In this simulation study, we investigate the power and Type I error rate of a procedure based on the mean and covariance structure analysis (MACS) model in detecting differential item functioning (DIF) of graded response items with five response categories. The following factors were manipulated: type of DIF (uniform and non-uniform), DIF magnitude (low, medium and large), equality/inequality of latent trait distributions, sample size (100, 200, 400, and 800) and equality or inequality of the sample sizes across groups. The simulated test was made up of 10 items, of which only 1 contained DIF. One hundred replications were generated for each simulated condition. Results indicate that the MACS-based procedure showed acceptable power levels (≥ .70) for detecting medium-sized uniform and non-uniform DIF, when both groups’ sample sizes were as low as 200/200 and 400/200, respectively. Power increased as sample sizes and DIF magnitude increased. The analyzed procedure tended to better control for its Type I error when both groups’ sizes and latent trait distribution were equal across groups and when magnitude of DIF and sample size were small.
A procedure for DIF detection based on Sörbom’s (1974) mean and covariance structure analysis (MACS) model has been recently used in empirical studies to detect uniform and non-uniform differential item functioning (DIF) on graded response items (e.g., Chan, 2000; González-Romá, Tomás, Ferreres, & Hernández, 2005; Wasti, Bergman, Glomb, & Drasgow, 2000). Because the MACS model is a model for continuous variables, this procedure has been applied assuming that the aforementioned items approximate to a continuous scale. But there is a scarcity of studies assessing the power of this procedure to detect true DIF when the procedure is used with polytomous ordered response items. As noted in Chan’s article, to begin to clarify this issue “one possibility is to conduct simulation studies with known DIF manipulations to check the sensitivity of the DIF detection method with different sample sizes” (p. 194). The aim of the present simulation study is to assess the sensitivity of the DIF detection procedure based on the MACS model when this procedure is used with polytomous ordered response items.

THE MACS MODEL-BASED PROCEDURE TO DETECT DIF

Assuming that the item responses on the test are accounted for by one latent variable, in the MACS model the item response of individual $i$ to item $j$, $x_{ij}$, can be explained by means of the linear regression of $x_{ij}$ on the latent trait variable $\xi_i$ as:

$$x_{ij}^{(g)} = \mu_j^{(g)} + \lambda_j^{(g)} \xi_i^{(g)} + \delta_{ij}^{(g)}$$

(1)

The regression intercept, $\mu_j$, represents the expected mean response to item $j$ for subjects at the latent trait value of zero. The regression coefficient or factor loading, $\lambda_j$, refers to the expected change in the item response $x_{ij}$ per unit change in $\xi_i$. $\delta_{ij}$ is the random error term, which is assumed to be normally distributed. Finally, $g$ refers to group membership. Within the MACS model, the item intercept ($\mu_j$) corresponds to the item location or attractiveness parameter, whereas the item factor loading ($\lambda_j$) corresponds to the item discrimination parameter (Ferrando, 1996; Mellenbergh, 1994).

An item shows DIF when individuals with equal levels on the latent trait respond differently to the item depending on group membership. DIF can either be uniform or non-uniform, depending on which item parameter is not invariant across the groups of interest. Non-uniform DIF exists when the item discrimination parameter differs across groups, regardless of whether the item location parameter is invariant. Uniform DIF exists when only the item location parameter differs across groups. Within the MACS model, the location or attractiveness parameter corresponds to the expected item score for a trait value of zero. The item discrimination parameter refers to the ability of the item to differentiate among people with different latent trait...
levels. The higher the discrimination parameter, the better the item distinguishes among people with similar levels on the latent trait.

The MACS model allows researchers to test both uniform and non-uniform DIF according to group membership. In general terms, testing for uniform DIF using the MACS model involves testing the null hypothesis that parameter $\mu_j$ is invariant across groups [e.g., $\mu_j^{(1)} = \mu_j^{(2)} = \ldots = \mu_j^{(G)}$], whereas testing for non-uniform DIF involves testing the null hypothesis that parameter $\lambda_j$ is invariant across groups [e.g., $\lambda_j^{(1)} = \lambda_j^{(2)} = \ldots = \lambda_j^{(G)}$]. When only the invariance of item intercepts cannot be maintained, then uniform DIF exists [i.e., $\mu_j^{(1)} = \mu_j^{(2)} = \ldots = \mu_j^{(G)}$]. When the invariance of item factor loadings cannot be maintained, regardless of whether the intercepts are invariant or not, then non-uniform DIF exists [i.e., $\lambda_j^{(1)} = \lambda_j^{(2)} = \ldots = \lambda_j^{(G)}$].

The DIF detection procedure based on the MACS model analyzed here uses modification indices (MI) for detecting which specific items function differentially across groups (see Chan, 2000). An MI shows the reduction in the model chi-square value, if the implied constrained parameter is freely estimated. Because this chi-square difference is distributed with one degree of freedom, it is easy to determine whether the reduction in chi-square is statistically significant. The procedure starts with a fully equivalent model, in which all the item factor loadings and the intercepts are constrained to be equal across groups. This model corresponds to Meredith’s (1993) strong factorial invariance model. The investigated DIF detection procedure does not impose invariance restrictions on the item residual variances because the two item parameters of substantive interest in the DIF literature are the location and the discrimination parameter (in our case, the intercept and the factor loading). Should the researcher wish to investigate invariance of item precision across groups, then an invariance restriction on the item residual variances should be imposed. This model would correspond to Meredith’s strict factorial invariance model.

Once the model is fitted to data, non-uniform DIF is examined. The factor loadings that show statistically significant MIs are identified, and the corresponding items are said to show non-uniform DIF. Next, uniform DIF is investigated. The intercepts that show statistically significant MIs are identified, and the corresponding items are said to show uniform DIF.

There is a scarcity of published simulation studies that have evaluated the procedure described above. Hernández and González-Romá (2003) evaluated this procedure using simulated graded responses generated by means of Samejima’s (1969) Graded Response Model. They manipulated only two factors to establish the DIF conditions: type of DIF (uniform and non-uniform) and amount of DIF (low, medium, and high). Sample size was set to 800 subjects for all the study conditions, and the simulated test length was 10 items, one of which was a DIF item. In the non-DIF condition, the proportion of false positive (FP) or incorrect DIF iden-
tifications in the intercept (i.e., the proportion of items that were detected as showing uniform DIF) and in the factor loading (i.e., the proportion of items that were detected as showing non-uniform DIF) equaled .008 (the expected FP proportion was .005). In the uniform DIF conditions, the proportions of FPs in which DIF was wrongly detected in the intercept were less than .006. The true positive (TP) proportion in the intercept in which DIF was correctly identified increased as the amount of DIF increased. Hernández and González-Romá observed that differences in the intercept as small as .19 in a graded response scale ranging from 1 to 5 yielded TP proportions of .88. When the difference generated was .31, the proportion of TPs was 1. Therefore, they concluded that the MACS procedure was very sensitive to small differences across groups in the item intercept. In the non-uniform DIF conditions, the proportions of FPs in which DIF was wrongly detected in the factor loadings were less than .008. The proportion of TPs in the factor loading increased as the amount of DIF increased, but it did not exceed .44, probably because the amount of DIF generated was too small, even in the high DIF condition. This study showed that, under the simulated conditions, the MACS procedure maintained reasonable control of its Type I error, and it showed a satisfactory sensitivity to uniform DIF. However, as the authors acknowledged, the very few factors were manipulated in their study, and this situation precluded sound generalizations. They suggested investigating the performance of the analyzed procedure when the comparison groups differ in latent trait distribution and in size (Hernández & González-Romá, 2003).

