Keywords:** situational judgment test, nominal factor analysis, multidimensional nominal categories model, multi-group factor analysis, Monte Carlo simulation, gender stereotypes.

## **Nominal Factor Analysis of Situational Judgment Tests: Evaluation of Latent Dimensionality and Factorial Invariance**

Items from a situational judgment test (SJT) present a situation followed by a set of hypothetical response categories that the individual might choose in that situation. The task of the test takers consists of selecting the response option that best represents their way of behaving in that particular scenario (Lievens, Peeters, & Schollaert, 2008; McDaniel & Nguyen, 2001). Situational judgement tests have been widely used in industrial and organizational psychology (Ployhart & Ward, 2013), but they have gained popularity over recent years and have been extended to education and other contexts for the assessment of constructs, such as interpersonal skills, leadership development, etc. Unlike other test items formats, in a SJT the response categories consist of a list of behaviors that lacks an explicit order, thus the items are scored as nominal categories. However, the basic questions of psychological measurement are still relevant for such a test. This paper shows how to respond to the question of what is the factorial structure of a SJT scored in nominal categories.

The traditional approach to factor analysis is not immediately applicable to the SJTs because the factor analysis model assumes that scores have at least an ordinal measurement level (Floyd & Widaman, 1995). One solution to this problem is to assign arbitrary scores to the response categories of a SJT and assume that these scores are suitable for classical statistical analyses (Sharma, Gangopadhyay, Austin, & Mandal, 2013). However, apart from being arbitrary, the weights are implicitly based on the premise that the data are one-dimensional and afterward, they are used to test dimensionality, which is a rather circular argument. More recently, Zu and Kyllonen (2018) proposed to use Bock's model to score SJT to avoid the assignment of arbitrary scores. However, Bock's model assumes a one-dimensional latent structure and does not solve the problem of estimating the number of latent factors. This paper builds on Zu and Kyllonen (2018) and uses a multidimensional extension of Bock's model to estimate the number of factors and the factorial structure.

Multidimensional extensions of Bock's model were suggested by Takane and de

Leeuw (1987), Thissen, Cai, and Bock (2010), Revuelta (2014) and Thissen and Cai (2016). The multidimensional nominal categories model (MNCM) explains the distribution of the manifest variables by appealing to a vector of latent factors, and the relation between response categories and latent factors is estimated from the matrix of individual by item responses. The MNCM has been applied by Revuelta, Maydeu-Olivares, and Ximénez (2020) to perform exploratory and confirmatory factor analyses from item responses that do not have an explicit order.

This paper presents procedures based on the MNCM for the analysis of the factorial structure in SJTs. It also extends the MNCM to the multi-group analysis of factorial invariance for nominal data provided by a SJT. The paper is organized in five sections. Firstly, a Scale of Gender Stereotypes (SGS) is presented to illustrate a SJT and the type of data that motivates the subsequent theoretical developments. The SGS is a SJT in which each item presents a scenario for a group of characters and contains response categories representing different ways of behaving, thus generating nominal response data. Two different forms of the SGS were created, differing in the wording with which the group of characters is presented, and both forms have been applied to different samples of individuals, thus motivating a multi-group analysis. The second section contains the theoretical aspects, including the mathematical formulation of the MNCM and the different levels at which factorial invariance can be investigated. Thirdly, a simulation study is presented to gather information about the capabilities of the methods to estimate the number of latent factors, test hypotheses about factor invariance and the recovery of the parameters in realistic conditions. The fourth section of the paper includes a real data analysis that consists of an empirical study estimating the number of latent factors and investigating factorial invariance for the SGS. Finally, a general discussion summarizes the results of the simulation and the empirical studies and concludes the paper.

All the analyses were performed using the computer program Mplus (Muthén & Muthén, 1998-2017). The estimation algorithms and fit statistics are purposely restricted to those included in Mplus to facilitate the dissemination of these methods, as

Mplus is one of the most accessible and widely used software for fitting latent variable models. The Mplus code is available in the supplemental material so that the readers may adapt it to their own data.

### **A situational judgment test of gender stereotypes**

The SGS consists of 10 items presenting brief scenarios. Each scenario includes a group of characters in a certain situation, and the item contains three response categories that represent behaviors that such characters may follow in that situation. The test takers must choose which behavior is more plausible for the group in the situation presented. The content of the scenarios has been designed based on the gender stereotypes of the *traits* and *roles* dimensions collected in López-Sáez, Morales, and Lisbona (2008). Five items are extracted from the stereotypes related to the personality traits socially assigned to men and women (items X1 to X5), and the other five items with the stereotypical social roles of both genders (items X6 to X10). The response categories are designed to represent a supposedly male, female, or neutral stereotypical way of behaving in the situation posed by the item.

Two forms of the test were created differing in the terms used to present the characters of the story. For both forms, the group of characters is assumed to be gender-mixed. The SGS is written in Spanish language, and the two forms differ in the linguistic markers used to denote the characters of the story. For example, imagine that you want to write about your group of friends. In Spanish, the morpheme “-o” is reserved for the masculine grammatical gender as well as for the generic *he* (i.e., the *masculine generics*, MG), so that the Spanish word “*amigos*” (friends) denotes both a group of friends composed only of men, or a group composed of men and women. The word “*amigas*” denotes a group of friends composed only of women, and the *alternative generics* (AG) “*amigos/as*” explicitly emphasizes that the group is gender-mixed.

In the first form of the SGS, the group is mentioned using the MG (e.g., “*amigos*”); and in the second form, using the AG (e.g., “*amigos/as*”). The purpose of developing two forms of the test is to investigate whether or not the wording of the

generic (which is currently a strong international debate) affects the gender stereotypes that they elicit in the individuals responding to the test. An English translation of the SGS is presented in the supplemental material of this paper, as well as the Spanish wording of the MG and AG forms.

Responses to the SGS consists of nominal categories. However, as the task of the respondent is to select the most plausible category, we can assume that there is an underlying order of preference between categories that can be estimated from the data. The purpose of the data analysis of the SGS is twofold: First, we want to know how many latent factors are measured by the test, and second, we want to investigate whether the factorial structure of the SGS differs from the MG and AG forms of the questionnaire. According to previous findings on gender stereotypes, we formulated two hypotheses as follows. First, latent structure will not differ from MG to AG, as both generic forms are designed to measure the same phenomenon: the male over-representational bias (*Hypothesis 1*). However, congruent with the literature (e.g., Kaufmann & Bohnert, 2014), this bias will not be equally elicited in both generic forms. Hence, we rather expect that the response frequencies will differ between generics (*Hypothesis 2*). More specifically, our interest is to know whether the stereotypically male-categories will be more frequently selected when using MG as compared to AG. Most of the previous studies in the literature about gender stereotypes have addressed these questions by using simple frequency comparisons (Hamilton, 1988; Kaufmann & Bohnert, 2014; Merritt & Kok, 1995; Nissen, 2013). However, the application of the MNCM makes it possible to explore these hypotheses in a novel way, based on latent variable models for nominal responses.

### Theoretical background

The MNCM is an extension of the (one-dimensional) nominal categories model by Bock (1972) and Bock and Gibbons (2010) to the multidimensional latent space. The MNCM has been successfully applied to conduct exploratory and confirmatory factor analyses from nominal responses (Revuelta, 2014; Revuelta et al., 2020). In this paper,

the MNCM is extended to the multi-group research design and applied to a real data sample from the SGS.

### Multinomial probit and logit models

The MNCM assumes that the individuals who respond to a nominal item can rank order the categories according to their preferences. In this nominal response format, which is called *first-choice data*, the selected category is the most preferred one. More specifically, suppose that an item is scored in  $K$  categories. The item elicits  $K$  latent responses that indicate the utility (preference or attractiveness) of each category. Let  $\mathbf{u} = (u_1, \dots, u_k, \dots, u_K)$  be the vector of utilities of the item categories, where  $u_k$  is the utility of category  $k$ . The manifest response,  $Y$ , is the category with maximal utility:

$$Y = \underset{k}{\operatorname{argmax}}(u_k).$$

Utilities depend on a vector of  $D$  latent factors,  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_D)'$ , by the equation (Bock, 1975; Fahrmeir & Tutz, 1991):

$$u_k = \tau_k + \boldsymbol{\lambda}_k' \boldsymbol{\eta} + e_k, \quad (1)$$

where  $\tau_k$  is a scalar intercept parameter for Category  $k$ ,  $\boldsymbol{\lambda}_k = (\lambda_{k1}, \dots, \lambda_{kD})'$  is the vector of slopes of the category, and  $e_k$  is a normally distributed error with zero mean and variance  $\sigma^2/2$ , where  $\sigma^2$  is arbitrary.

Category  $k$  will be selected if the difference between its utility ( $u_k$ ) and the utility of any other category ( $u_j$ , for  $j \neq k$ ) is positive. Let  $d_{kj} = u_k - u_j$  be the difference between two utilities. In matrix terms, the vector of differences between  $u_k$  and the other  $K - 1$  utilities is  $\mathbf{d}_k = \mathbf{C}\mathbf{u}$ , where  $\mathbf{C}$  is a matrix of fixed coefficients that depends on the categories being compared. For example, Category 1 for an item with  $K = 3$  categories will be selected if the differences  $d_{12} = u_1 - u_2$  and  $d_{13} = u_1 - u_3$  are positive. The vector of differences for such an item is based on the following matrix of coefficients



$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \quad (2)$$

For a fixed value of  $\boldsymbol{\eta}$ , the probability of selecting Category 1 is

$P(Y = 1 | \boldsymbol{\eta}) = P(d_{12} \geq 0, d_{13} \geq 0)$ , and is given by the probability in the first quadrant of the bivariate normal distribution of  $(d_{12}, d_{13})$ . In general, the probability of selecting category  $k$  for an item with  $K$  categories is given by

$$\begin{aligned} P(Y = k | \boldsymbol{\eta}) &= P\left(\bigcap_{j \neq k} (d_{kj} \geq 0)\right) \\ &= \int_0^\infty \int_0^\infty \cdots \int_0^\infty f(\mathbf{d}_k; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{d}_k \end{aligned} \quad (3)$$

where  $f()$  represents a multivariate-normal density function with mean  $\boldsymbol{\mu} = \mathbf{C}(\boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta})$  and variance-covariance matrix  $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{E}\mathbf{C}'$ , the symbol  $\boldsymbol{\Lambda}$  represents the matrix of item slopes, and  $\mathbf{E}$  is the variance-covariance matrix of the errors. Because of the above assumptions,  $\mathbf{E}$  is a diagonal matrix of values  $\sigma^2/2$ , and matrix  $\boldsymbol{\Sigma}$  has the value  $\sigma^2$  on the diagonal and its non-diagonal elements are  $\sigma^2/2$ . The model in Equation (3) is referred to as the multinomial probit model.

Equation (3) is difficult to compute because it involves a  $K - 1$  multi-dimensional integral. Although the equation could be approximated using numerical methods such as Gauss-Hermite quadrature, a more convenient analytical approximation is available. If one sets  $\sigma^2 = \pi^2/3$ , an approximation to Equation (3) can be computed by the multivariate logit model (Bock, 1972, 1975):

$$P(Y = k | \boldsymbol{\eta}) = \frac{\exp(\tau_k + \boldsymbol{\lambda}'_k \boldsymbol{\eta})}{\sum_{j=1}^K \exp(\tau_j + \boldsymbol{\lambda}'_j \boldsymbol{\eta})} \quad (4)$$

In practice, response probabilities are computed using Equation (4) for mathematical simplicity. McFadden (1974) showed that the approximation (4) to (3) is exact when the errors,  $e_k$ , follow the extreme-value distribution.

