PERGAMON

Personality and Individual Differences 36 (2004) 1459–1470

PERSONALITY AND INDIVIDUAL DIFFERENCES

www.elsevier.com/locate/paid

# Sex differential item functioning in the Raven's Advanced Progressive Matrices: evidence for bias

Francisco J. Abad*, Roberto Colom, Irene Rebollo, Sergio Escorial

*Facultad de Psicología, Universidad Autónoma de Madrid, 28049 Madrid, Spain*

Received 15 July 2002; received in revised form 8 April 2003; accepted 8 June 2003

**Abstract**

There are no sex differences in general intelligence or *g*. The Progressive Matrices (PM) Test is one of the best estimates of *g*. Males outperform females in the PM Test. Colom and García-López (2002) demonstrated that the information content has a role in the estimates of sex differences in general intelligence. The PM test is based on abstract figures and males outperform females in spatial tests. The present study administered the Advanced Progressive Matrices Test (APM) to a sample of 1970 applicants to a private University (1069 males and 901 females). It is predicted that there are several items biased against female performance, by virtue of their visuo-spatial nature. A double methodology is used. First, confirmatory factor analysis techniques are used to contrast one and two factor solutions. Second, Differential Item Functioning (DIF) methods are used to investigate sex DIF in the APM. The results show that although there are several biased items, the male advantage still remains. However, the assumptions of the DIF analysis could help to explain the observed results.


*Keywords:* Sex differences; Advanced Progressive Matrices Test; General intelligence; Differential Item Functioning

## 1. Introduction

There are several meta-analyses demonstrating that there is a sex difference in some cognitive abilities. The first meta-analysis was published by Hyde (1981) from the data summarized by Maccoby and Jacklin (1974) and showed that boys outperform girls in spatial and mathematical ability, but that girls outperform boys in verbal ability. Hyde and Linn (1988) found that females outperform males in several verbal abilities. Hyde, Fennema, and Lamon (1990) found a male

* Corresponding author. Tel.: +34-397-41-14.
*E-mail address:* fjose.abad@uam.es (F.J. Abad).

0191-8869/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0191-8869(03)00241-1

advantage in quantitative ability, but those researchers noted that many quantitative items are expressed in a spatial form. Linn and Petersen (1985) found a male advantage in spatial rotation, spatial relations, and visualization. Voyer, Voyer, and Bryden (1995) found the same male advantage in spatial ability, being the most important sex difference in spatial rotation. Feingold (1988) found a male advantage in reasoning ability. Thus, research findings support the idea that the main sex difference may be attributed to overall spatial performance, in which males outperform females (Neisser et al., 1996).

However, verbal, quantitative, or spatial abilities explain less variance than general cognitive ability or *g*. *g* is the most general ability and is common to all the remaining cognitive abilities. *g* is a common source of individual differences *in all cognitive tests*. Carroll (1997) has stated "*g*′... is likely to be present, in some degree, in nearly all measures of cognitive ability. Furthermore, it is an important factor, because on the average over many studies of cognitive ability tests it is found to constitute more than half of the total common factor variance in a test" (p. 31).

A key question in the research on cognitive sex differences is whether, on average, females and males differ in *g*. This question is technically the most difficult to answer and has been the least investigated (Jensen, 1998). Colom, Juan-Espinosa, Abad, and García (2000) found a negligible sex difference in *g* after the largest sample on which a sex difference in *g* has ever been tested ($N = 10{,}475$). Colom, García, Abad, and Juan-Espinosa (2002) found a null correlation between *g* and sex differences on the Spanish standardization sample of the WAIS-III. Those studies agree with Jensen's (1998) statement: "in no case is there a correlation between subtests' *g* loadings and the mean sex differences on the various subtests the *g* loadings of the sex differences are all quite small" (p. 540). This means that cognitive sex differences result from differences on specific cognitive abilities, but not from differences in the core of intelligence, namely, *g*.