Other researchers have evaluated other CFA-based methods for DIF detection. Meade and Lautenschlager (2004) conducted a simulation study to investigate to what extent differences in factor loadings across groups were detected by a nested chi-square difference test between a model of configural invariance (i.e., a model with no invariance constraints that posits that the same structure holds across groups; see Meredith, 1993) and a model in which factor loadings were constrained to be equal across groups. They manipulated sample size, the number of scale items, the number of items showing measurement invariance (that is, DIF), and the pattern of these differences. They only simulated differences in the item factor loadings, not in the item intercepts (in other words, they only generated non-uniform DIF conditions), and those differences equaled .25 in all conditions. They observed that differences in factor loading matrices were correctly detected in most conditions (in all but two conditions, the proportion of correct detections was higher than .9). However, in the conditions in which no differences in factor loadings were simulated, the proportion of false positive detections ranged between .33 and .50. Therefore, the findings reported by Meade and Lautenschlager point out that whereas the studied procedure showed a high power, it did not maintain control of its Type I error rate. Another limitation of this procedure is that in empirical applications it does not show which items have noninvariant loadings.

Oort (1998) carried out a simulation study to check the power of the method he proposed for detecting DIF (see Oort, 1992) using polytomous graded response
items. Within this method, the common factor model serves as an item response model, and a different factor is included for each of the potential causes of DIF. The results of this study showed that the model was adequate for the evaluation of DIF in items with seven response categories, especially when the sample size was large, the mean trait difference between the focal and the reference groups was small, the sizes of both groups were equal, and the amount of bias was large. However, one of the disadvantages of Oort’s approach is that it does not make it possible to distinguish between uniform and non-uniform DIF when it is applied to empirical data.

As stated before, in this study we assessed the power and Type I error rate of the procedure for DIF detection based on the MACS model that uses modification indices to detect both uniform and non-uniform DIF in polytomous ordered response items. Considering that previous studies have demonstrated that the power of this and other procedures for DIF detection increases as the magnitude of DIF is increased (Hernández & González-Romá, 2003; Narayanan & Swaminathan, 1996), we manipulated the magnitude of DIF, expecting to observe the same effect. Sample size is also known to be positively related to the power of DIF detection procedures (Narayanan & Swaminathan, 1994, 1996). Moreover, Kaplan and George (1995) examined the power of the Wald test of factor mean differences under violations of factorial invariance within the multigroup CFA method. They found that the marginal effect of inequality of sample size across groups was to decrease power. Therefore, we manipulated sample size and created conditions of equality and inequality of sample sizes. Finally, a number of DIF studies have shown that, for different DIF detection procedures, when latent trait distributions were identical across groups, power was greater than when latent trait distributions were distinct [e.g., reference group: \( N(0, 1) \) and focal group: \( N(-1, 1) \)] (Ankemann, Witt, & Dunbar, 1999; Clauser, Mazor, & Hambleton, 1993; Narayanan & Swaminathan, 1996). Moreover, when latent trait distributions were different, the Mantel-Haenszel and the SIBTEST procedures yielded inadequate Type I error rates (Roussos & Stout, 1996; Uttaro & Millsap, 1994). Thus, we also manipulated the latent trait distributions of both the reference and the focal groups.

By means of this simulation study, we extend previous research on the procedure for DIF detection based on the MACS model, and contribute to clarifying one of the questions raised in Chan’s (2000) article: ascertaining the power of this procedure to detect true DIF in polytomous ordered response items under different conditions.

METHOD

Simulation of Data

To simulate realistic conditions, the simulated test was made up of 10 items with five ordered response options. Ten items is a common number for the measurement of psychological constructs (e.g., see the Fifth Edition of the 16PF Questionnaire).
Five categories is also a frequent number in psychological tests and questionnaires, and it is the minimum recommended by different authors to adequately represent subjects’ scores on graded response items by means of the MACS model (Bollen & Barb, 1981; Dolan, 1994).

To use realistic item parameter values for the reference group, we estimated the item parameters of Anderson and West’s (1998) support for innovation scale. This scale is part of the Team Climate Inventory (Anderson & West, 1998). The scale is composed of 11 items that are responded to using a 5-point Likert-scale. The unidimensionality of this questionnaire was tested in a sample made up of 1112 employees of health centers. Data were submitted to a principal component analysis. We only obtained one component with an eigenvalue greater than 1, which explained 53% of the variance. One of the 11 items was randomly discarded to fit the length of the simulated test.

The MACS model was fitted to the 10 selected items by means of LISREL 8.30 (Jöreskog & Sörbom, 1993). Taking into account that in the aforementioned sample the 10 selected items showed skewness and kurtosis values within the range from −1 to +1 (skewness ranged from −.95 to .02 and kurtosis ranged from −.86 to .48), the assumption of approximate normality was tenable, justifying the use of normal theory maximum likelihood (ML) estimation methods (Bollen, 1989; Muthén & Kaplan, 1985). The model showed an acceptable fit to data ($\chi^2 = 331.34$, $df = 35$; $NNFI = .92$; $AGFI = .91$). The parameter estimates we obtained were used to generate the continuous responses to the 10 items according to Equation 1. Specifically, the standardized solution estimates provided by LISREL for both factor loadings and intercepts were used as the item parameters for the reference group (see Table 1). We recall that in LISREL, under the standardized solution, only latent variables are standardized, whereas the variances and covariances

<table>
<thead>
<tr>
<th>Item</th>
<th>$\sigma^2$</th>
<th>$\mu$</th>
<th>$\lambda$</th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.74</td>
<td>3.44</td>
<td>0.65</td>
<td>0.02</td>
<td>0.13</td>
<td>0.29</td>
<td>0.51</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.77</td>
<td>3.31</td>
<td>0.64</td>
<td>0.02</td>
<td>0.16</td>
<td>0.33</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>0.76</td>
<td>3.44</td>
<td>0.66</td>
<td>0.02</td>
<td>0.15</td>
<td>0.25</td>
<td>0.52</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>0.78</td>
<td>3.04</td>
<td>0.56</td>
<td>0.04</td>
<td>0.19</td>
<td>0.51</td>
<td>0.21</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>0.65</td>
<td>3.36</td>
<td>0.59</td>
<td>0.01</td>
<td>0.16</td>
<td>0.33</td>
<td>0.48</td>
<td>0.03</td>
</tr>
<tr>
<td>6</td>
<td>0.80</td>
<td>3.17</td>
<td>0.64</td>
<td>0.02</td>
<td>0.23</td>
<td>0.32</td>
<td>0.40</td>
<td>0.03</td>
</tr>
<tr>
<td>7</td>
<td>0.62</td>
<td>3.50</td>
<td>0.56</td>
<td>0.01</td>
<td>0.11</td>
<td>0.26</td>
<td>0.59</td>
<td>0.03</td>
</tr>
<tr>
<td>8</td>
<td>0.92</td>
<td>2.97</td>
<td>0.49</td>
<td>0.05</td>
<td>0.31</td>
<td>0.30</td>
<td>0.31</td>
<td>0.03</td>
</tr>
<tr>
<td>9</td>
<td>0.60</td>
<td>3.46</td>
<td>0.56</td>
<td>0.01</td>
<td>0.12</td>
<td>0.30</td>
<td>0.54</td>
<td>0.03</td>
</tr>
<tr>
<td>10</td>
<td>0.73</td>
<td>3.03</td>
<td>0.48</td>
<td>0.03</td>
<td>0.24</td>
<td>0.44</td>
<td>0.27</td>
<td>0.03</td>
</tr>
</tbody>
</table>

*Note.* $P_i$ is the probability of selecting response category $i$. The reference group's item parameters and response category probabilities.
of the observed variables maintain their original metric. It is under LISREL’s completely standardized solution when both latent and observed variables are standardized.