Model parameters are unidentifiable because the probabilities depend on differences between utilities, and if a constant value is added to all the utilities, the

differences remain unaltered. To remove this arbitrariness, one of the categories (say category  $K$ ) is used as a reference and its parameters are set to zero ( $\tau_K = 0$ ,  $\lambda_K = \mathbf{0}$ ). This constrains uniquely identifies the parameters for the other categories, which are interpreted as log-odds against the reference category:

$$\log \frac{P(Y = k | \boldsymbol{\eta})}{P(Y = K | \boldsymbol{\eta})} = \tau_k + \lambda'_k \boldsymbol{\eta} \quad (5)$$

The intercept parameter,  $\tau_k$ , determines the log-odds of categories  $k$  and  $K$  for an individual with factor scores  $\boldsymbol{\eta} = \mathbf{0}$ . The loadings vector  $\lambda_k$  is the rate of change of the log-odds in relation to the individual parameter  $\boldsymbol{\eta}$ . The property (5) is called *independence from irrelevant alternatives* (Luce, 1959), and means that the odds of choosing  $k$  over  $K$  do not depend on the other alternatives in the item. Thissen et al. (2010), Falk and Cai (2016) and Falk and Ju (2020) developed a newer parameterization to facilitate interpretation and to implement particular cases of the model between an unconstrained nominal model and a model for ordinal responses (Muraki, 1992). The new parameterization decomposes the slopes of the categories into a set of parameters that represent the the implicit ordering of the categories and a vector of item slopes.

### Removing latent indeterminacy in exploratory nominal factor analysis

The distribution of the latent factors is assumed to be

$$\boldsymbol{\eta} \sim \text{mnormal}(\boldsymbol{\kappa}, \boldsymbol{\Phi}), \quad (6)$$

where  $\boldsymbol{\kappa}$  is the mean of the latent factors, and  $\boldsymbol{\Phi}$  is its variance-covariance matrix. The scale of the factor scores and rotational indeterminacy shall be fixed before estimating the model because otherwise there would be infinite modes of the likelihood function, and the iterative estimation process may not converge. More details about theses issues can be seen in Revuelta and Ximénez (2017) and Revuelta et al. (2020).

Let  $p$  be the number of items,  $\boldsymbol{\tau}$  the vector of  $p(K - 1)$  item intercepts, and  $\mathbf{\Lambda}$  the loadings matrix with  $p(K - 1)$  rows and  $D$  columns (note that this is a slight redefinition of  $\mathbf{\Lambda}$ , which in Equation 3 was the matrix of item slopes and here it is the

matrix of slopes for all the items). Response probabilities remain invariant under linear translation and rotation of the factor scores, and according to the equations:

$$\begin{aligned}\boldsymbol{\eta}^* &= \mathbf{L}(\boldsymbol{\eta} - \mathbf{c}) \\ \boldsymbol{\tau}^* &= \boldsymbol{\tau} + \boldsymbol{\Lambda}\mathbf{c} \\ \boldsymbol{\Lambda}^* &= \boldsymbol{\Lambda}\mathbf{L}^{-1},\end{aligned}\tag{7}$$

where  $\mathbf{c}$  is a vector of constants and  $\mathbf{L}$  is an orthogonal matrix of order  $D \times D$ . This transformation of parameters does not alter the response probability because the exponential term in Equation (4) remains the same:

$$\begin{aligned}\boldsymbol{\tau}^* + \boldsymbol{\Lambda}^*\boldsymbol{\eta}^* &= (\boldsymbol{\tau} + \boldsymbol{\Lambda}\mathbf{c}) + \boldsymbol{\Lambda}\mathbf{L}^{-1}\mathbf{L}(\boldsymbol{\eta} - \mathbf{c}) \\ &= \boldsymbol{\tau} + \boldsymbol{\Lambda}(\boldsymbol{\eta} - \mathbf{c})\end{aligned}\tag{8}$$

The practical consequence of this indeterminacy is that  $\boldsymbol{\kappa}$  and  $\boldsymbol{\Phi}$  are arbitrary because they can be transformed to any other values of  $\boldsymbol{\kappa}$  and  $\boldsymbol{\Phi}$  by an appropriate choice of  $\mathbf{c}$  and  $\mathbf{L}$  without altering model fit. The most common solution to this problem is to assume that factors are independent and standardized (i.e.,  $\boldsymbol{\kappa} = \mathbf{0}$  and  $\boldsymbol{\Phi} = \mathbf{I}$ ) and that  $\boldsymbol{\Lambda}$  is a lower triangular matrix; that is, the first  $d - 1$  loadings are set to zero for factors  $d = 2, \dots, D$  (Erosheva & Curtis, 2017; Geweke & Zhou, 1996).

These constraints in  $\boldsymbol{\kappa}$ ,  $\boldsymbol{\Phi}$ , and  $\boldsymbol{\Lambda}$  remove indeterminacy in the origin and rotation of the factors without putting any restriction in the model because they do not affect the response probabilities. The lower triangular  $\boldsymbol{\Lambda}$  is just a simple method to avoid rotational indeterminacy during estimation, although this structure may not be particularly interesting for interpreting the meaning of the factors. A more interpretable solution, orthogonal or oblique, may be obtained by applying standard rotation methods to the estimated  $\boldsymbol{\Lambda}$  (Browne, 2001).

Note that these constraints still leave the problem of reflection indeterminacy unresolved. That is, the model remains the same if the factor scores and the slopes are multiplied by -1. However, this indeterminacy can be easily resolved after estimation by inverting the sign of the slopes if desired.

### Multi-group factor analysis for nominal data and levels of invariance

The multi-group MNCM assumes that the population of individuals is divided into  $G$  groups. The parameters for group  $g = 1, \dots, G$  are  $\boldsymbol{\tau}^{(g)}$ ,  $\boldsymbol{\Lambda}^{(g)}$ ,  $\boldsymbol{\kappa}^{(g)}$  and  $\boldsymbol{\Phi}^{(g)}$ , whereas  $\boldsymbol{\eta}^{(g)}$  is a random effect that is integrated out in the process of estimation.

A multi-group analysis is useful to compare factorial structures and test hypotheses of parameter equality across groups (Raju, Laffitte, & Byrne, 2002). However, these tests need to be performed in a specific order to obtain meaningful results. For example, it is obvious that factor means cannot be compared across groups if the meaning of the factors varies. For these reasons, tests of equivalence are arranged in a sequence of nested models with increasingly strong equivalence levels (Vandenberg & Lance, 2000) as follows:

- *No invariance* (NI). The factorial structure differs between groups, and all parameters are different. This is the baseline model, and it is useful to test hypotheses of invariance by comparing its fit with the one of a more constrained model.
- *Configural invariance* (CI). The same factor structure holds across groups. The matrices  $\boldsymbol{\Lambda}^{(1)}, \dots, \boldsymbol{\Lambda}^{(G)}$  have the same number of columns (factors), and fixed elements are located in the same positions. However, the elements of  $\boldsymbol{\Lambda}^{(g)}$  may vary from one group to another.
- *Weak invariance* (WI). The factor loadings are equal across groups,  $\boldsymbol{\Lambda}^{(1)} = \dots = \boldsymbol{\Lambda}^{(G)}$ . This is a necessary condition for  $\boldsymbol{\eta}$  to have the same interpretation for each group.
- *Strong invariance* (SI). Factor loadings and intercepts ( $\boldsymbol{\tau}^{(1)} = \dots = \boldsymbol{\tau}^{(G)}$ ) are equal across groups. This is a requisite for comparing factor means across groups, because it ensures that a higher mean is associated with higher log-odds.

A multi-group MNCM with external covariates is a type of model known as MIMIC (multiple indicator, multiple cause), and it regresses factor scores on a vector of manifest exogenous variables,  $\boldsymbol{x}$ , by the linear equation

$$\boldsymbol{\eta}^{(g)} = \boldsymbol{\alpha}^{(g)} + \boldsymbol{\gamma}^{(g)}\mathbf{x} + \mathbf{h}^{(g)}, \quad (9)$$

where  $\boldsymbol{\alpha}^{(g)}$  and  $\boldsymbol{\gamma}^{(g)}$  are regression intercepts and slopes, and  $\mathbf{h}^{(g)}$  is a random error that follows a normal distribution with mean  $\mathbf{0}$  and diagonal variance-covariance matrix  $\boldsymbol{\Psi}^{(g)}$ . The off-diagonal elements of  $\mathbf{h}^{(g)}$  are zero and the diagonal contains the error variances that are estimated from the data. Regarding the MIMIC model, three levels of invariance can be compared:

- *No invariance.* The regression intercepts and slopes vary from one group to another.
- *Invariance of slopes.* The regression slopes,  $\boldsymbol{\gamma}^{(g)}$ , are equal across groups. It requires WI to set a common unit of measurement for the latent factors across groups.
- *Invariance of slopes and intercepts.* Both the regression intercepts and slopes are equal across groups. This level of invariance requires SI; otherwise it would be indeterminate if the differences in response frequencies across groups were due to the item intercepts or to the regression intercepts.

### Inferential aspects

In this paper, the inferential algorithms will be restricted to those included in the Mplus (version 8) computer program so that the readers can use them as a reference for their own potential applications. Bayesian estimation algorithms have also been proposed in the psychometric literature to overcome some of the limitations of the classic methods (Revuelta & Ximénez, 2017) but they will be not used in this paper to keep the methods within the capabilities of Mplus.

### *Parameter estimation and likelihood ratios*

The MNCM is estimated from the matrix of responses of *individuals*  $\times$  *items* using a marginal maximum-likelihood/EM estimation algorithm (Bock & Aitkin, 1981). The marginal/EM algorithm requires numerical integration over the latent factor space, which limits the number of latent factors that can be estimated to about seven.

The output provided by Mplus when fitting latent variable models for nominal data includes the maximum value of the log-likelihood, the number of free parameters, and the AIC, BIC and BICc statistics (Vrieze, 2012). AIC and BIC are the Akaike information criterion and the Bayesian information criterion, and BICc is the BIC statistic adjusted by sample size proposed by Sclove (1987). The AIC, BIC and BICc have been compared in the context of factor variable models by Tein, Coxe, and Cham (2013), Chen, Luo, Palardy, Glaman, and McEnturff (2017), and Dziak, Coffman, Lanza, Li, and Jermin (2020).

The likelihood-ratio chi-square is computed as  $-2$  times the difference in the log-likelihood values of the two models being compared. Likelihood-ratio absolute fit of the fitted model against a saturated model is in general not applicable because of the exponential growth rate of the number of parameters for the saturated model. The saturated model is a multinomial distribution with unrestricted parameters, that are the probabilities of the different response patterns that can be given to the questionnaire. The number of parameters for the saturated model is  $K^I - 1$  for a questionnaire with  $I$  items and  $K$  categories. In real applications there is usually a huge number of parameters, and the saturated model cannot be estimated. Absolute model fit testing is applicable only with very short questionnaires.

The test of relative fit assumes that the mathematical form of the model is correct and uses the likelihood-ratio chi-square to compare models with different number of latent factors in exploratory analyses, or models with different levels of invariance in a multi-group analysis (Bentler, 1990).

### ***Estimating the number of latent factors***

In an exploratory analysis, the number of latent factors could be estimated by comparing the fit of the model with  $m$  factors against the model with  $m + 1$  factors until the likelihood-ratio chi-square statistic is non-significant. However this method of estimating the number of latent factors is not advisable for reasons to be explained below in this paper. Instead, the AIC, BIC and BICc indices can be used to compare exploratory models with different number of factors, and select the most appropriate

model for the sample data. Other popular methods for testing latent dimensionality, such as parallel analysis or the RMSEA and its test of close fit, are not currently available for the MNCM.

### ***Multi-group Comparisons***

The likelihood ratio tests can also be applied in a multi-group analysis to compare models at different levels of factorial invariance. In our experience, this method is reliable in the investigation of invariance. However, the likelihood-ratio chi-square requires to estimate the model twice, with and without the parameter constraint to be tested. In many cases, a simpler approach based on a single estimation is more practical.

Contrast tests allow for great flexibility in the comparison of parameters; for example, to test the invariance item-by-item. Parameters can be compared within an item, from one item to another, and across groups. The null hypothesis for a contrast test is that a linear combination of item parameters is zero. Two contrast tests that will be applied below in this paper are:

$$\begin{aligned} H_{0(A)} : L_A &= \lambda_1^{(1)} - \lambda_2^{(1)} = 0 \\ H_{0(B)} : L_B &= (\tau_1^{(1)} - \tau_2^{(1)}) - (\tau_1^{(2)} - \tau_2^{(2)}) = 0 \end{aligned} \tag{10}$$

The test of  $H_{0(A)}$  evaluates whether the slopes of categories 1 and 2 of a given item for the group 1 are equal. The hypothesis  $H_{0(B)}$  compares the difference between the intercepts of two categories in Group 1 with the difference of the same categories in Group 2.