If there is not a sex difference in *g*, then the sex difference in the best measures of *g* must be non existent. The Progressive Matrices (PM) Test (Raven, Court, & Raven, 1996) is one of the most widely used measures of cognitive ability. PM scores are considered one of the best estimates of general intelligence or *g* (Jensen, 1998; McLaurin, Jenkins, Farrar, & Rumore, 1973; Paul, 1985). If there is not a sex difference in *g*, males and females must obtain similar scores in the PM Test. However, Lynn (1998) has reported evidence supporting the view that males outperform females in the Standard Progressive Matrices Test (SPM). He considered data from England, Hawaii, and Belgium. The average difference was equivalent to 5.3 IQ points favouring males. Colom and García-López (2002), and Colom, Escorial, and Rebollo (submitted) found a sex difference in the APM (Advanced Progressive Matrices) favouring males: 4.2 IQ and 4.3 IQ points, respectively. Those findings do not support the view that males and females do not differ in *g*.

Previous findings show that there is no sex difference in *g*. However, there is a small but consistent sex difference in one of the best measures of general intelligence, namely, the PM Test. Colom and García-López's (2002) findings support the view that the information content has a role in the estimates of sex differences in general intelligence. They concluded that "researchers must be careful in selecting the markers of central abilities like fluid intelligence, which is supposed to be the core of intelligent behavior . A "gross" selection can lead to confusing results and misleading conclusions" (p. 450). Although the PM test is routinely considered the "essence" of fluid *g*, this is a doubtful. Gustaffson (1984, 1988) has demonstrated that the PM Test loads on a first order factor which he nominates as "Cognition of Figural Relations" (CFR). This evidence is supported by our own research (Colom, Palacios, Rebollo, & Kyllonen, submitted). We performed a hierarchical factor analysis and obtained a first order factor loaded by Surface development, Identical pictures, and the APM. This factor is a mixture of Gv and Gf. Thus, the male advantage on the Raven could come from its Gv ingredient. It must be remembered that the highest difference between the sexes is in spatial performance. Could the spatial content of the PM Test explain the sex difference?

The factors underlying performance on the PM Test have been analysed from both the psychometric and cognitive perspectives. Carpenter, Just, and Shell (1990) suggest that several items can be solved by perceptually based algorithms such as line continuation, while other items involve goal management and abstraction. There is some evidence to argue that the PM test is a multi-componential measure. Embretson (1995) distinguishes the working memory capacity aspects from the general control processes related to the meta-ability to allocate cognitive resources. Verguts, De Boeck, and Maris (2000) explored the abstraction ability. Those researchers applied a non-compensatory multidimensional model, the conjunctive Rasch model, in which higher scores on one factor cannot compensate low scores on other factors. Anyway, these studies conceive performance across items as a function of a homogeneous set of basic operations.

However, the most studied type of multidimensionality is related to the visuo-spatial basis of the PM test. Hunt (1974) identified two general problem solving strategies that could be used to solve the items, one visual—applying operations of visual perception, such as superimposition of images upon each other—and one verbal—applying logical operations to features contained within the problem elements. Carpenter et al. (1990) found five rules governing the variation among the entries of the items: constant in a row, quantitative pairwise progression, figure addition or substraction, distribution of three values, and distribution of two values. DeShon, Chan, and Weissbein (1995) consider that Carpenter et al. (1990) discount the importance of the visual format of the PM test. Following Hunt (1974) those researchers developed an alternative set of visuospatial rules that may be used to solve several items: superimposition, superimposition with cancellation, object addition/subtraction, movement, rotation, and mental transformation. They classified 25 APM Set II items as purely verbal-analytical or purely visuo-spatial. The remaining items required both types of processing or were equally likely to be solved using both strategies.