The latent trait scores required to generate the continuous responses to the 10 items for the reference group following Equation 1 were generated according to a normal distribution, \( N(0, 1) \). The error components were consistently drawn from a standard normal distribution for all items and conditions. These components were weighted to reproduce the observed error variance components.

Once the continuous item responses were obtained, they were categorized into five response categories, according to the four threshold estimates obtained for each item in the empirical sample by means of the PRELIS program. To categorize the continuous item responses \( (x_{ij}) \), the standardized threshold estimates provided by PRELIS were transformed into their original nonstandardized metrics. For the five-category items simulated, graded item responses \( (k) \) were assigned as follows: \( k = 1 \) if \( x_{ij} \leq \tau_1 \); \( k = 2 \) if \( \tau_1 < x_{ij} \leq \tau_2 \); \( k = 3 \) if \( \tau_2 < x_{ij} \leq \tau_3 \); \( k = 4 \) if \( \tau_3 < x_{ij} \leq \tau_4 \) and \( k = 5 \) if \( \tau_4 < x_{ij} \). The probabilities associated with the five response categories according to the standard normal distribution are shown in Table 1.

Apart from a non-DIF condition, in which all the 10 item parameters were equal for both the reference and the focal groups, four factors were varied to create different DIF-conditions (see Table 2): the equality or inequality of the group latent distributions (i.e., presence or absence of impact), the type of DIF (uniform and non-uniform), the magnitude of DIF (low, medium, and high), and the sample sizes of both the reference and the focal groups (800, 400, 200, and 100 subjects), which were combined in different ways.

<table>
<thead>
<tr>
<th>Factors Manipulated</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>Combination of sample sizes results in 10 conditions: 4 conditions with</td>
</tr>
<tr>
<td>100, 200, 400, 800</td>
<td>equal sample sizes for both groups, and 6 conditions with unequal</td>
</tr>
<tr>
<td></td>
<td>sample sizes (in these cases, the reference group was always bigger</td>
</tr>
<tr>
<td></td>
<td>than the focal group)</td>
</tr>
<tr>
<td>Latent trait distribution</td>
<td>Equal: RG ( N(0,1) ) FG ( N(0,1) )</td>
</tr>
<tr>
<td></td>
<td>Unequal: RG ( N(0,1) ) FG ( N(0,1) )</td>
</tr>
<tr>
<td>Type of DIF</td>
<td>No DIF</td>
</tr>
<tr>
<td></td>
<td>Uniform DIF</td>
</tr>
<tr>
<td></td>
<td>Non-Uniform DIF</td>
</tr>
<tr>
<td>Magnitude of DIF</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>High</td>
</tr>
</tbody>
</table>

Note. RG = reference group; FG = focal group; DIF = differential item functioning.
Equality/Inequality of the latent distributions. When the latent variable distributions were equal across groups, both the reference and focal group latent scores were generated according to a standard normal distribution, $N(0,1)$. Consequently, there was no impact. When the latent distribution was unequal across groups, the focal group’s distribution was modified to produce impact. Specifically, the mean of the focal group was one standard deviation below the mean of the reference group, $N(-1,1)$.

Type of DIF. The intercept parameter or the discrimination parameter of one item (Item 1) was varied across groups to produce uniform and non-uniform DIF conditions, respectively. Specifically, for the uniform DIF conditions, the intercept of Item 1 for the focal group was obtained by subtracting a constant from the Item 1 intercept for the reference group. Consequently, Item 1 was less attractive (i.e., harder to agree with) for the focal group than for the reference group. For the non-uniform DIF conditions, the factor loading of Item 1 for the focal group was obtained by subtracting a constant from the Item 1 loading for the reference group. Consequently, this item was less discriminative for the focal group than for the reference group.

Magnitude of DIF. For the uniform DIF conditions, the differences in the intercepts of Item 1 were .10, .25, and .50 for the so-called low, medium, and high DIF conditions. These constant values were subtracted from the reference group intercepts to obtain the focal group intercepts. For the non-uniform DIF conditions, the constant values were selected in such a way that the differences under LISREL’s standardized solution used to generate the data were equivalent to differences of .10, .25, and .50 under LISREL’s completely standardized solution. These values extended factor loading differences considered previously. For instance, Meade and Lautenschlager (2004) only considered differences of .25 in factor loadings. The corresponding values under the standardized solution were .08, .21, and .43, respectively. These constant values were subtracted from the reference group’s standardized factor loadings to obtain those of the focal group. The Item 1 parameters for the focal group for the types and magnitudes of DIF are shown in Table 3.

Sample size. Four sample sizes were considered in the simulation: 100, 200, 400, and 800 subjects. These sizes are representative of those available to most researchers and practitioners, and they provide a wide range of conditions for testing the influence of sample size. These sample sizes were combined in such a way that two additional conditions could be specified: the equality and inequality of sample sizes in both the reference and the focal groups. In four conditions the reference and the focal groups showed equal sample sizes (800–800, 400–400, 200–200, and
100–100), and in six conditions both groups showed unequal sample sizes (800–400, 800–200, 800–100, 400–200, 400–100, 200–100). In these latter six conditions, the reference group was the largest group. We chose this alternative because in empirical DIF studies the reference group frequently is a majority group that is compared to a minority group (e.g., a linguistic or ethnic minority group).

Thus, 140 conditions were created and evaluated ([2 latent distribution conditions × 10 sample size combinations × 2 types of DIF × 3 magnitudes of DIF] + [2 latent distribution conditions × 10 sample size combinations × 1 non-DIF condition]). For each condition, 100 samples were generated.

At this point, we want to bring up a concern related to the magnitudes of factor loadings in categorized data. When continuous responses are categorized, factor loadings tend to be attenuated to some degree (see Distefano, 2002, for a more detailed discussion). Consequently, the between-group differences in factor loadings in our categorized responses (i.e., non-uniform DIF) might not have the same magnitude as in the original loadings used to generate the continuous data.