Contrast tests can be easily implemented in Mplus using the MODEL CONSTRAINT command. The output of Mplus includes the test statistic, the standard error, and the p-value for such a hypothesis. The test statistic provided by Mplus is the standardized estimated linear combination:  $Z_{\hat{L}} = \hat{L}/S_e(\hat{L})$  and the  $p$ -value is taken from a standard normal distribution.

### Simulation study

A simulation study was designed to investigate the performance of the goodness-of-fit statistics in estimating the number of latent factors and detecting different degrees of model invariance between groups. As a secondary aim, we investigated the recovery of the parameters.

#### Method

The simulation scenarios consisted of two groups with different levels of invariance. The number of items was 10, with three categories per item to recreate a realistic situation. The simulating model was one-dimensional. There are no external covariates (i.e., the MIMIC model) in the simulation study. The true values of  $\eta$  were taken from a standard normal distribution for each simulated data matrix.

The true parameter values appear in Table 1. The symbols  $\Delta\tau$  and  $\Delta\lambda$  in Table 1 indicate the difference in the parameter values between the groups.  $\Delta\tau$  and  $\Delta\lambda$  were set to 0 for Group 1 and take a non-negative value for Group 2. In the condition of model invariance,  $\Delta\tau = \Delta\lambda = 0$  for Group 2. In the conditions where model invariance fails,  $\Delta\tau$  or  $\Delta\lambda$  are positive quantities. In sum,  $\Delta\tau$  and  $\Delta\lambda$  are population effect sizes that represent the difference between groups.

\*\*\* Insert table 1 \*\*\*

The simulated conditions were:

- *Total sample size*,  $N = 200, 400, 800$ , divided into two groups of equal size so that group sample size is 100, 200, and 400.
- The *effect size for parameter  $\tau$*  took the values  $\Delta\tau = 0, 0.1, 0.2, \dots, 1.5$ .  $\Delta\lambda$  was set to 0.
- *Effect size for parameters  $\tau$  and  $\lambda$* , with values  $\Delta\tau = \Delta\lambda = 0.1, 0.2, \dots, 1.5$ .

The total number of simulated conditions was

$$3(\text{sample sizes}) \times [16(\text{values of } \Delta\tau) + 15(\text{values of } \Delta\tau \text{ and } \Delta\lambda)] = 93$$



The values of  $\Delta\tau$  and  $\Delta\lambda$  were manipulated smoothly to obtain the graphical representation of the power curve that indicates the sensitivity of the chi-square, and the values of AIC, BIC and BICc associated with different levels of invariance. The number of simulated samples for each condition was 200.

### **Analytic strategy**

The following models were estimated for each sample:

1. Configural invariance with one factor. This model assumes that all parameters  $\tau$  and  $\lambda$  vary between groups. The configural invariance model includes the simulated model as a particular case in all conditions of the simulation.
2. Configural invariance with two factors. This model was fitted to evaluate the type-I error rate when comparing fit against a model with unnecessary latent factors and to investigate whether invariance can be confounded with multi-dimensionality.
3. Model of weak invariance. The loadings are held constant across groups but the intercepts are allowed to vary. It is equivalent to the generating model in the conditions with  $\Delta\lambda = 0$ , and more constrained than the generating model when  $\Delta\lambda > 0$ .
4. Model of strong invariance. It assumes equal loadings and intercepts across groups. It is equivalent to the generating model when  $\Delta\tau = \Delta\lambda = 0$ , and a too restrictive model otherwise.

The analyzed statistics were those provided by Mplus:

1. Likelihood ratio chi-square statistic. Model 1 was compared against Model 2, Model 3 against Model 1, and Model 4 against Model 1. The interpretation of a significant chi-square depends on the models being compared. Models 1 and 2 are correct for all the conditions because the simulating model falls within their parameter space. Models 3 and 4 are incorrect in general as they are more constrained than the simulating model. In the comparison 1-2, a significant

chi-square was interpreted as a type-I error that consists of rejecting the true common-factor model in favor of a model with an spurious factor. In the comparisons 3-1 and 4-1, a significant chi-square is interpreted as an indication of statistical power. Note that in comparison 1-2, both models were always correct and the comparison gave an estimate of the type-I error rate when comparing a one-dimensional model against a two-dimensional model with a spurious latent factor. In comparisons 3-1 and 4-1, the Models 3 and 4 were generally wrong and were tested against Model 1, which was always correct; thus, these comparisons inform about statistical power.

Type-I error rate and statistical power were estimated by the empirical proportion of rejection (EPR), which is the proportion of samples in which the  $p$  – value associated with the chi-square statistic is smaller than or equal to the nominal level ( $\alpha = 0.05$ ).

2. Information measures: AIC, BIC, and BICc. For each condition, the model that minimized any given information measure was the model selected by that measure. The empirical proportion of selection (EPS) was the proportion of samples in which each of the four models minimized the AIC. The EPS associated to BIC and BICc were also computed.
3. Contrast tests were computed in relation to Model 1 to investigate the type-I error rate and power curves associated with these tests. Table 2 summarizes the contrast tests. The column labeled *status of  $H_0$*  indicates whether the EPR associated with theses tests represents type-I error or statistical power. For example, Test  $L(1)$  compared the intercepts of item 1, Category 1, across groups; because these intercepts are always equal in the simulated conditions,  $H_0$  is always true, and the EPR is an estimate of type-I error. Test  $L(3)$  compared the intercepts of item 6, Category 1, across groups; because these intercepts are different when  $\Delta\tau > 0$ , the EPR for the conditions with  $\Delta\tau = 0$  estimates type-I error, and the EPR for the conditions with  $\Delta\tau > 0$  estimates statistical power.

\*\*\* Insert table 2 \*\*\*

## Results of the simulation study

### *Results for chi-square*

Results for the likelihood-ratio chi-square are summarized by the graphical representation of the EPR as a function of effect size (Figure 1), which is an estimate of the power curve. The upper panel represents the conditions with  $\Delta\tau = 0, 0.1, \dots, 1.5$  and  $\Delta\lambda = 0$ , and the lower panel stands for  $\Delta\tau = \Delta\lambda = 0, 0.1, \dots, 1.5$ . The green lines are the EPR for the comparison of Models 1 and 2. Because the simulated model has one factor, the green lines are estimates of the type-I error rate for the model with one factor against a model with a spurious factor. In this regard, the type-I error rate was about 0.9 for all conditions, clearly indicating that the chi-square was extremely liberal and little reliable for estimating the number of factors for an MNCM in these conditions.

The blue lines represent the EPR for the comparison between Models 3 and 1. In the upper panel, these lines are type-I error rate when comparing the true weak invariance model with a model that unnecessarily allows group differences both in the intercepts and in the slopes (i.e., the configural model). The results showed that chi-square was too liberal with the small sample size, with an EPR about 0.2, and approached the nominal level as the sample size increased. Thus, in real applications with small samples, there is a danger of erring in the direction of rejecting a true model of weak invariance in favor of a model with unnecessary free slopes. In contrast, the blue lines in the lower panel are an estimate of statistical power, as this time, slopes varied across groups and therefore Model 3 was false. The EPR increased sharply in relation to the effect size, approaching 1 for values of  $\Delta$  between 0.5 and 1, depending on sample size.

The red lines represent the comparison between Model 4 (which was true only when  $\Delta\tau = \Delta\lambda = 0$ ) and Model 1. Because Model 1 was true in all conditions, the red lines are power curves. The EPR showed a sharp increase of statistical power, which reached values close to 1 for an effect size about 0.5.

All in all, the chi-square was very sensitive to detect model violations. When it

comes to test model dimensionality, the chi-square lacks its utility because of its extremely liberal tendency. However, the chi-square proved to be useful to detect differences about 0.5 points or more between the model parameters of the two groups.

\*\*\* Insert figure 1 \*\*\*

### ***Results for AIC, BIC, and BICc***

The EPS results for the conditions with  $\Delta\lambda = 0$  appear in Figure 2. The upper, middle, and lower panels refer to AIC, BIC, and BICc, respectively. The three statistics select the true model with high frequency, whereas Models 1 and 2 are seldom selected. When the effect size is low, the model of strong invariance is selected, and the model of weak invariance is selected for effect sizes of about 0.5 and above. The AIC was more liberal than the BIC and selects the weak invariance model for smaller effect sizes than the BIC and BICc.

\*\*\* Insert figure 2 \*\*\*

The results for  $\Delta\tau = \Delta\lambda > 0$  appear in Figure 3. The three statistics select the strong invariance model for low or no effect size, the (incorrect) weak invariance model for intermediate effect sizes, and the configural model for effect size above 1. The two-factor model is seldom selected. The AIC is more liberal than the two variants of the BIC and has a tendency towards more sophisticated models.

\*\*\* Insert figure 3 \*\*\*

### ***Results for contrast tests***

**Contrast tests for  $\Delta\lambda = 0$ .** Contrast tests were computed to investigate the type-I error rate and statistical power for Model 1. The  $H_0$  for the contrast tests  $L(3)$  and  $L(6)$  is false when  $\Delta\tau = 0$ , and its EPR is an estimate of statistical power. The null hypotheses for  $L(3)$  and  $L(6)$  are equivalent because  $H_{0,L(3)} : \tau_{611} - \tau_{612} = 0$  and  $H_{0,L(6)} : (\tau_{611} - \tau_{612}) - (\tau_{621} - \tau_{622}) = 0$ , where  $\tau_{611} - \tau_{612} = \Delta\tau$  and  $\tau_{621} - \tau_{622} = 0$ . Thus,  $L(3)$  and  $L(6)$  differ not in the status of the hypothesis but in the test statistic. The  $H_0$  for the other contrast test is true as long as  $\Delta\lambda = 0$ , and hence the EPR is type-I error.

Figure 4 contains the EPR for all the contrast tests as a function of effect size. The figure shows that both  $L(3)$  and  $L(6)$  reached high statistical power for moderate effect size but the power curve for  $L(6)$  rose sharply. The EPR remained close to the nominal type-I error for all the conditions when  $H_0$  was not rejected.

\*\*\* Insert figure 4 \*\*\*

**Contrast tests for  $\Delta\lambda > 0$ .** In the conditions with  $\Delta\tau = \Delta\lambda > 0$ , the  $H_0$  for the contrast tests  $L(3)$ ,  $L(6)$ ,  $L(8)$ , and  $L(11)$  is false. The  $H_0$  is equivalent for  $L(8)$  and  $L(11)$ , and these contrasts differ in the test statistics. The  $H_0$  for the other contrast tests is true, and its EPR estimates Type-I error.

Figure 5 shows that the contrast tests for the slopes ( $L(8)$  and  $L(11)$ ) had less power than the contrast test for the intercepts ( $L(3)$  and  $L(6)$ ). The general trend was that contrast tests performed well, with an EPR close to the nominal alpha-level when the null hypothesis was true and detected differences in model parameters when they existed.

\*\*\* Insert figure 5 \*\*\*

### ***Parameter recovery***

The simulation study also provided information about the recovery of true parameter values in realistic conditions regarding sample size and number of items. Recovery was evaluated by the mean bias and the root mean squared error (RMSE) between the true and estimated parameters for Model 1.

The means and standard deviations of the recovery statistics appear in Table 3. In order to keep the presentation to a minimum, the results were averaged over parameter types. In this sense, recovery was better for intercepts than for slopes. Moreover, the simulation indicated that at least 400 individuals seem necessary to obtain interpretable item parameters because of the poor recovery with smaller samples.

\*\*\* Insert table 3 \*\*\*

### **Conclusions of the simulation study**

In summary, the goodness-of-fit statistics did a good job of detecting different levels of invariance. While the chi-square statistic was sensitive to violations of the assumption of invariance, the information statistics (AIC, BIC, and BICc) successfully identified the correct model, with a slight tendency for the AIC to select a too complex model. In addition, the contrast tests rendered reliable results about the item parameters that differ between groups and kept the type-I error rate at the nominal level when no such differences exist.

However, our results showed that the estimation of the number of factors cannot be based on chi-square. This statistic showed an extremely liberal tendency and selected an over-factored model in about 90% of the samples. Fortunately, the information statistics did not share this drawback and penalized the two-factor model in comparison to the models with one factor. All in all, chi-square is useful for testing model invariance, but when it comes to evaluating dimensionality, the decision should be based on information statistics.

### **Empirical study**

An empirical study was conducted to test the hypotheses presented in Section 1 with a sample of real data and to put the recommendations extracted from the simulation study into practice.