Lim's (1994) factor analysis suggests that APM could measure different abilities in males and females. Some APM item factor analyses were conducted by Dillon, Pohlmann, and Lohman (1981) suggesting that two factors are needed to explain item correlations. One factor was interpreted to be an ability to solve problems whose solutions required adding or subtracting patterns, while the other factor was interpreted as an ability to solve problems whose solutions required detecting a progression in a pattern. However, several researchers (Alderton & Larson, 1990; Arthur & Woehr, 1993; Bors & Stokes, 1998; Deshon et al., 1995) reported results indicating that the APM is unidimensional. But there are some problems in these studies. Alderton and Larson (1990) used two samples of male Navy recruits, while Deshon et al. (1995) and Bors and Stokes (1998) administered the APM to a sample composed mostly of females (64%). Furthermore, they administered the APM with a time limit of 40 minutes. Bors and Stokes's (1998) two-factor solution suggests that the second factor was a speed factor. Additionally, Bors and Stokes (1998), Arthur and Woehr (1993), and Deshon et al. (1995) studied small samples to estimate the tetrachoric correlation matrices they analysed.

Although Dillon et al.'s (1981) bi-factor structure has been validated by others, Deshon et al. (1995) proposal has not been investigated further. Their results make it plausible that some APM items could be biased by its visuo-spatial content (see the classical study by Burke, 1958). We propose that several APM items claim for visuo-spatial strategies. This fact could help to explain sex differences on the PM Test. To test this possibility, we used a double methodology. First, we applied traditional confirmatory factor analysis techniques to contrast one and two factor solutions. Second, we applied current Differential Item Functioning methods (Holland & Wainer, 1993; Thissen, Steinberg, & Gerrard, 1986) to investigate sex Differential Item Functioning (DIF) in APM items. The finding of sex DIF in one item means that after grouping participants with respect to the measured ability, sex differences on item performance remains. It must be emphasized that, to our knowledge, DIF analysis has never been applied to the PM Test.

## 2. Method

### 2.1. Participants, measures, and procedures

The participants were applicants for admissions to a private university. They were 1970 adults (1069 males and 901 females), ranging in age from 17 to 30 years. Each participant completed the Advanced Progressive Matrices Test, Set II, in a group self-administered format. Following general instructions and practice problems, the APM was administered with a 40-min time limit.

The mean APM score for the total sample was 23.53 (S.D. = 5.47). The mean score for males was 24.19 (S.D. = 5.37) and for females it was 22.73 (S.D. = 5.47). The sex difference was equivalent to 4.03 IQ points. Of the sample, 65.3% completed the test and 93% (irrespective of sex) completed the first 30 items. In order to avoid a processing speed factor, we selected these 30 items and excluded all the participants that did not complete the test. The final sample comprised 1820 participants (985 males and 835 females). The mean score for the total sample was 21.87 (S.D. = 4.65). For males the mean score was 22.45 (S.D. = 4.52) and for females it was 21.19 (S.D. = 4.72). The sex difference in IQ points was unaffected by the data selection (4.06 IQ points). The correlation between APM scores and sex was significant ($r = -0.134$; $P < 0.000$) and similar to previous studies (Arthur & Woehr, 1993; Bors & Stokes, 1998).

### 2.2. Statistical analyses

#### 2.2.1. Structural equation modelling

A matrix of tetrachoric interitem correlations calculated by the PRELIS computer program (Joreskog & Sorbom, 1989) was used as input for the confirmatory factor analyses (diagonally weighted least squares). The LISREL computer program was used (Joreskog & Sorbom, 1989). Three models were directly evaluated. Dillon et al.'s and DeShon et al.'s two factor models (correlated or independent) were evaluated against a one dimensional model. Our predictions are that Dillon et al.'s model (First factor: items 7, 9, 10, 11, 16, 21 & 28; second factor: items 2, 3, 4, 5, 17 & 26) will not fit data better than the one dimensional model, while DeShon et al.'s model (Verbal analytical factor: items 8, 13, 17, 21, 27, 28, 29 & 30; visuo-spatial factor: items 7, 9, 10, 11, 12, 16, 18, 22, 23 & 24) could fit data slightly better.