To deal with this concern, we first evaluated to what extent factor loadings were attenuated as a result of data categorization. Specifically, a 100,000 subject sample was generated for each of the two non-DIF conditions (equality/inequality of the latent distributions) according to the parameters previously presented. Once the categorized data were obtained, they were used to estimate the item parameters by means of the MACS model with equality constraints across groups in both the factor loadings and the intercepts. When both groups had equal latent distributions, the factor loadings estimated for the categorized data were attenuated by an average value of 0.045 (in comparison with the factor loadings used to generate the continuous data). When both groups had different latent distributions, the factor loadings estimated for the categorized data were attenuated by an average value of 0.02.
Second, we evaluated to what extent the non-uniform DIF magnitudes in the categorized data were comparable with the magnitudes used to generate the continuous data. Again, a 100,000 subject sample was generated for every non-uniform DIF condition, using the reported parameters and procedure. Once the data were categorized, we fitted the MACS model with equality constraints across groups for every intercept and every factor loading, except those corresponding to the DIF item. Results showed that the maximum difference between the DIF magnitude simulated for the continuous data and the DIF magnitude observed after categorizing the data was 0.04, regardless of the equality or inequality of the latent distribution. Considering these results, we concluded that the attenuation effects of categorizing the simulated continuous data, and the impact of this categorization on the magnitude of non-uniform DIF, was not severe in any case.

Analysis

Analyses were carried out by means of LISREL 8.30 (Jöreskog & Sörbom, 1993). As in practical applications of the MACS model for DIF detection (e.g., Chan, 2000), we used ML estimation methods. The MACS model was fitted to the $10 \times 10$ item variance-covariance matrices and vectors of 10 means for both the reference and the focal groups. To detect uniform and non-uniform DIF, the procedure described in the introduction was applied. For all the models, a number of constraints were imposed for model identification and scaling purposes. First, an item was chosen as the reference indicator (specifically, the item with the highest discrimination parameter). This item factor loading was set to 1 in both groups in order to scale the latent variable and provide a common scale in both groups. Second, the factor mean was fixed to zero in the reference group for identification purposes, whereas the factor mean in the focal group was freely estimated. Finally, the reference indicator intercepts were constrained to be equal in both groups, in order to identify and estimate the factor mean in the focal group and the intercepts in both groups (Sörbom, 1982).

To ensure that DIF was detected only when it existed and rarely by chance alone, the Bonferroni correction was used to test the statistical significance of MIs. Considering that in the procedure proposed by Chan (2000) uniform and non-uniform DIF are examined separately, and that the simulated test was made up of 10 items, the alpha value for determining the statistical significance of each MI was .005 (i.e., .05/10).

RESULTS

To assess the power of the MACS-based procedure to detect DIF, we computed the TP ratio or proportion of correct identifications of Item 1. To assess Type I error
rate, we computed the FP ratio or proportion of incorrect DIF identifications. Specifically, for the non-DIF condition, DIF identifications of Items 1 to 10 were classified as FPs, regardless of the kind of parameters implied (intercept or factor loading). For the uniform DIF conditions, DIF identifications of loadings of Items 1 to 10 and DIF identifications of intercepts of Items 2 to 10 were classified as FPs. Finally, for the non-uniform DIF conditions, DIF identifications of loadings of Items 2 to 10, and DIF identifications of intercepts of Items 1 to 10, were identified as FPs. An item was considered to show DIF when the decrement in chi-square shown by the corresponding MI was statistically significant (i.e., greater than $X^2_{(0.005,1)} = 7.88$). The results are considered separately in the following paragraphs, according to the different DIF conditions.

**Non-DIF Conditions**

Focusing on the non-DIF condition, the results obtained when the latent distribution was equal across groups showed that both the proportions of FPs in the intercept (i.e., the proportion of items that were detected as showing uniform DIF), and the proportions of FPs in the factor loading (i.e., the proportion of items that were detected as showing non-uniform DIF) across different combinations of sample sizes, were generally equal to or less than the assumed nominal alpha level of .005 (see Table 4). In only 3 of 10 cases was the proportion of FPs in the factor loading greater than .005. In these cases, the FP rate was close$^1$ to the nominal alpha level (it equaled .006). Likewise, in only 3 of 10 cases was the proportion of FPs in the intercept greater than .005. In 2 of these cases, the FP proportion was close to the nominal alpha level (it equaled .006), and in 1 case it was not close to the nominal alpha level (it equaled .008).

This was the case in 14 out of the 20 considered conditions. In 5 conditions the proportion of FPs was close (see footnote 1) to the nominal alpha level (it equaled .006), and in 1 condition it equaled .008. Thus, we concluded that the investigated procedure maintained control of its Type I error rate when there was no DIF and the latent distributions were equal across groups.

When the latent distribution was unequal across groups, the proportion of FPs in the intercept and in the factor loading depended on the equality/inequality of sample sizes. When sample sizes were equal, the proportion of FPs was equal to or less than .005 (see Table 4). When sample sizes were unequal, the proportion of FPs generally notably exceeded .005. This was particularly noticeable for the pro-

---

$^1$We operationalized closeness following Bradley’s (1978) liberal criterion of robustness. According to Bradley, an observed proportion of FPs ($p$) is robust at a specific nominal significance level ($\alpha$) if it is within the following interval: $0.5 \alpha \leq p \leq 1.5\alpha$. In our case, the interval is $0.0025 \leq p \leq 0.0075$. Therefore, we concluded that an observed proportion of FPs was close to the nominal significance level ($\alpha = .005$) when $p \leq .0075$. 
portions of FPs in the factor loading, where only 1 out of the 6 computed proportions was close to the alpha level. The proportions of FPs in the intercept when sample sizes were unequal were not as high as those observed for the factor loadings. In 4 of the 6 conditions, these proportions ranged between .009 and .008, and in 2 cases they were .007 and .004. These results point out that when there was no DIF and the latent distributions were unequal across groups, the analyzed procedure controlled for Type I error only when the sample sizes of the reference and the focal groups were equal.

Uniform DIF Conditions

Equality of latent distributions. When the latent distribution was equal across groups, the proportion of FPs in the factor loading tended to increase as the magnitude of DIF increased (see Table 5). When the sample size of both groups was equal, the proportion of FPs in the factor loading generally exceeded the nominal alpha value. Only in 4 of the 12 conditions was the FP ratio equal to or close to .005. When sample sizes were unequal, 12 of 18 proportions of FPs in the factor loading were smaller than or close to 0.005. The remaining 6 proportions ranged between .008 and .023, and they appeared in conditions in which the magnitude of DIF was medium or high.