#### **Method**

Two samples of 200 individuals were recruited, making a total of 400 participants, conforming a representative sample of young Spanish men and women. Each sample contains 100 men and 100 women who responded to one of the SGS forms (MG or AG), which were already explained in Section 1. The ages of the participants ranged from 17 to 30 years. The test was administered online through the Qualtrics platform. The presentation order of both the items and the categories was randomized. The sample contains no missing data because the online platform does not allow omitting responses. The recruitment of the participants was done through social networks like Twitter and

Instagram, as they are the most accessible media for our population of interest. The recruitment method allowed obtaining data from a less restricted sample, as participants could respond from different locations within Spain, so the results could be more generalizable. However, recruitment was limited to the Spanish participants. Each participant was randomly assigned to one of the two conditions (MG or AG), controlling that gender was balanced between both forms of the test. The conditions in the multi-group analysis (MG and AG) refer to test form, not to men and women. The purpose of collecting these data was to test the hypotheses previously presented in an empirical scenario.

### **Analytic strategy**

Five data analyses were conducted on the SGS data:

1. Preliminary *descriptive analyses* were used to obtain initial information about the data before running the inferential methods.
2. An *exploratory nominal factor analysis* (Revuelta et al., 2020) was conducted to identify the number of factors and to gather evidence about the theoretical assumption that the SGS is a one-dimensional instrument.
3. *Measurement invariance* between the two generic forms of the SGS (MG and AG) was tested to identify the level of invariance between the forms (Vandenberg & Lance, 2000). Supposedly, the use of one generic or another should not modify the construct measured by the test, and the data may thus support at least the level of weak invariance. The purpose of the analysis is to evaluate this hypothesis.
4. The third analysis consisted of delving further information by applying *contrast tests* to compare both forms of the SGS item-by-item. In other words, intercept and factor loadings for each item were compared between the two forms.
5. Lastly, factor scores were regressed on gender and age using the MIMIC model (Jöreskog & Goldberger, 1975).

### **Results of the empirical study**

***Descriptive analysis***

As can be seen in the bar-plot of Figure 6, at a descriptive level, there were little differences between the observed response frequency of the categories in the two groups. Specifically, the female-category had a higher frequency in the AG condition than in the MG condition for Item X7, whereas the reverse pattern occurred in Item X8, where the male-category had a higher frequency when using MG. The differences between conditions were small in the rest of the items.

\*\*\* Insert figure 6 \*\*\*

Regarding the covariable Age, the mean was 22.0, the standard deviation was 3.27, and the quartiles were  $Q_1 = 20$ ,  $Q_2 = 21$ , and  $Q_3 = 24$ . Recall that, under the design of the data-gathering procedure, the proportion of men and women in the sample was fifty-fifty.

***Exploratory nominal factor analysis to identify the number of factors***

Exploratory nominal factor analysis models with one, two, and three factors were fitted to the SGS data. Model fit was compared by the likelihood ratio chi-square, and the AIC, BIC and BICc indices. The results appear in Table 4. In this regard, significant likelihood-ratio chi-square statistics were found between the one-factor and two-factors models, but not between the two- and three-factors models. However, the simulation study revealed that chi-square is too liberal, and information statistics are more reliable to estimate dimensionality. The model with two factors minimized the AIC, but the difference in AIC between the one-factor and the two-factor models was smaller than 2, which is the value suggested by Burnham and Anderson (2004) as an indication of substantial difference between AIC values. The BIC and BICc also supported the one-factor model. All in all, there was some evidence that there could be two factors underlying the item responses. However, this evidence was not strong and we retained the one-factor model for further interpretation of the results. A larger sample could provide further information about this issue.

\*\*\* Insert table 4 \*\*\*



*Measurement invariance*

The three levels of invariance (CI, WI, and SI) were compared for the one-factor model. The goodness-of-fit statistics appear in Table 5, and indicate that the chi-square statistic was not significant for the comparison between the CI and WI levels, but it was significant when comparing the WI and SI levels. In addition, the AIC favored the WI model, whereas the BIC and BICc indexes suggested the SI model.

\*\*\* Insert table 5 \*\*\*

Item parameters and response functions were analyzed to investigate the differences between generic forms in more detail. Figure 7 contains the path diagram and item parameters for the WI model. We have slightly modified the conventions for a path diagram to accommodate the peculiarities of the nominal variables. In this respect, the path that goes from the factor to each item splits into two arrows, one for each of the categories that have estimated parameters. The arrow of each category is labeled with the slope and the intercept (in brackets). The parameters for the third category (the neutral option) are structural zeros and are not represented in the path diagram.

\*\*\* Insert figure 7 \*\*\*

The item response functions in Figure 8 represent the probabilities of the categories as a function of the factor score. The solid lines correspond to the MG group, and the dotted lines stand for the AG group. Similarly to the descriptive analyses, the figure reveals not substantial differences between groups except for items X7 and X8. Also, some small differences appeared for items X2, X3, X6, and X9. For both items X7 and X8, the male-category was selected more frequently in the MG group than in the AG group, suggesting that these items may contain a relevant feature that elicits a masculine stereotype.

\*\*\* Insert figure 8 \*\*\*

*Contrast tests*

The preceding analyses indicate that there is not a general lack of invariance between the generic forms, but there could be differences related to specific items. We have applied contrast tests to further clarify these issues and compare forms item-by-item. Four contrast tests were applied to each item:

- The intercept associated to the male-category in the MG form was compared to the same intercept in the AG form ( $L_{Mj}$ ).
- The intercept for the female-category in the MG form was compared to the same intercept in the AG form ( $L_{Fj}$ ).
- Difference between the intercept for the male and female category of each item was compared across the MG and AG forms ( $L_{Dj}$ ).
- The loading of the male and female categories of each item was compared ( $L_{Lj}$ ).

Table 6 summarizes the comparisons involved in these tests.

\*\*\* Insert table 6 \*\*\*

The total number of comparisons was 40, so the correction criteria for the Benjamini-Hochberg type-I error rate (Benjamini & Hochberg, 1995) was applied using the `p.adjust()` function from the *stats* package in R language.

The estimates of the comparisons and their corrected p-values are shown in Table 7. No significant differences appeared for the comparison of the male category between the MG and AG forms ( $L_{Mj}$ ) except for Item X8. The female category did not present significant differences across forms ( $L_{Fj}$ ). However, the comparison  $L_{Dj}$  was significant for items X7, X8, and X9. More specifically, the difference between the male and female intercepts was higher in the MG form than in the AG one.

The comparison between the factor loadings of the male and female categories of each item ( $L_{Lj}$ ) was significant for all the items, except for the last three (X7, X8, and X9). The sign of the slopes indicated that positive factor scores were associated with a higher propensity to the male stereotyped categories, factor scores about zero were

associated with the neutral categories, and the female categories were selected at negative factor scores.

\*\*\* Insert table 7 \*\*\*

### ***MIMIC model: gender and age***

The MIMIC model regressed factor score on gender and age. The analysis was run for the tree levels of MIMIC invariance described in Section 2.3. The likelihood ratio chi-square statistics were non-significant, and the simplest model of invariance of slopes and intercepts was retained. This conclusion is supported also by the AIC, BIC and BICc. The results appear in table 8.

\*\*\* Insert table 8 \*\*\*

Regarding the model of invariance, the slope for age was -0.027 ( $p = 0.315$ ), and for gender it was -0.264 ( $p = 0.110$ ). Thus, we did not find a significant effect for the covariates.

### **Conclusions of the empirical study**

The MNCM allowed us to reach several substantive conclusions about the SGS. First, this model enabled us to study the dimensionality of the SGS, suggesting a one-factor model that can serve as evidence of construct validity. Second, the hypothesis of factorial invariance for both generic forms of the questionnaire (MG and AG) was not rejected on the basis of our analyses. The factorial structure includes a latent factor, which can be interpreted as the male over-representational bias. Indeed, factor loadings remained unchanged between forms, indicating that the relation between categories and the latent factor is not altered by the generic form, as stated in *Hypothesis 1*. However, the intercepts of the categories showed differences for some specific items because of a different response frequency from one form to another. These differences are related to stereotypically masculine activities, which were selected more frequently when the item is written in masculine generic, thus retaining our *Hypothesis 2*. These results conform with the theory supporting the SGS. In addition, the MNCM

provides an alternative approach to the study of invariance that has allowed drawing substantive conclusions regarding the initial hypotheses.

Contrast tests provided more detailed item-by-item information about the differences between the forms of the questionnaire. The simulation study has shown that increased statistical power is achieved when comparing the differences between the two item intercepts from one form to another. The analysis showed little differences between the generic forms, except for three items: X7, X8, and X9. This could be related to the fact that these items were designed based on gender *roles* (i.e., "beliefs about the activities considered more appropriate for men and women" López-Sáez et al. 2008), whereas the first five represented differences between gender *traits* (i.e., psychological characteristics expected for each gender). From a substantive point of view, these differences between situations that represent *roles* and *traits* could be decisive in testing the effect of the use of generics in the situations presented in the SJT. These results are in line with the independence found between these two gender-stereotype components in López-Sáez et al. (2008). However, it was not among the objectives of this work to deepen the study of the differences between these two social aspects, and future research will be necessary to draw more concrete conclusions about this potential phenomenon. Another interesting fact to highlight in this analysis is that the sign of the estimates among male-categories is positive for most of the items, whereas for the female-categories, it is negative. This indicates that, in general, the male-categories are more frequently selected than the female-categories in the MG as compared to the AG form. These results suggest that the use of alternative generics reduces stereotypically male expectations of mixed groups, and are congruent with previous research on this topic (e.g., Kaufmann & Böhner, 2014).

Item-by-item analyses of factor loadings revealed a clear pattern. The slope of the male-category is higher than that of the female-category for all of the items. Then, the latent factor could be conceptualized as a continuum whose ends range from *preference for stereotypically female expectations for the group presented in the item* (low factor score) to *preference for stereotypically male expectations* (high factor score). These

conclusions can be interpreted as a source of construct validity evidence, as this description appropriately illustrates what was intended to be measured: male over-representational bias.

Finally, the part of the MIMIC model does not render significant results. No evidence of age- or gender-related differences are found in male over-representational bias. Regarding the covariate age, this could probably be due to a rather homogeneous sample, with ages between 17 and 30 years old. Furthermore, the application of the questionnaire via internet may also induce a selection bias by excluding participants with a lower educational level or fewer technological resources. Moreover, the linguistic use of alternative generics is somewhat more frequent among young and educated people. A more comprehensive study would have an explicit focus on collecting participants with a variety of ages and educational levels. However, we are not aware of previous empirical findings about these issues, as the vast majority of studies in this field are based on samples of students. Regarding the lack of a significant gender effect, previous evidence is diverse. Works such as those by Merritt and Kok (1995) and Nissen (2013) did not record significant differences between genders in their respective tests of the influence of MG, whereas Hamilton (1988) and Kaufmann and Bohner (2014) did find them. For these latter authors, the explanation of these differences refers to the bias "People = Me" described by Silveira (1980), which could explain that men, in addition to the bias "People = Men", manifested a higher male over-representation in mixed groups. In any case, it remains to be seen if these differences could appear in other alternative generics not covered in this study, such as the generic with "-x" or with "-e", which are increasingly used nowadays.

## Discussion

This paper addresses the problem of how to factor analyze the nominal responses originated from SJTs. A situational judgment item presents a hypothetical scenario and a list of actions, and the individuals are asked to select their most likely action for that scenario. Because actions have no explicit order, the item generates nominal responses

consisting of the actions selected by the individuals. We present procedures based on the MNCM (Revuelta et al., 2020), which is a multidimensional extension of the nominal categories model by Bock. The MNCM allows the analysis of the factorial structure for item responses that do not have an explicit order, as is the case of a SJT. Previous research has addressed the problem of factor analyzing nominal data in SJTs but has not considered the nominal categories model in the multidimensional case. The paper also demonstrates the application of the MNCM in exploratory and multi-group factor analysis from the nominal data provided by a SJT.

We present the results of two studies. First, a simulation study is conducted to investigate the statistical properties of the inferential methods. The likelihood-ratio chi-square, AIC, BIC, and BICc statistics showed good results in the recovery of the invariance level of the simulated data. However, the chi-square statistic was too liberal and, in practice, useless for estimating the number of factors for the MNCM. Based on these findings we recommend the use of the indices AIC, BIC, and BICc, which do not share this problem and, for the time being, can be used to guide the selection of the number of factors.