### 2.2.2. DIF analysis

We analysed one dimensionality criteria in order to apply IRT (Item Response Theory) DIF methods. We fitted the two parameter logistic model (2PL). In the 2PL models the probability of a correct response is a function of the item difficulty $b$ (the point on the item curve characteristic—ICC—at which there is a 50% probability of correct response), the item discrimination $a$ (proportional to the slope of the ICC) and the person ability $\theta$:

$$P(1|\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

We followed an iterative and conceptually guided procedure. To compute a DIF procedure, an unbiased variable to group participants must be used as an anchor criterion.

First, the BILOG computer program was used (Mislevy & Bock, 1990) to check the 2PL model item's fit to the data (separately for males and females). Second, we selected the eight items from APM that only can be solved by analytical strategies (items 8, 13, 17, 21, 27, 28, 29 and 30—see Deshon et al., 1995). In fact, items 1 and 4 were analytical too, but Deshon et al. (1995) excluded them from their analysis. Third a multi-step procedure was applied to find one purified anchor criterion. The steps were (Kim & Cohen, 1992):

1. Estimate item parameters separately for males and females for these 8 items.
2. Link parameter metrics across groups. Items parameters estimated in different groups will differ by a linear transformation from one group to another. Thus, one needs to equate the metrics. The linear coefficients were computed by the test characteristic curve method (Stocking & Lord, 1983).
3. Estimate DIF indexes and remove DIF items. We used the Lord's chi-square statistic that simultaneously tests the hypothesis that $a$ and $b$ for one item are identical across groups (Lord, 1980).
4. Re-link group metrics using only non-DIF items.
5. Re-estimate DIF indexes and remove DIF items.

Steps 4 and 5 are continued until either no DIF items are detected or until the same items are identified as DIF items on two successive iterations. The Iterlink program (Stark & Chernyshenko, 2001) automatically performs steps 2 to 5.

After identifying a subset of items without DIF (the anchor variable), we performed two analyses:

1. We re-estimated parameters by the MULTILOG computer program (Thissen, 1991) and verified the assumption of invariance checking the model fit loss when we fix equal parameters across groups.
2. Once this assumption is checked, we performed an item by item DIF analysis. The procedure we used is well described in Thissen et al. (1986) and Maller (2001). It is based in the likelihood ratio goodness of fit statistic, $G^2$, which is distributed as chi-square. For each item we checked (separately for $a$ and $b$ parameters) the model fit loss (changes in $G^2$) when we fix equal parameters across groups (which is the same as the nested models comparison

in structural equation modelling). The first hypothesis tested was about the discrimination parameters equality. If we reject this equality it means the presence of non uniform DIF (e.g., items discriminate worse for females than for males). If there was not "non uniform DIF" we test the second hypothesis about difficulty parameters equality. Rejecting this equality for one item we confirm "uniform DIF" for it (e.g., items are easier for males than for females).

To examine the magnitude of the DIF effect, the root mean squared probability difference (RMSD) was computed as follows:

$$\mathrm{RMSD} = \sqrt{\sum_{q=1}^{25} g(\theta_q)(p_{\mathrm{males}}(\theta_q) - p_{\mathrm{females}}(\theta_q))^2}$$

where squared probability difference between groups is evaluated across 25 ability points assuming that $\theta$ is normally distributed, $N(0,1)$. The value of 0.05 or greater is frequently used by the Educational Testing Service to indicate DIF (O'neill & McPeck, 1993; Maller, 2001).