TABLE 4
Type I Error Rate in Non-DIF Conditions

<table>
<thead>
<tr>
<th>Focal Group</th>
<th>800</th>
<th>400</th>
<th>200</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>λ</td>
<td>μ</td>
<td>λ</td>
<td>μ</td>
<td>λ</td>
</tr>
<tr>
<td>Reference group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal latent distributions</td>
<td>[N (0,1)]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>.003</td>
<td>.005</td>
<td>.004</td>
<td>.005</td>
</tr>
<tr>
<td>400</td>
<td>.006</td>
<td>.005</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>200</td>
<td>.002</td>
<td>.005</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>100</td>
<td>.003</td>
<td>.004</td>
<td>.003</td>
<td>.004</td>
</tr>
<tr>
<td>Unequal latent distributions</td>
<td>[N (0,1) vs N (–1,1)]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>.005</td>
<td>.005</td>
<td>.026</td>
<td>.008</td>
</tr>
<tr>
<td>400</td>
<td>.001</td>
<td>.000</td>
<td>.016</td>
<td>.009</td>
</tr>
<tr>
<td>200</td>
<td>.000</td>
<td>.000</td>
<td>.007</td>
<td>.007</td>
</tr>
<tr>
<td>100</td>
<td>.000</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. DIF = differential item functioning.
<table>
<thead>
<tr>
<th></th>
<th>Reference Group [N (0,1)]</th>
<th>800</th>
<th>400</th>
<th>200</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FP λ</td>
<td>FP μ</td>
<td>TP μ</td>
<td>FP λ</td>
</tr>
<tr>
<td>Equal latent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distributions</td>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>.007</td>
<td>.009</td>
<td>.580</td>
<td>.006</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td>.007</td>
<td>.021</td>
<td>1</td>
<td>.008</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>.046</td>
<td>.130</td>
<td>1</td>
<td>.023</td>
</tr>
<tr>
<td>400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>.005</td>
<td>.007</td>
<td>.220</td>
<td>.003</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td>.016</td>
<td>.018</td>
<td>.970</td>
<td>.005</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>.024</td>
<td>.058</td>
<td>1</td>
<td>.009</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>.008</td>
<td>.003</td>
<td>.070</td>
<td>.007</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td>.008</td>
<td>.003</td>
<td>.720</td>
<td>.006</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>.011</td>
<td>.020</td>
<td>1</td>
<td>.008</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unequal latent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distributions</td>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>.009</td>
<td>.001</td>
<td>.440</td>
<td>.058</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td>.099</td>
<td>.007</td>
<td>1</td>
<td>.115</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>.126</td>
<td>.154</td>
<td>1</td>
<td>.154</td>
</tr>
<tr>
<td>400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>.001</td>
<td>.000</td>
<td>.110</td>
<td>.022</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td>.037</td>
<td>.002</td>
<td>.990</td>
<td>.048</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>.103</td>
<td>.036</td>
<td>1</td>
<td>.115</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>.000</td>
<td>.000</td>
<td>.020</td>
<td>.005</td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td>.002</td>
<td>.000</td>
<td>.760</td>
<td>.021</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>.068</td>
<td>.008</td>
<td>1</td>
<td>.061</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. DIF = differential item functioning; FP = false positives; TP = true positives.
Regarding the proportion of FPs in the intercept, it tended to increase as the magnitude of DIF increased (see Table 5). When sample sizes were equal, the proportion of FPs in the intercept tended to increase as sample size increased. The FP ratio was less than the nominal alpha level, or close to it in the cases of sample sizes of 100 and 200 with a low or medium magnitude of DIF, and when sample sizes equaled 400 and the magnitude of DIF was low (5 of 12 conditions). In the remaining conditions, the proportion of FPs in the intercept varied between .009 and .13. When sample sizes were unequal, the proportion of FPs in the intercept generally exceeded the nominal alpha value. However, when the size of the focal group was 200 or 100 and the DIF magnitude was low, the observed proportions were less than or close to the alpha value.

In general terms, these results indicate that under uniform DIF conditions, when the latent distribution was equal across groups, the analyzed procedure showed difficulties in maintaining control of its Type I error rate, especially as DIF magnitude and sample sizes increased.

The proportion of TPs in which DIF was correctly detected in the intercept of Item 1 increased as DIF magnitude and sample sizes increased (see Table 5). When the magnitude of DIF was high, the DIF item was almost always detected, regardless of the sample sizes (the TP ratio ranged between .99 and 1). When the magnitude of DIF was medium and both groups’ sizes were 200, the TP ratio equaled .72. This ratio increased when the sample size of the reference group or the sample sizes of both the reference and the focal groups increased, ranging from .89 (when the reference group’s size was 400 and the focal group’s size was 200, i.e., 400/200), to 1 (when sample sizes were 800/400 and 800/800). However, when the sample size of any of the groups was 100, the power ranged between .34 and .60. Finally, when the DIF magnitude was low, the TP ratio did not exceed .25, except when both groups’ sizes were 800; in this case, the TP ratio was .58. Thus, when the magnitude of uniform DIF was medium, and both groups’ sample sizes were as low as 200/200, the procedure investigated showed an acceptable power, which increased as the magnitude of DIF and the reference group’s size or both groups’ sizes increased.

**Inequality of latent distributions.** When the latent distribution was unequal across groups, the proportion of FPs in the factor loading tended to increase as the magnitude of DIF increased (see Table 5). When the sample size of both groups was equal, the proportion of FPs in the factor loading increased as sample size increased. This proportion was less than .005 only when both sample sizes equaled 200 or 100 and the DIF magnitude was medium or low, and when both sample sizes were 400 and the DIF magnitude was low. In the remaining conditions, it var-

---

\(^2\)According to Cohen and Cohen (1983), when a researcher has to set the power of a statistical test, “he is likely to set some value in the .70–.90 range” (p. 162).
ied between .009 and .126. When sample sizes were unequal, in all the conditions but one (sizes 200/100, low DIF), the FP ratio clearly exceeded the nominal alpha value. It was also observed that for a specific size of the reference group, the greater the (unequal) size of the focal group, the greater the proportion of FPs.

As to the proportion of FPs in the intercept, it increased as the DIF magnitude increased (see Table 5). When sample sizes were equal, the FP ratio generally was less than or close to .005. Only when the DIF was high, and both groups’ sample sizes greater than 100, was the FP ratio greater than .0075. When sample sizes were unequal, in all the conditions but two (200/100, low DIF; and 400/100, low DIF), the FP ratio clearly exceeded the nominal alpha value. It was also observed that for a specific size of the reference group, the greater the size of the focal group, the greater the proportion of FPs. The results reported point out that under uniform DIF conditions, when the latent distribution was unequal across groups, and the factor loading was the item parameter analyzed, the investigated procedure showed some difficulties in maintaining control of its Type I error rate, specifically when sample sizes were unequal. When sample size was equal, and the intercept was the item parameter analyzed, the investigated procedure generally controlled for Type I error.

With regard to the proportion of TPs in which DIF was correctly detected in the intercept of Item 1, the results obtained showed the same pattern as in the case of equality of latent distributions. The TP ratio increased as DIF magnitude and sample size increased (see Table 5). When the magnitude of DIF was high, the DIF item was almost always detected (the TP ratio ranged between .99 and 1). When the magnitude of DIF was medium, and both groups’ size was 200, the TP ratio equaled .76. This ratio increased when the sample size of the reference group or the sample sizes of both the reference and the focal groups increased, ranging from .86 (400/200) to 1 (800/800). Finally, when the DIF magnitude was low, the TP ratio did not exceed .44. These results indicate that the power of the procedure investigated was acceptable when the sample sizes involved were as low as 200/200 and the magnitude of uniform DIF was medium. Power increased as the magnitude of DIF and the reference group’s or both groups’ sizes increased.