The simulation study is focused on investigating how the goodness-of-fit statistics behave depending on the magnitude of the difference between the item parameters in the two groups. However, the simulating model is one-dimensional in all the conditions. One topic that requires further investigation is to compare results for simulating models of different dimensionality and factorial structure.

The second study is a real data example that serves as an illustration of how the MNCM can be used to factor analyze nominal responses from an SJT about gender stereotypes. The purpose of the study is to investigate latent dimensionality and to test hypotheses about the factorial structure and parameter values. Results of exploratory analyses showed that responses to the SJT are associated to a single latent factor. Concerning the analysis of the factorial invariance of the questionnaire, consisting of the comparison of the factor structure in two groups of individuals who received a different test form, the results supported the weak invariance level. From a substantive point of

view, the procedure allowed us to draw conclusions concerning the use of generics and their relation with stereotypically expectations.

The contribution of the paper is largely based on its focus on the SJT and the consideration of the MNCM to study its dimensionality. Previous research has only focused on the assignation of arbitrary scores to the response categories of an SJT, assuming that these scores are suitable for defining the dimensions of the SJT. A notable step further is that of Zu and Kyllonen (2018), who consider the use of a one-dimensional nominal categories model for SJT. To our knowledge, the extension to the multi-dimensional model and multi-group designs has not been attempted before. Future research should continue investigating the applicability of the MNCM to other item formats. For instance, multiple-choice items, that also generate nominal responses could be analyzed using the same procedures described here.

The main challenge with nominal data is that several response categories are grouped together to create an item, and the parameters of the categories are interpreted in relation to the other categories of the same item instead of comparing them to the parameters of the other items. Additional investigation is necessary to determine how the item parameters of the MNCM can be compared across items.

We chose the Mplus computer program to run all these analyses because of its widespread use for structural models. Notwithstanding, more research about Bayesian methods for the nominal model (Revuelta & Ximénez, 2017) would be welcome, due to the current relative scarcity of model-fit techniques in the classical framework. For example, Bayesian methods are applicable to monitoring the posterior distribution of goodness-of-fit statistics that are not implemented in Mplus. We are aware that the Bayesian analyses require computer coding using languages such as JAGS (Plummer, 2003) or RStan (Stan Development Team, 2018), which are not routinely used by many applied investigators. Therefore, in this paper, we limited our analyses to the possibilities provided by a computer program of general usage. We hope that in the near future this popular software also considers including the Bayesian procedures for the nominal model. We also hope that this research provides additional information to

assist researchers when conducting the task of fitting latent variable models to nominal data.



## References

- Benjamini, Y., & Hochberg, Y. (1995). Bayesian measures of model complexity and fit. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300. doi: <https://doi.org/10.1111/1467-9868.00353>
- Bentler, P. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238-246. doi: <https://doi.org/10.1037/0033-2909.107.2.238>
- 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. doi: <https://doi.org/10.1007/BF02291411>
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York, NY: McGraw-Hill.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4), 443-459. doi: <https://doi.org/10.1007/BF02293801>
- Bock, R. D., & Gibbons, R. (2010). Factor analysis of categorical item responses. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (p. 43-75). New York, NY: Taylor & Francis.
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36(1), 111-150. doi: [https://doi.org/10.1207/S15327906MBR3601\\_05](https://doi.org/10.1207/S15327906MBR3601_05)
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2), 261-304. doi: <https://doi.org/10.1177/0049124104268644>
- Chen, Q., Luo, W., Palardy, G. J., Glaman, R., & McEnturff, A. (2017). The efficacy of common fit indices for enumerating classes in growth mixture models when nested data structure is ignored: A monte carlo study. *SAGE Open*, 7(1), 1-19. doi: <https://doi.org/10.1177/2158244017700459>
- Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jeremiin, L. S. (2020). Sensitivity

- and specificity of information criteria. *Briefings in Bioinformatics*, 21(2), 553-565.  
doi: <https://doi.org/10.1093/bib/bbz016>
- Erosheva, E. A., & Curtis, S. M. (2017). Dealing with reflection invariance in Bayesian factor analysis. *Psychometrika*, 82(2), 295—307. doi:  
<https://doi.org/10.1007/s11336-017-9564-y>
- Fahrmeir, L., & Tutz, G. (1991). *Multivariate statistical modelling based on generalized linear models*. New York: Springer.
- Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, 328:347. doi:  
<https://doi.org/10.1037/met0000059>
- Falk, C. F., & Ju, U. (2020). Estimation of response styles using the multidimensional nominal response model: A tutorial and comparison with sum scores. *Frontiers in Psychology*, 11:72. doi: <https://doi.org/10.3389/fpsyg.2020.00072>
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, 7(3), 286-299. doi: <https://doi.org/10.1037/1040-3590.7.3.286>
- Geweke, J., & Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studie*, 9(2), 557-587. doi:  
<https://doi.org/10.1093/rfs/9.2.557>
- Hamilton, M. C. (1988). Using masculine generics: Does generic he increase male bias in the user's imagery? *Sex roles*, 19(11-12), 785-799. doi:  
<https://doi.org/10.1007/BF00288993>
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a), 631-639. doi:  
<https://doi.org/10.1080/01621459.1975.10482485>
- Kaufmann, C., & Bohner, G. (2014). Masculine generics and gender-aware alternatives in spanish. *IFFOnZeit—Online Journal of the Interdisciplinary Center for Research on Women and Gender at the University of Bielefeld*, 4(3), 8-17.

- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: a review of recent research. *Personnel Review*(37), 426-441. doi: <https://doi.org/10.1108/00483480810877598>
- López-Sáez, M., Morales, J. F., & Lisbona, A. (2008). Evolution of gender stereotypes in Spain: Traits and roles. *The Spanish Journal of Psychology*, 11(2), 609-617. doi: <https://doi.org/10.1017/S1138741600004613>
- Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.
- McDaniel, M. A., & Nguyen, N. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9(1-2), 103-113. doi: <https://doi.org/10.1111/1468-2389.00167>
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (p. 105–142). New York, NY: Academic Press.
- Merritt, R. D., & Kok, C. J. (1995). Attribution of gender to a gender-unspecified individual: An evaluation of the people = male hypothesis. *Sex Roles*, 33(3-4), 145-147. doi: <https://doi.org/10.1007/BF01544608>
- Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, 16(2), 159-176. doi: <https://doi.org/10.1177/014662169201600206>
- Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide. eighth edition*. Los Angeles, CA: Muthén & Muthén.
- Nissen, U. K. (2013). Is Spanish becoming more gender fair? a historical perspective on the interpretation of gender-specific and gender-neutral expressions. *Linguistik online*, 58(1). doi: <https://doi.org/10.13092/lo.58.241>
- Ployhart, R. E., & Ward, A.-K. (2013). Situational judgment measures. In K. F. Geisinger et al. (Eds.), *APA handbooks in psychology. APA handbook of testing and assessment in psychology, vol. 1. test theory and testing and assessment in industrial and organizational psychology* (p. 551-564). Washington, DC: American Psychological Association. doi: <https://doi.org/10.1037/14047-030>

- Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Retrieved from <http://mcmc-jags.sourceforge.net/>
- Raju, S., Laffitte, L., & Byrne, B. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87(3), 517-529. doi: <https://doi.org/10.1037//0021-9010.87.3.517>
- Revuelta, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, 38(7), 549-562. doi: <https://doi.org/10.1177/0146621614536272>
- Revuelta, J., Maydeu-Olivares, A., & Ximénez, C. (2020). Factor analysis for nominal (first choice) data. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(5), 781-797. doi: <https://doi.org/10.1080/10705511.2019.1668276>
- Revuelta, J., & Ximénez, C. (2017). Bayesian dimensionality assessment for the multidimensional nominal response model. *Frontiers in Psychology*, 8, 961. doi: <https://doi.org/10.3389/fpsyg.2017.00961>
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3), 333-343. doi: <https://doi.org/10.1007/BF02294360>
- Sharma, S., Gangopadhyay, M., Austin, E., & Mandal, M. (2013). Development and validation of a situational judgment test of emotional intelligence. *International Journal of Selection and Assessment*, 21(1), 57-73. doi: <https://doi.org/10.1111/ijsa.12017>
- Silveira, J. (1980). Generic masculine words and thinking. *Women's Studies International Quarterly*, 3, 165-178. doi: [https://doi.org/10.1016/S0148-0685\(80\)92113-2](https://doi.org/10.1016/S0148-0685(80)92113-2)
- Stan Development Team. (2018). *RStan: the R interface to Stan*. R package version 2.17.3. Retrieved from <http://mc-stan.org>
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory

- and factor analysis of discretized variables. *Psychometrika*, 52(3), 393–408. doi: <https://doi.org/10.1007/BF02294363>
- Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural equation modeling: A multidisciplinary journal*, 20(4), 640–657. doi: <https://doi.org/10.1080/10705511.2013.824781>
- Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory, vol. 1: Models* (p. 51–73). Boca Raton, FL: Chapman & Hall/CRC press.
- Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (p. 155–184). New York, NY: Taylor & Francis.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1), 4–70. doi: <https://doi.org/10.1177/109442810031002>
- Vrieze, S. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2), 228–243. doi: <https://doi.org/10.1037/a0027127>
- Zu, J., & Kyllonen, P. C. (2018). Nominal response model is useful for scoring multiple-choice situational judgment tests. *Organizational Research Methods*, 23(2), 342–366. doi: <https://doi.org/10.1177/1094428118812669>

Table 1  
*True Parameter Values for the Simulation Study*

Item	$\tau_1$	$\lambda_1$	$\tau_2$	$\lambda_2$
1	0.5	1.0	1.0	$0.5 + \Delta\lambda$
2	0.5	1.0	1.0	$0.5 + \Delta\lambda$
3	0.5	1.0	1.0	$0.5 + \Delta\lambda$
4	0.5	1.0	1.0	$0.5 + \Delta\lambda$
5	0.5	1.0	1.0	$0.5 + \Delta\lambda$
6	$0.5 + \Delta\tau$	1.0	1.0	0.5
7	$0.5 + \Delta\tau$	1.0	1.0	0.5
8	$0.5 + \Delta\tau$	1.0	1.0	0.5
9	$0.5 + \Delta\tau$	1.0	1.0	0.5
10	$0.5 + \Delta\tau$	1.0	1.0	0.5

Note:  $\Delta\tau$  and  $\Delta\lambda$  represent the difference between parameter values for the two groups.

Table 2  
*Contrast Tests for the Simulation Study*

Contrast	Comparison	Status of $H_0$
$L(1)$	$\tau_{111} - \tau_{112}$	True for any $\Delta\tau$
$L(2)$	$\tau_{121} - \tau_{122}$	True for any $\Delta\tau$
$L(3)$	$\tau_{611} - \tau_{612}$	True when $\Delta\tau = 0$ . False when $\Delta\tau > 0$
$L(4)$	$\tau_{621} - \tau_{622}$	True for any $\Delta\tau$
$L(5)$	$L(1) - L(2)$	True for any $\Delta\tau$
$L(6)$	$L(3) - L(4)$	True when $\Delta\tau = 0$ . False when $\Delta\tau > 0$
$L(7)$	$\lambda_{111} - \lambda_{112}$	True for any $\Delta\lambda$
$L(8)$	$\lambda_{121} - \lambda_{122}$	True when $\Delta\lambda = 0$ . False when $\Delta\lambda > 0$
$L(9)$	$\lambda_{611} - \lambda_{612}$	True for any $\Delta\lambda$
$L(10)$	$\lambda_{621} - \lambda_{622}$	True for any $\Delta\lambda$
$L(11)$	$L(7) - L(8)$	True when $\Delta\lambda = 0$ . False when $\Delta\lambda > 0$
$L(12)$	$L(9) - L(10)$	True for any $\Delta\lambda$

Note: The subindexes of  $\tau$  and  $\lambda$  refer to item, category, and group.