## 3. Results

Fit results for the model evaluations are presented in Table 1. The Dillons et al.'s two-factor solution in which the factors are assumed to be independent does not provide an adequate fit to the data (see AGFI and NNFI indexes). If we free the inter-factors correlation, fit model increases significantly with respect to the one factor solution fit ($\chi^2(1) = 10.42$; $P = 0.001$). However the inter-factors correlation is high (0.89). Furthermore, the other fit indices (NNFI and AGFI), less

Table 1
Goodness of Fit Indices for the single-factor and two-factors models[a]

| Model comparison: | $\chi^2$ | df | $P$ | AGFI | NNFI | RMSEA |
|---|---|---|---|---|---|---|
| *Dillon et al. model (13 items)* | | | | | | |
| One factor | 75.67 | 65 | 0.17 | 1.00 | 1.00 | 0.010 |
| Two correlated factors | 65.25 | 64 | 0.43 | 1.00 | 1.00 | 0.00 |
| Two independent factors | 537.76 | 65 | 0.00 | 0.85 | 0.89 | 0.030 |
| *DeShon et al. model (18 items)* | | | | | | |
| One factor | 315.64 | 135 | 0.00 | 0.99 | 0.99 | 0.024 |
| Two correlated factors | 314.58 | 134 | 0.00 | 0.99 | 0.99 | 0.024 |
| Two independent factors | 1246.14 | 135 | 0.00 | 0.83 | 0.85 | 0.036 |
| *Final model (30 items)* | | | | | | |
| One factor | 954.78 | 405 | 0.00 | 0.98 | 1.00 | 0.020 |

[a] Diagonally Weighted Least Squares Confirmatory factor analysis with the tetrachoric correlation matrix.

sensitive to sample size, do not increase. Thus the one factor solution fits enough to this APM subset composed of 13 items. On the other hand, DeShon's et al. bifactorial model only fits data if we assume factors are correlated. But then, the inter-factor estimated correlation increases to 0.97 [i.e, the solution is almost one-dimensional; the chi-square's difference between both of them was non significant $\chi^2$ (1) = 1.06; $P = 0.30$]. The one factor solution fits nicely (See NFFI and RMSEA indexes)

It seems that these results do not support DeShon's et al. classification rules for items. With respect to Dillon et al.'s taxonomy, differences between one-dimensional and two-oblique-factor models are small. Thus, the one-factor model has an adequate fit. The first factor explained 28.72% of the variance. It could be concluded that the essential one-dimensionality assumption applied to IRT models is satisfied in this APM's data.

The 2PL model fits the data [for males, $\chi^2(199) = 263.0$; $\chi^2/\text{gl} = 1.32$; for females, $\chi^2$ $(220) = 265.7$; $\chi^2/\text{gl} = 1.21$]. The 2PL model fits all the items for females and 28 out of 30 items ($\alpha = 0.01$) for males.

We then applied the 2PL model to the eight items that, following Deshon et al.'s taxonomy, can only be solved by analytic strategies. After applying the multi-step procedure, we deleted items showing DIF [item 8: $\chi^2(2) = 6.08$; $P = 0.048$; item 3: $\chi^2(2) = 10.11$; $P = 0.006$]. For the remaining six anchor items, we tested the equality $a$ and $b$ parameters' assumption using MULTILOG and concluded that differences are not significant [$\chi^2(12) = 10.5$; $P = 0.57$].

Once we obtained an analytic items anchor test, we performed DIF analysis with MULTILOG. Results are shown in Table 2. The second and third columns show the rules needed to solve the item. The second column corresponds to Carpenter et al.'s (1990) taxonomy, while the third column describes visual strategies from Deshon et al. (1995). The fourth column indicates the factor classification by Dillon et al. (1981).

Table 2 presents the difference in $G^2$ between the free model and the model with the parameter constraint. The fifth and sixth columns show the test for the hypothesis $a_{\text{boys}} = a_{\text{girls}}$ and $b_{\text{boys}} = b_{\text{girls}}$. The $b$ differences are made conditional to the $a$ differences contrast (we assume the $a$ parameter contrast results establish the null model). The last columns show $a$ and $b$ final parameters for each group (conditioned to the previous statistical tests).