Non-Uniform DIF Conditions

Equality of latent distributions. When the latent distribution was equal across groups, the proportion of FPs in the factor loading tended to increase as the magnitude of DIF increased (see Table 6). When the sample size of both groups was equal, the FP ratio tended to increase as sample size increased. It was also observed that when DIF magnitude was low, the FP ratio was less than or close to the alpha level; when DIF magnitude was medium, the FP ratio was less than the nominal alpha value only when sample size was equal to or less than 200; and finally, when DIF magnitude was high, the FP ratio ranged between .011 and .022. When
TABLE 6
Type I Error Rate and Power in Non-Uniform DIF Conditions

\[
\begin{array}{cccc|cccc|cccc|cccc|cccc|cccc}
\hline
& \multicolumn{4}{c|}{\text{Focal Group} [N (0,1)]} & \multicolumn{4}{c|}{\text{Focal Group} [N (-1,1)]} \\
\hline
\text{Reference Group} [N (0.1)] & \text{FP } \lambda & \text{FP } \mu & \text{TP } \lambda & \text{FP } \lambda & \text{FP } \mu & \text{TP } \lambda & \text{FP } \lambda & \text{FP } \mu & \text{TP } \lambda & \text{FP } \lambda & \text{FP } \mu & \text{TP } \lambda & \text{FP } \lambda & \text{FP } \mu & \text{TP } \lambda & \text{FP } \lambda & \text{FP } \mu & \text{TP } \lambda \\
\hline
\text{Equal latent distributions} & & & & & & & & & & & & & & & & & & \\
800 & Low & .001 & .017 & .290 & .006 & .011 & .150 & .006 & .007 & .090 & .007 & .007 & .06 & .001 & .017 & .290 & .006 & .011 & .150 & \\
& High & .022 & .008 & 1 & .030 & .010 & 1 & .008 & .004 & 1 & .008 & .007 & .98 & .022 & .008 & 1 & .030 & .010 & 1 & \\
400 & Low & .007 & .010 & .060 & .009 & .006 & .070 & .007 & .005 & .700 & .008 & .009 & .430 & .007 & .010 & .060 & .009 & .006 & .070 & \\
& Medium & .010 & .007 & .810 & .007 & .005 & .700 & .008 & .009 & .430 & .008 & .009 & .430 & .010 & .007 & .810 & .007 & .005 & .700 & \\
& High & .014 & .007 & 1 & .019 & .004 & 1 & .006 & .004 & .950 & .006 & .004 & .950 & .014 & .007 & 1 & .019 & .004 & 1 & \\
& Medium & .003 & .006 & .440 & .008 & .005 & .280 & .008 & .005 & .280 & .008 & .005 & .280 & .003 & .006 & .440 & .008 & .005 & .280 & \\
100 & Low & .003 & .007 & .020 & .009 & .003 & .400 & .008 & .004 & .090 & .008 & .004 & .090 & .003 & .007 & .020 & .009 & .003 & .400 & \\
& Medium & .004 & .004 & .090 & .008 & .004 & .090 & .004 & .004 & .090 & .008 & .004 & .090 & .004 & .004 & .090 & .008 & .004 & .090 & \\
& High & .012 & .004 & .820 & .012 & .004 & .820 & .012 & .004 & .820 & .012 & .004 & .820 & .012 & .004 & .820 & .012 & .004 & .820 & \\
\hline
\text{Unequal latent distributions} & & & & & & & & & & & & & & & & & & \\
& Medium & .039 & .086 & 1 & .069 & .097 & .970 & .041 & .078 & .930 & .023 & .044 & .720 & .039 & .086 & 1 & .069 & .097 & .970 & \\
& High & .072 & .100 & 1 & .096 & .117 & 1 & .062 & .113 & 1 & .027 & .103 & 1 & .072 & .100 & 1 & .096 & .117 & 1 & \\
400 & Low & .000 & .008 & .000 & .016 & .014 & .030 & .012 & .010 & .040 & .009 & .012 & .020 & .000 & .008 & .000 & .016 & .014 & .030 & \\
& High & .009 & .097 & 1 & .044 & .117 & 1 & .019 & .089 & 1 & .019 & .089 & 1 & .009 & .097 & 1 & .044 & .117 & 1 & \\
200 & Low & .000 & .003 & .000 & .009 & .012 & .020 & .009 & .012 & .020 & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .003 & .000 & \\
& Medium & .000 & .008 & .530 & .016 & .018 & .410 & .016 & .018 & .410 & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .003 & .000 & \\
& High & .000 & .062 & 1 & .019 & .078 & .990 & .019 & .078 & .990 & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .003 & .000 & \\
100 & Low & .000 & .001 & .000 & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .000 & .003 & .000 & .000 & .003 & \\
& Medium & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .000 & .003 & .000 & .000 & .003 & .000 & .000 & .003 & \\
& High & .001 & .017 & .990 & .001 & .017 & .990 & .001 & .017 & .990 & .001 & .017 & .990 & .001 & .017 & .990 & .001 & .017 & .990 & \\
\hline
\end{array}
\]

**Note.** DIF = differential item functioning; FP = incorrect DIF identification; TP = correct DIF identification.
the sample sizes of the two groups were unequal, it was generally observed that for a specific size of the reference group, the greater the size of the focal group, the greater the proportion of FPs. Moreover, when DIF magnitude was low, the FP ratio tended to be close to the alpha level (in 4 out of 6 conditions). However, when DIF magnitude was medium or high, the FP ratio tended to be greater than .0075 (in 9 out of 12 conditions). Therefore, under non-uniform DIF conditions, and when the latent distribution and sample size were the same across groups, the investigated procedure maintained control of its Type I error when DIF magnitude was low, and when it was medium and sample size was not greater than 200. However, when sample sizes were unequal, the investigated procedure only showed some control of its Type I error when DIF magnitude was low.

With regard to the proportion of FPs in the intercept, the values obtained were generally less than or close to the nominal alpha level, except when sample sizes were 800/800 and 800/400, and when sample sizes were 400/400 and DIF magnitude was low. In these cases, the PF ratio was higher than .0075. These results indicate that the investigated procedure controlled for its Type I error, except when sample sizes were equal to or greater than 400.

The proportion of TPs in which DIF was correctly detected in the factor loading of Item 1 increased as DIF magnitude and sample sizes increased (see Table 6). When DIF was high, the power was excellent. The TP ratio was .82 when sample sizes were 100/100. When the sample size of the reference group increased, the power increased to values that ranged between .93 (200/100) and .98 (800/100). For the remaining conditions with bigger sample sizes (200/200, 400/200, 800/200, 400/400, 800/400, 800/800), the TP ratio equaled 1. When the magnitude of DIF was medium, the TP ratio reached the value of .70 when sample sizes were 400/200. This ratio increased when the sample size of the reference group or the sample sizes of both the reference and the focal groups increased, ranging from .77 (for the 800/200 condition) to 0.99 (for the 800/800 condition). When the DIF magnitude was low, the TP ratio did not exceed .29. These results indicate that the power of the investigated procedure was acceptable when the sample sizes involved were as large as 400/200 and the magnitude of non-uniform DIF was medium. Power increased as the magnitude of DIF and the reference group’s or both groups’ sizes increased.

**Inequality of latent distributions.** The proportion of FPs in the factor loading tended to increase as the magnitude of DIF increased (see Table 6). When the sample size of both groups was equal, the FP ratio tended to increase as sample size increased. When both groups’ sizes were 200 or 100, the FP ratio was less than .005 across all levels of DIF magnitude. When both groups’ sizes were 400, and DIF magnitude was low or medium, the FP ratio was less than or close to the alpha level. However, when the sample size of both groups was 800, the FP ratio was greater than .01. When sample sizes were unequal, it was observed that for a spe-
specific size of the reference group, the greater the size of the focal group, the greater
the proportion of FPs. Moreover, under sample size inequality conditions, the FP
ratio was equal to or greater than .009 across all DIF magnitude levels. Thus, under
non-uniform DIF conditions, when latent distributions were unequal, and sample
size was the same across groups, the investigated procedure maintained control of
its Type I error when sample size was small (100 or 200 subjects), and when it was
medium (400 subjects) and DIF magnitude was medium or low. When sample
sizes were unequal, the DIF detection procedure lost control of its Type I error.