Table 3

*Recovery of the Parameters for the Configural Invariance Model*

Effect	Statistic	Parameter	$N = 200$	$N = 400$	$N = 800$
$\Delta\tau$	Bias	$\tau$	0.109(0.087)	0.034(0.025)	0.017(0.015)
		$\lambda$	0.118(0.083)	0.037(0.031)	0.018(0.019)
	RMSE	$\tau$	0.746(0.376)	0.284(0.052)	0.183(0.023)
		$\lambda$	1.001(0.309)	0.407(0.066)	0.255(0.028)
$\Delta\tau, \Delta\lambda$	Bias	$\tau$	0.099(0.050)	0.031(0.023)	0.014(0.013)
		$\lambda$	0.122(0.076)	0.040(0.034)	0.021(0.020)
	RMSE	$\tau$	0.652(0.219)	0.285(0.045)	0.186(0.019)
		$\lambda$	0.903(0.282)	0.400(0.062)	0.253(0.026)

Note: The table contains the mean of the recovery statistics. Standard deviations appear in brackets.



Table 4  
*Goodness-of-fit Indices for the One-, Two-, and Three-Factors Models*

Sample	Solution	log Lik.	pars.	$G^2$	df	p-value	AIC	$\Delta$ AIC	BIC	BICc
Complete	1-factor	-3727.45	40	—	—	—	7534.91	1.43	7694.56	7691.39
	2-factors	-3707.74	59	39.43	19	0.004	7533.47	0.00	7768.97	7765.79
	3-factors	-3695.52	77	24.44	18	0.141	7545.03	11.56	7852.37	7849.20
MG	1-factor	-1827.79	40	—	—	—	3735.60	0.40	3867.53	3864.36
	2-factors	-1808.60	59	38.39	19	0.005	3735.19	0.00	3929.79	3926.62
	3-factors	-1791.81	77	33.56	18	0.014	3737.62	2.43	3991.60	3988.43
AG	1-factor	-1859.41	40	—	—	—	3798.82	0.66	3930.76	3927.59
	2-factors	-1840.08	59	38.66	19	0.005	3798.16	0.00	3992.76	3989.59
	3-factors	-1826.94	77	26.28	18	0.094	3807.88	9.71	4061.85	4058.68

Note:  $G^2$  is the likelihood-ratio chi-square between pairs of nested models.  $\Delta$ AIC is the difference between the AIC of the model and the minimum AIC.

Table 5  
*Goodness-of-fit Indices for the Configural, Weak, and Strong Invariance Models*

Level of invariance	log Lik.	pars.	$G^2$	df	p-value	AIC	$\Delta$ AIC	BIC	BICc
Configural	-5295.89	86	—	—	—	10763.79	22.58	11107.05	11103.88
Weak	-5304.60	66	17.42	20	0.626	10741.20	0.00	11004.64	11001.47
Strong	-5336.13	46	63.07	20	0.000	10764.27	23.07	10947.87	10944.70

Note:  $G^2$  is the likelihood-ratio chi-square between pairs of nested models.  $\Delta$ AIC is the difference between the AIC of the model and the minimum AIC.

Table 6  
*Contrast Tests for Item Parameters*

Parameters	Comparisons
Intercepts	$L_{Mj} = \tau_{mj}^{(MG)} - \tau_{mj}^{(AG)}$ , where $m$ is the male-option of the item $j$ .
	$L_{Fj} = \tau_{fj}^{(MG)} - \tau_{fj}^{(AG)}$ , where $f$ is the female-option of the item $j$ .
	$L_{Dj} = \left( \tau_{mj}^{(MG)} - \tau_{fj}^{(MG)} \right) - \left( \tau_{mj}^{(AG)} - \tau_{fj}^{(AG)} \right)$
Factor loadings	$L_{Lj} = \lambda_{mj} - \lambda_{fj}$

Table 7  
Results of Contrast Tests for Item Parameters

Comparison	$\hat{L}$	$Se(\hat{L})$	$Z$	$p - value$	Comparison	$\hat{L}$	$Se(\hat{L})$	$Z$	$p - value$
$L_{M1}$	-0.186	0.283	-0.659	.76	$L_{D1}$	0.144	0.276	0.524	.74
$L_{M2}$	-0.225	0.252	-0.893	.60	$L_{D2}$	0.215	0.419	0.515	.74
$L_{M3}$	0.041	0.272	0.151	.96	$L_{D3}$	0.340	0.328	1.035	.52
$L_{M4}$	0.251	0.342	0.735	.68	$L_{D4}$	0.259	0.408	0.634	.73
$L_{M5}$	0.113	0.262	0.429	.79	$L_{D5}$	0.242	0.297	0.815	.64
$L_{M6}$	0.259	0.248	1.043	.52	$L_{D6}$	0.709	0.349	2.031	.13
$L_{M7}$	0.650	0.602	1.080	.52	$L_{D7}$	1.397	0.266	5.255	.00
$L_{M8}$	0.943	0.311	3.035	.01	$L_{D8}$	0.920	0.231	3.982	.00
$L_{M9}$	0.134	0.225	0.597	.74	$L_{D9}$	1.030	0.404	2.550	.04
$L_{M10}$	-0.045	0.319	-0.140	.96	$L_{D10}$	-0.062	0.345	-0.179	.96
$L_{F1}$	-0.331	0.345	-0.959	.56	$L_{L1}$	0.889	0.251	3.542	.00
$L_{F2}$	-0.440	0.355	-1.239	.45	$L_{L2}$	1.389	0.391	3.557	.00
$L_{F3}$	-0.299	0.270	-1.107	.52	$L_{L3}$	1.003	0.313	3.208	.01
$L_{F4}$	-0.008	0.259	-0.029	.98	$L_{L4}$	2.228	0.626	3.562	.00
$L_{F5}$	-0.129	0.245	-0.526	.74	$L_{L5}$	1.027	0.259	3.961	.00
$L_{F6}$	-0.450	0.303	-1.486	.39	$L_{L6}$	1.014	0.322	3.147	.01
$L_{F7}$	-0.747	0.563	-1.327	.42	$L_{L7}$	0.681	0.204	3.338	.01
$L_{F8}$	0.023	0.330	0.069	.97	$L_{L8}$	0.237	0.180	1.316	.42
$L_{F9}$	-0.895	0.384	-2.333	.07	$L_{L9}$	0.395	0.272	1.451	.39
$L_{F10}$	0.017	0.228	0.075	.97	$L_{L10}$	0.356	0.269	1.319	.42

Note: Benjamini-Hochberg correction was applied to  $p - values$ .

Table 8

*Goodness-of-fit Indices for the MIMIC Model with and without Invariance between Forms*

Level of invariance	log Lik.	pars.	$G^2$	df	p-value	AIC	$\Delta$ AIC	BIC	BICc
No invariance	-5290.75	71	—	—	—	10723.50	2.25	11006.89	11003.71
Slopes	-5293.35	69	5.21	2	0.074	10724.71	3.46	11000.12	10996.94
Slopes and intercepts	-5293.63	67	0.56	2	0.756	10721.25	0.00	10988.68	10776.08

Note:  $G^2$  is the likelihood-ratio chi-square between pairs of nested models.  $\Delta$ AIC is the difference between the AIC of the model and the minimum AIC.

# FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

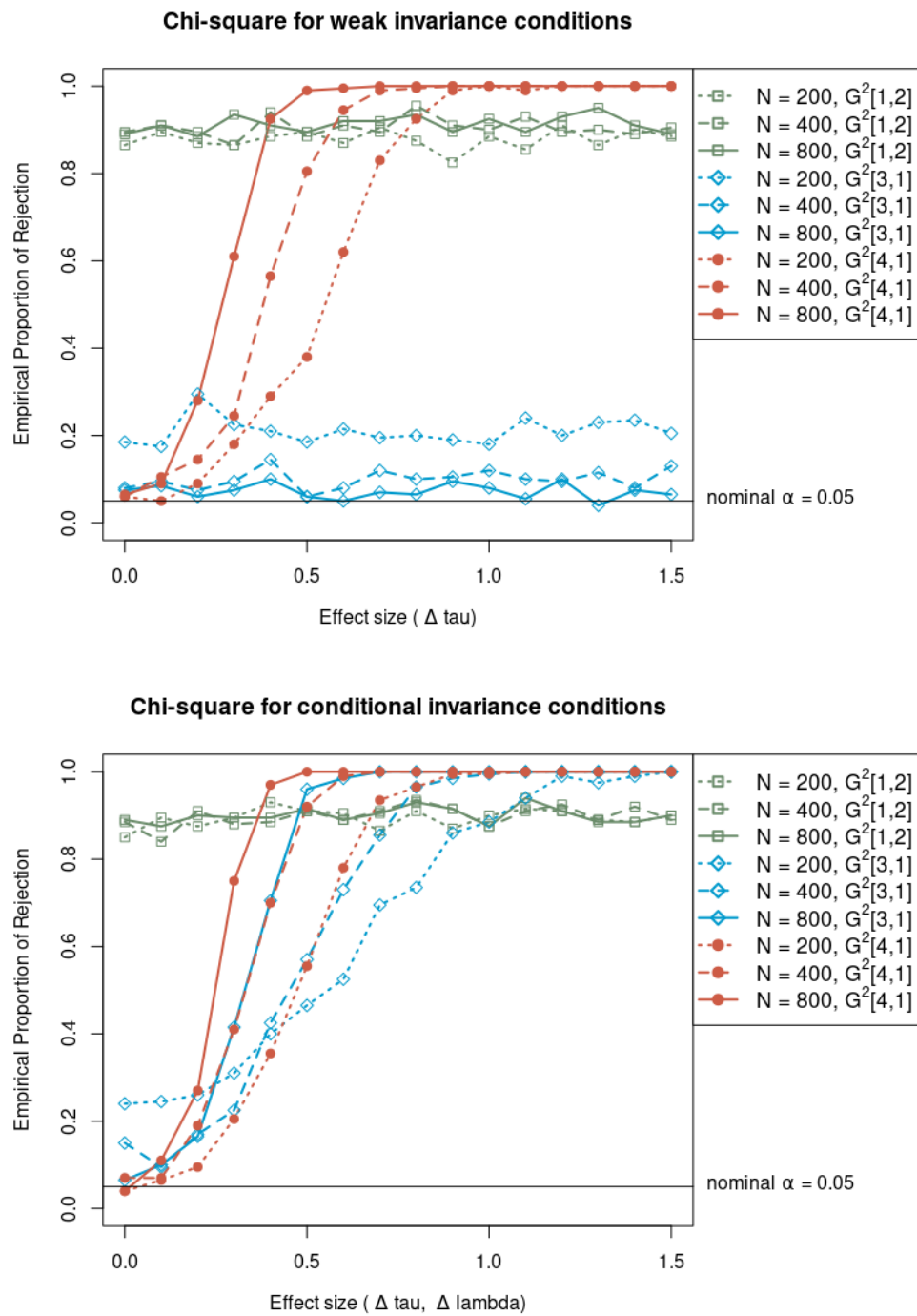


Figure 1. Empirical Proportion of Rejection for chi-square likelihood ratio statistic as a function of effect size.

# FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

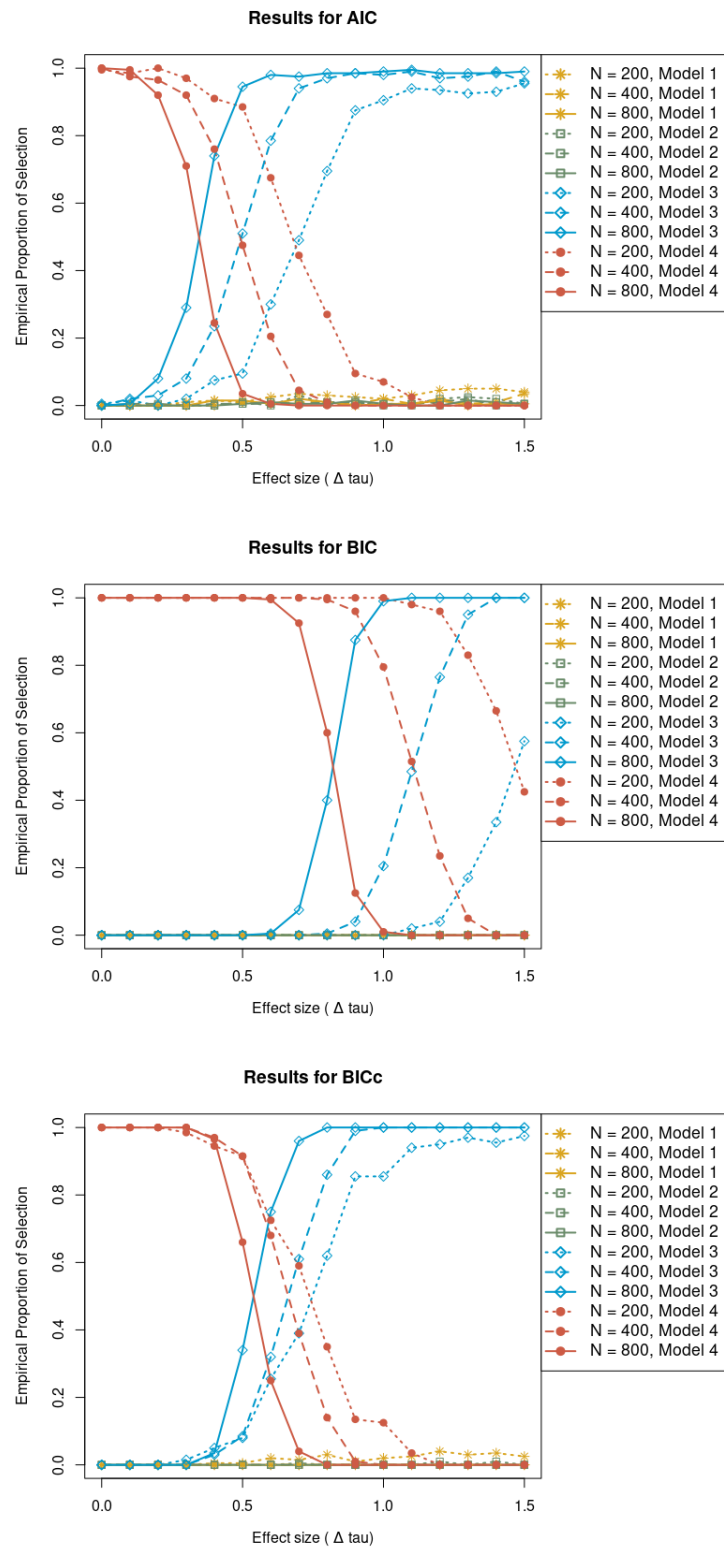


Figure 2. Empirical Proportion of Selection for the AIC, BIC and BICc as a function of effect size ( $\Delta\tau$ ).

# FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

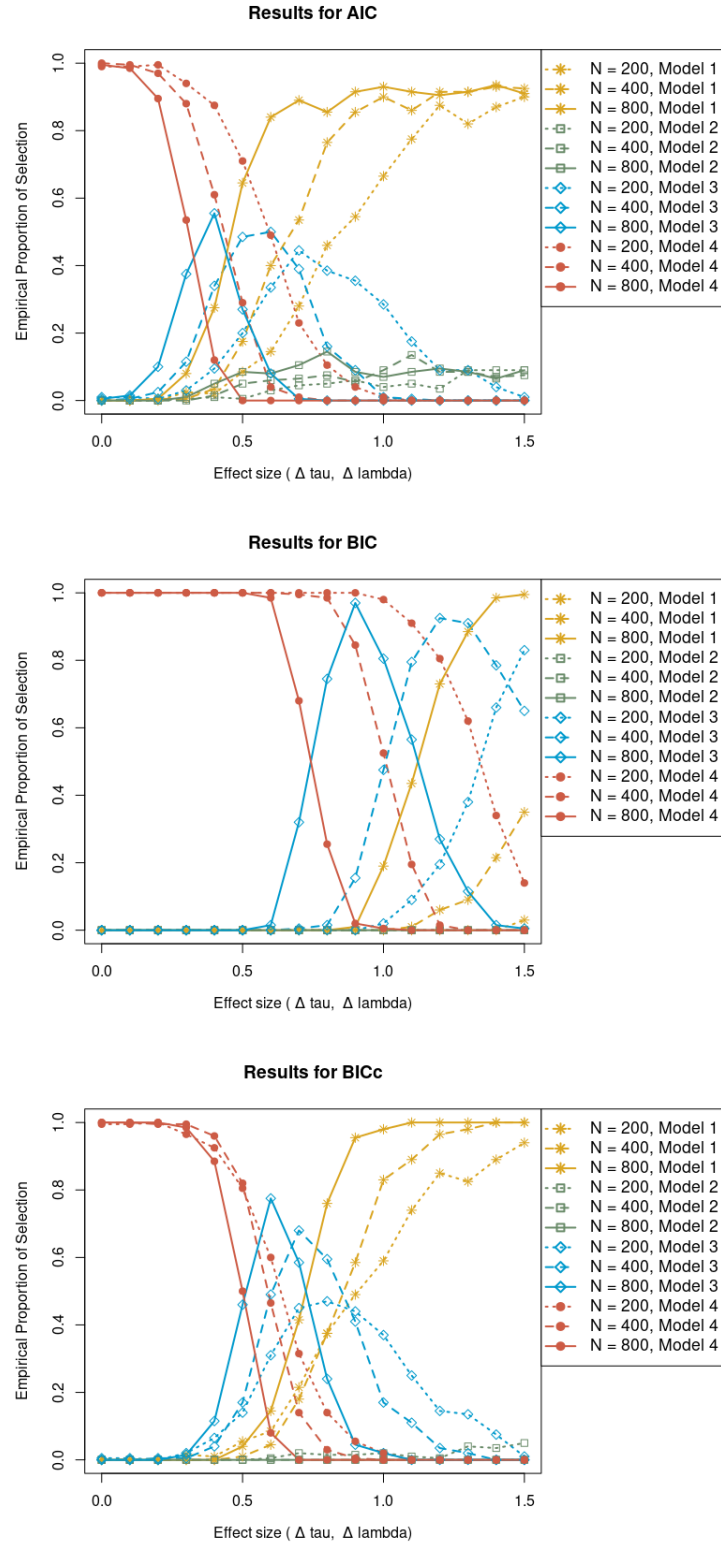


Figure 3. Empirical Proportion of Selection for the AIC, BIC, and BICc as a function of effect size ( $\Delta\tau$  and  $\Delta\lambda$ ).



# FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

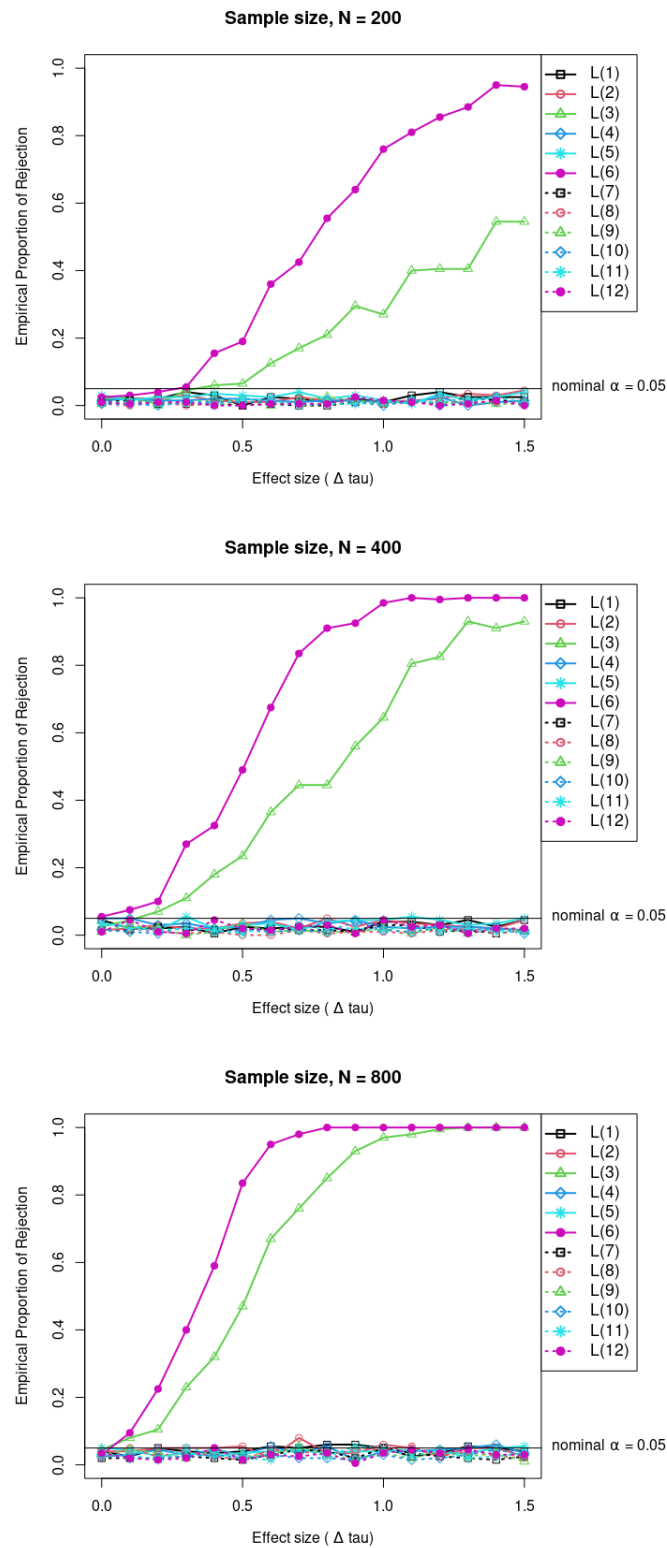


Figure 4. Empirical Proportion of Rejection for  $L$  as a function of effect size ( $\Delta\tau$ ).

# FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

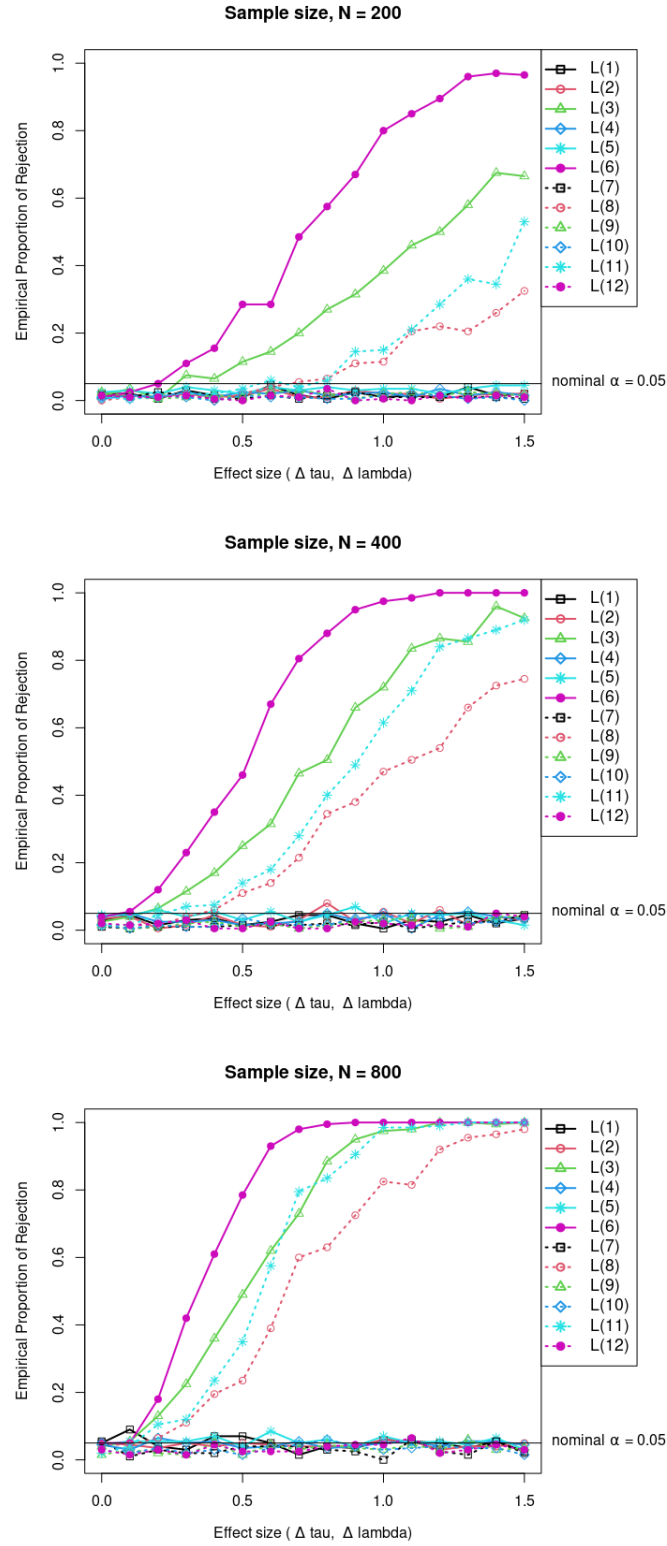


Figure 5. Empirical Proportion of Rejection for  $L$  as a function of effect size ( $\Delta\tau$  and  $\Delta\lambda$ ).

## FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

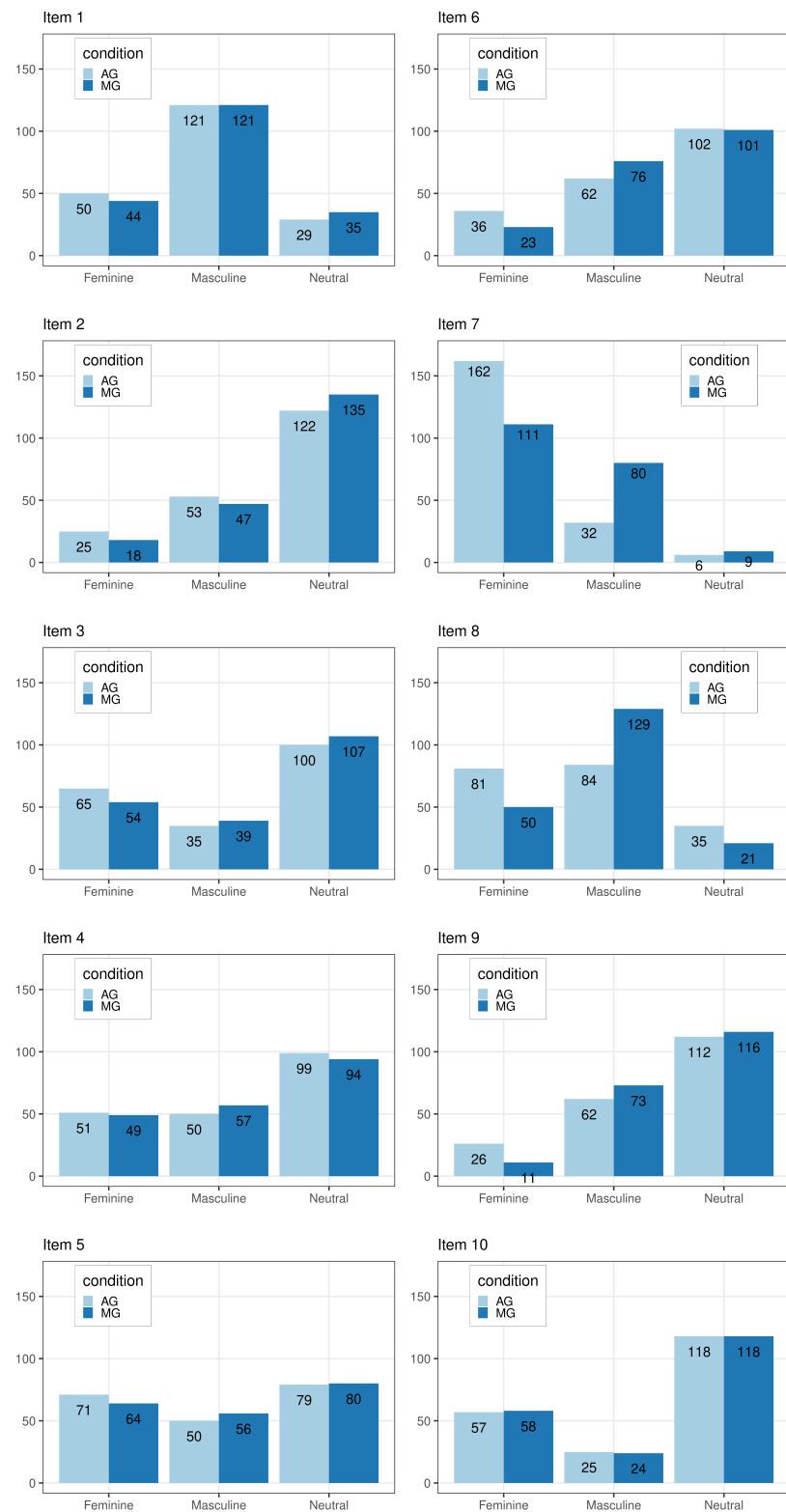


Figure 6. Response frequencies for the item categories in both groups.

# FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

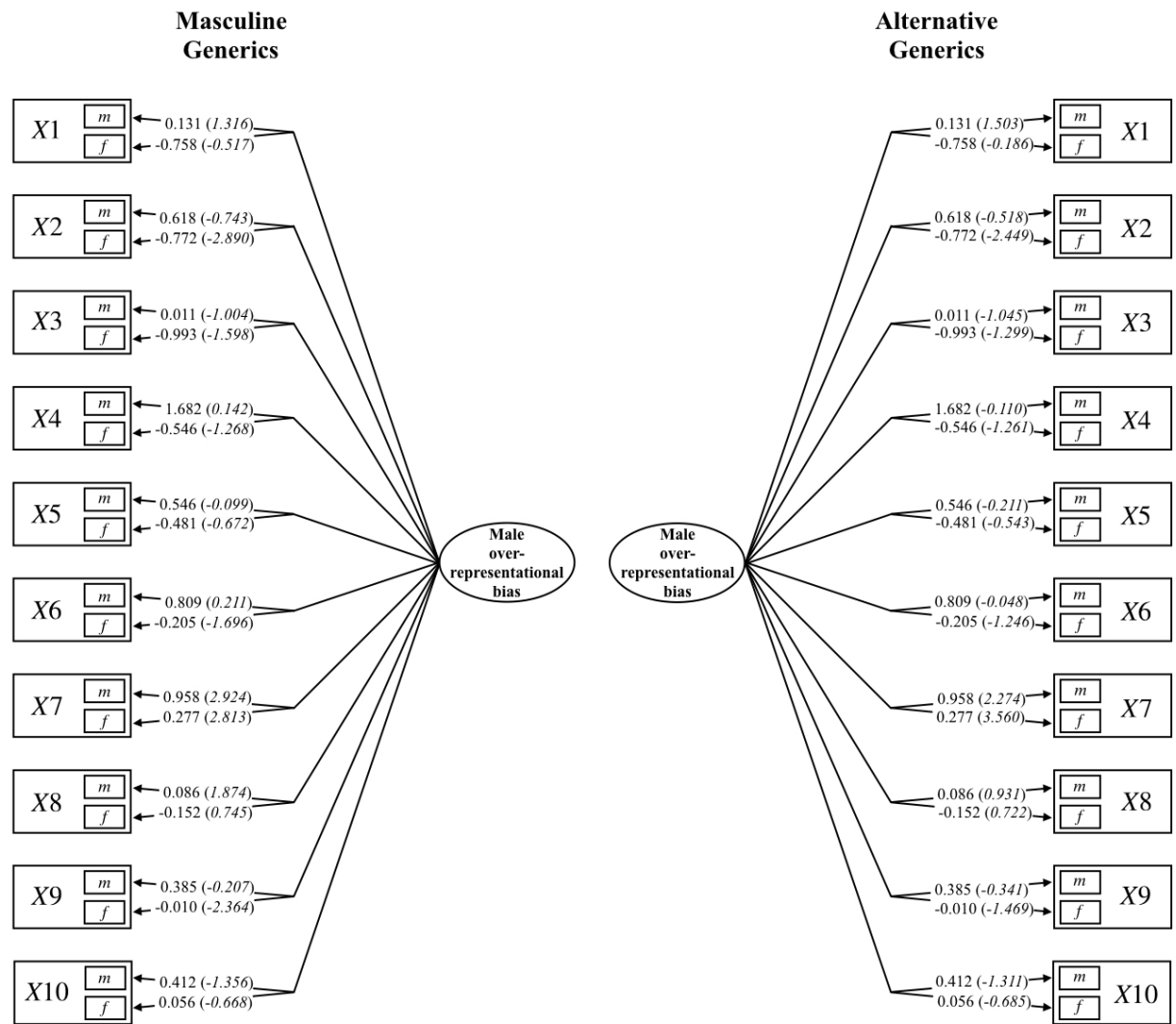


Figure 7. Path diagram of the fitted model. The arrows show the category slope and the intercept is in brackets.

# FACTOR ANALYSIS OF SITUATIONAL JUDGMENT TESTS

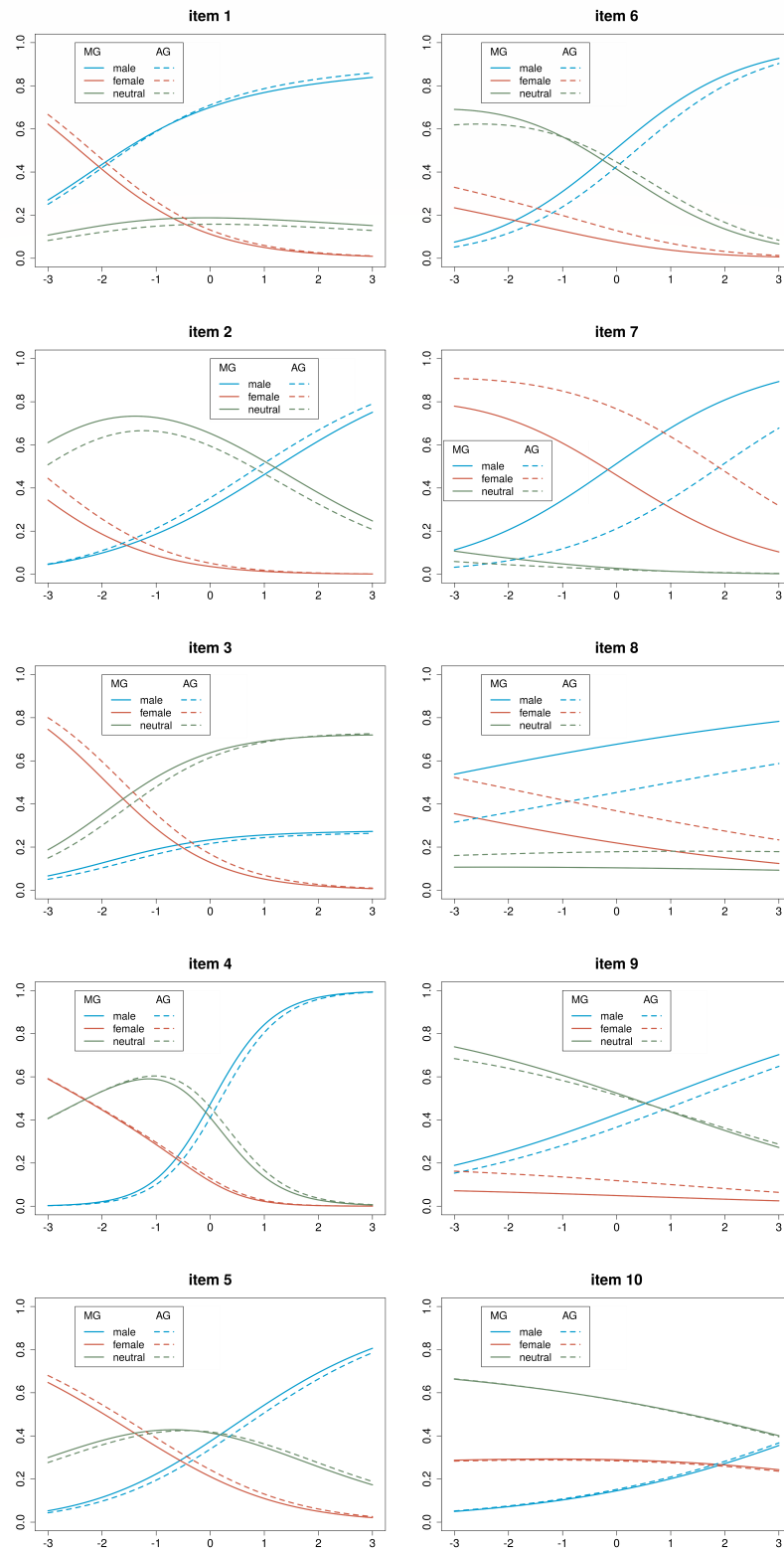


Figure 8. Category Response Functions for the MG and AG forms. The  $x$ -axis represent the factor score ( $\eta$ ) and the  $y$ -axis is the probability of the categories.