The proportion of FPs in the intercept tended to increase as DIF magnitude in-
creased (see Table 6). When the sample size of both groups was equal, the FP ratio
increased as sample size increased. The FP ratio was small (i.e., less than .005)
only when sample size equaled 100 and DIF magnitude was low or medium, and
when sample size was 200 and DIF magnitude was low. When sample sizes were
unequal, as in previous cases, it was observed that for a specific size of the refer-
ence group, the greater the size of the focal group, the greater the proportion of
FPs. However, the FP ratio was always equal to or greater than .01. Therefore, the
studied procedure generally did not control for its Type I error in the intercept un-
der non-uniform DIF conditions when latent distributions were distinct.

Finally, the proportion of TPs in which DIF was correctly detected in the factor
loading of Item 1 increased as DIF magnitude and sample size increased (see Table
6). When DIF magnitude was high, the DIF item was practically always detected
(the TP ratio ranged between .99 and 1). When DIF was medium, the TP ratio
reached a value of .72 when sample sizes were 800/100. This ratio increased as the
size of the focal group increased, showing values that ranged from .93 (800/200) to
1 (800/800). When the sample sizes of both groups were 400/200, the TP ratio was
.80. This ratio improved to .99 when the size of the focal group was 400 (400/400
condition). Finally, when DIF magnitude was low, the TP ratio was always close to
zero. These results point out that the power of the studied procedure was ac-
ceptable when the sample sizes involved were as large as 400/200 and the magnitude of
non-uniform DIF was medium. Power increased as the magnitude of DIF and ref-
ference group’s or both groups’ sizes increased.

DISCUSSION

The aim of the present simulation study was to assess the sensitivity of the DIF de-
tection procedure based on the MACS model when this procedure is used with
polytomous ordered response items. We carried out a simulation study in which a
number of factors were manipulated: type of DIF, magnitude of DIF, equality/in-
equality of the latent distributions, and the size of the reference and the focal
groups.
Regarding Type I error rates, the results obtained under non-DIF conditions showed that the analyzed procedure maintained control of its FP ratio when the latent distribution was equal across groups, regardless of the equality or inequality of both groups’ sample sizes. However, when the latent distribution was different, the procedure only controlled for its type I error when both the reference and the focal group were the same size. This means that when there is no DIF, there exists a certain risk of incorrectly flagging a non-DIF item as a DIF item when latent distributions and sample sizes are different across groups. Fortunately, this risk is not large, since in the worst case it amounts to 3%.

The FP ratios observed under uniform and non-uniform DIF conditions revealed that the analyzed procedure tended to better control for its Type I error when the sizes of the two groups were the same than when they were different, and when the latent distribution was equal across groups rather than when it was distinct. Regarding the impact of the latent trait distribution, some studies also found that when this distribution was different across groups, the Mantel-Haenszel and the SIBTEST procedures yielded inadequate Type I error rates (Roussos & Stout, 1996; Uttaro & Millsap, 1994). Magnitude of DIF and sample size affected the FP ratio, so that the latter tended to increase as the former increased. Thus, under uniform and non-uniform DIF conditions, the analyzed procedure lost control of its Type I Error when the latent distribution was unequal and the sizes of the two groups were different. In the other cases, the analyzed procedure tended to show some control of its Type I error, at least when DIF was low or medium and sample sizes were small (≤ 200).

As expected, the power of the analyzed procedure was affected by DIF magnitude and sample size. Power increased as the magnitude of DIF and sample size increased. With regard to sample size, it is interesting to note that when the focal group had the smallest size (100), the investigated procedure generally showed inadequate power levels when the magnitude of DIF was medium or low (although power increases as the reference group’s sample increases). When the focal group size was 200, the procedure started to show acceptable power levels when DIF magnitude was medium. A possible explanation for these results is that when the focal group’s sample is very small, item parameters would be poorly estimated if they were not constrained to be equal across groups (i.e., if they were freely estimated). Consequently, the MIs obtained under these circumstances are small, indicating that little improvement in fit could be obtained by freely estimating the constrained item parameters. If MIs are small, it is not possible to detect any DIF items by means of the investigated procedure.

Regarding the influence of the equality/inequality of the latent trait distributions, contrary to the findings reported by previous simulation studies in which the power of other DIF detection procedures was analyzed (Ankemann et al., 1999; Clauser et al., 1993; Narayanan & Swaminathan, 1996), this factor did not systematically affect the power of the MACS procedure. We computed the absolute dif-
ference between TP ratios across equality and inequality latent trait conditions. The average absolute differences were .031 and .069 for uniform and non-uniform conditions, respectively.

The MACS procedure showed an acceptable power for detecting medium uniform DIF when both groups’ sample sizes were as low as 200/200. Under these conditions, the TP ratios for equal and unequal latent trait distributions were .72 and .76, respectively. These ratios increased to .89 and .86 when sample sizes were 400/200, and to .97 and .99 when sample sizes were 400/400. When uniform DIF was high and sample sizes were the smallest considered here (i.e., 100/100), the TP ratio equaled .99; for the remaining sample size combinations, the TP ratio was 1. Considering that medium and high uniform DIF involved a difference of .25 and .50 in the item intercepts, respectively, and that the simulated items had a response scale ranging from 1 to 5, we can conclude that the MACS procedure analyzed here is very sensitive to uniform DIF, even when samples are not large. This result is congruent with previous research on the MACS procedure (Hernández & González-Romá, 2003).

In the case of non-uniform DIF, the MACS procedure showed an acceptable power for detecting medium non-uniform DIF, when both groups’ sample sizes were 400/200. Under these conditions, the TP ratios for equal and unequal latent trait distributions were .70 and .80, respectively. These ratios increased to .81 and .99 when sample sizes were 400/400. When non-uniform DIF was high and the sample sizes were 100/100, the TP ratios were .82 and .99. These ratios were .93 and .99 when sample sizes were 200/100, and they increased to 1 when the size of both groups was 200. Therefore, the power of the MACS procedure for detecting large differences (i.e., differences of .50) in the factor loading across groups is satisfactory, even when small samples are used. However, larger samples are required for adequately detecting differences of .25 in the factor loading across groups. Overall, the power results obtained point out that the MACS procedure provides adequate power levels for detecting medium uniform and non-uniform DIF, without requiring large samples.

In addition to the performance of the procedure analyzed here under the simulated conditions, other characteristics have to be considered when its use is being assessed. In comparison with Item Response Theory (IRT) methods for detecting DIF, the MACS procedure has a number of advantages (Chan, 2000; Reise, Widaman, & Pugh, 1993). First, programs performing CFA with MACS (i.e., Structural Equations Modeling [SEM] programs, such as LISREL) provide different indices of practical fit to data that are very useful for evaluating models that impose invariance constraints on item parameters across groups when sample size is large. IRT programs (e.g., MULTILOG, PARSCALE) only provide the likelihood ratio chi-square test as a measure of model fit, and this test is very sensitive to sample size. Consequently, even trivial differences on item parameters across groups may lead to model rejection. Second, when a model imposing invariance con-
straints on an item parameter cannot be maintained, modification indices provided by SEM software are very useful for detecting which particular items have parameters that are not invariant (see Chan, 2000). This facility allows researchers to specify models assuming partial invariance on the involved item parameters. IRT programs do not provide modification indices or analogues. When a model imposing invariance constraints on an item parameter cannot be maintained, a series of “item-by-item” statistical tests has to be performed to detect which items are responsible for the model misfit. In each of these tests, two models are compared: one assuming invariance across groups for all the items, and the other assuming invariance for all the items but one, whose parameters are freely estimated in each of the considered groups (see Reise et al., 1993). This procedure involves a large number of comparisons, which in turn increases the probability of Type I error.

Third, SEM programs allow researchers to work with multidimensional questionnaires (see Little, 1997), whereas IRT programs, such as MULTILOG and PARSSCALE, are suited for unidimensional questionnaires. Fourth, for many psychologists, factor analysis is a well-known technique, whereas IRT is still a much less familiar theory with complex methods. Taking into account that the MACS model is an extension of factor analysis, we think that the MACS-based procedure for DIF detection could be more easily learned by psychologists than the IRT-based methods. Finally, the MACS model only involves two parameters per item (the intercept or difficulty parameter and the factor loading or discrimination parameter), whereas IRT models for polytomous data involve a large number of parameters. This is because the latter include a number of category-related parameters instead of, or in addition to, the difficulty parameter. This implies that larger samples are needed in the latter case for an adequate item calibration.

However, the MACS-based procedure analyzed here also has some disadvantages for DIF detection compared with IRT methods. When the researcher is interested in examining between-group differences in the tendency to endorse specific response options in polytomous items, then IRT methods should be preferred (Chan, 2000). In those situations, IRT methods make it possible to test for uniform DIF associated with specific response categories.

The DIF detection procedure investigated and the IRT methods for DIF detection have a common problem. Both assume that the item chosen for parameter linkage is not a DIF item (Chan, 2000). In the case of the MACS-based procedure, an item is selected as the reference indicator, and its factor loading is set to 1 in both groups, to scale the latent variable and provide a common scale in both groups. If this assumption is incorrect (i.e., if the reference indicator is a DIF item), it can lead to inaccurate estimates of the other factor loadings (Bollen, 1989; Cheung & Rensvold, 1999), which in turn may increase the FP rate in tests of DIF involving factor loadings. Moreover, Cheung and Rensvold have shown that using a single, inappropriately chosen referent indicator can prevent identification of a DIF item, thus reducing power. These problems led Cheung and Rensvold (1999;
Rensvold & Cheung, 1998) to extend Byrne and colleagues’ procedure for testing factorial invariance (Byrne, Shavelson, & Muthén, 1989) by performing a series of tests in which all items are used as reference indicators. In practical applications of the MACS-based procedure for DIF detection (e.g., González-Romá et al., 2005), to rule out the possibility that the item used as the reference indicator might be a DIF item, the iterative procedure of DIF detection proposed by Chan (2000) was repeated using a randomly selected item. This strategy allows researchers to see whether results differ notably.

All these advantages and disadvantages, together with the performance of the method, and the type of DIF the researcher expects to detect, will have to be assessed before deciding whether using the MACS-based procedure is justified.

The present study has a number of limitations. First, data were simulated trying to produce low item skewness and kurtosis values, because under these conditions the use of normal theory ML estimation techniques in the MACS model is justified (Bollen, 1989). However, in some applications this may be an unrealistic assumption. Future research should investigate the performance of the MACS-based procedure for DIF detection when item variables show higher levels of skewness and kurtosis. Taking into account previous research on the application of ML factor analysis to ordinal data (Mooijaart, 1983; Muthén & Kaplan, 1985; Olsson, 1979), we expect that the analyzed procedure will perform worse when it is used with variables with skewness and kurtosis values larger than those considered here. In relation to this issue, we recall that the MACS model is a model for continuous variables that has been used in applied research to detect DIF in polytomous, graded response items (e.g., Chan, 2000; González-Romá et al., 2005; Wasti et al., 2000). The scarcity of studies assessing the power and Type I error of the MACS-based procedure for DIF detection when the test items are polytomous ordered response items is what motivated the present study. However, applied researchers should be aware that there is a multigroup factor model for ordinal responses (see Jöreskog, 2002). Future research should investigate the sensitivity of a DIF detection procedure based on the aforementioned model. Second, the MACS-based procedure analyzed here was a noniterative procedure. As such, all items showing significant MIs were flagged as DIF items. However, Chan (2000) applied the MACS procedure in an iterative way because Oort (1998) showed that his approach for DIF detection (the restricted factor analysis method of item bias detection) resulted in substantially more false positives when it was applied in a noniterative way than when it was applied in an iterative manner. In the MACS-based iterative procedure, initially only the item with the highest significant MI is flagged as a DIF item. After freeing the involved parameter of this DIF item, the model is re-estimated to evaluate whether there are any more items with significant MIs. If so, the item with the highest MI is flagged as a DIF item. This process continues until no sig-
significant MIs are obtained. Had this iterative approach been used in the present study, it might have reduced the number of false positives obtained. Item equality constraints were imposed across groups on the DIF item in our study, and this might have had an effect on the estimation of the parameters of the remaining non-DIF items, thus increasing the number of false positives. This effect could have been avoided by using the iterative approach. Future simulation studies will have to be carried out to assess the iterative version of the procedure analyzed here. Third, we only simulated one DIF item. This is not rare in DIF research (e.g., Ankemann et al., 1999; Zwick, Donaghue, & Grima, 1993), but it is a special case. Some studies focused on other DIF detection methods have shown that the percentage of DIF items in a test can positively affect their power and the FP ratio (e.g., Clauser et al., 1993; Narayanan & Swaminathan, 1994, 1996). Therefore, the impact of this factor on the power and Type I error rate of the MACS procedure should be addressed in future studies. Meanwhile, the findings reported here should be interpreted with caution. Fourth, as we only simulated one DIF item, we could not investigate the impact on both power and Type I error of using a DIF item as the reference indicator. Future studies, in which the percentage of DIF items is manipulated, could address this issue. Fifth, we did not manipulate test length. It is expected that with longer tests the analyzed procedure would perform better because latent scores are more accurately estimated. In Meade and Lautenschlager’s (2004) study, the multigroup CFA-based method they studied tended to show greater power when the test was made up of 12 items than when it had only 6 items, especially when the number of DIF items was low. Future studies will have to investigate the influence of test length on the functioning of the MACS procedure for DIF detection.

The results of our study have an important practical implication. Applied researchers and professionals interested in detecting medium-sized uniform and non-uniform DIF across groups, with satisfactory power levels (greater than .80) and a reasonable risk of incorrectly flagging a non-DIF item (smaller than .05), only need sample sizes of 400 for both the reference and the focal group. If the interest is only on uniform DIF, then samples of 400/200 will yield similar power and Type I error rates. Once DIF items are detected, it is advisable to assess the “practical significance” of the DIF detected (see Chan, 2000, for a simple procedure to do so), since previous research has shown that item-level DIF may not have any impact at the scale level (Chan, 2000; González-Romá et al., 2005; Reise, Smith, & Furr, 2001).

In previous empirical studies in which the MACS-based procedure for DIF detection was applied, an important issue was raised: the need to conduct simulation studies to check the power of the aforementioned procedure (Chan, 2000; González-Romá et al., 2005). In the present study, we have contributed to beginning to clarify this issue.
REFERENCES


Accepted January 2005