If we focalise items with RMSD greater than 0.04 we can observe that:

1. 45% of visual-strategy items have non-uniform DIF, and are easier for males. There is no significant discrimination difference. Other items that require visuo-spatial processing (item 20) or that are likely to be solved using a visuo-spatial strategy (item 14) show a similar result.
2. In the analytic test, two out of four items (items 8 and 30) from the non-anchor test show significant non ignorable DIF. These items are easier for women and, according to DeShon et al.'s (1995) classification, they can be solved relying in a verbal strategy. One of them shows little discrimination in this group.

Figs. 1–3 show representations for item characteristic curves in the DIF items of the Raven for three typical examples for the several groups we made.

Table 2
Results for differential item functioning

| It. | Verbal-analytic strategy (DeShon et al., 1995) | Visual strategy | Dillon (1981) | $G^2$ $a=a$ | $G^2$ $b=b$ | RMSD | a fem | a mal | b fem. | b mal. |
|---|---|---|---|---|---|---|---|---|---|---|
| *Verbal-analytic strategy items* | | | | | | | | | | |
| 1 | Distribution of 3, constant | | | 1.6 ($P=0.206$) | 2.4 ($P=0.121$) | 0.0000 | 0.92 | | −3.18 | |
| 4 | pairwise,constant | | D2 | 4.1 ($P=0.043$) | 1.4 ($P=0.237$) | 0.0344 | 1.36 | 1.76 | −2.1 | |
| 8 | Distribution of 3 | | | 6.2 ($P=0.013$) | 6.7 ($P=0.010$) | 0.0784 | 0.61 | 1.19 | −2.86 | −1.6 |
| 13 | Distribution of 3, constant | | | | | | 0.55 | | −1.80 | |
| 17 | Distribution of 3, constant | | D2 | | | | 0.79 | | −1.55 | |
| 21 | Distribution of 3, constant | | D1 | | Anchor test | | 1.07 | | −0.80 | |
| 27 | Distribution of 3 | | | | | | 1.12 | | 0.44 | |
| 28 | Distribution of 3 | | D1 | | | | 0.76 | | 1.32 | |
| 29 | Distribution of 3 | | | | | | 1.20 | | 1.09 | |
| 30 | Distribution of 2 | | | 0.8 ($P=0.371$) | 5.6 ($P=0.018$) | 0.0556 | 0.74 | | −0.63 | −0.29 |
| *Visual strategy items* | | | | | | | | | | |
| 3 | pairwise,constant | Movement | D2 | 0.1 ($P=0.752$) | 3.2 ($P=0.074$) | 0.0000 | 1.43 | | −2.45 | |
| 7 | Addition | Superimposition | D1 | 0.1 ($P=0.752$) | 4.4 ($P=0.036$) | 0.0433 | 1.47 | | −1.6 | −1.85 |
| 9 | Addition, constant | Superimposition | D1 | 2.8 ($P=0.094$) | 0.8 ($P=0.371$) | 0.0000 | 1.53 | | −2.32 | |
| 10 | pairwise, constant | Movement | D1 | 0.0 ($P=1$) | 5.7 ($P=0.017$) | 0.0537 | 1.65 | | −1.19 | −1.43 |
| 11 | | Addition/Subt. | D1 | 2.3 ($P=0.129$) | 0.0 ($P=1$) | 0.0000 | 1.78 | | −1.99 | |
| 12 | Substraction | Sup with cancel. | | 0.9 ($P=0.343$) | 0.5 ($P=0.480$) | 0.0000 | 1 | | −2.21 | |
| 16 | Substraction | Sup with cancel. | D1 | 0.3 ($P=0.584$) | 0.3 ($P=0.584$) | 0.0000 | 1 | | −1.56 | |
| 18 | | Mental Transf. | | 0.0 ($P=1$) | 0.1 ($P=0.752$) | 0.0000 | 0.85 | | −0.85 | |
| 22 | Distribution of 2 | Sup with cancel. | | 0.0 ($P=1$) | 5.2 ($P=0.023$) | 0.0535 | 0.93 | | −0.16 | −0.43 |
| 23 | Distribution of 2 | Sup with cancel. | | 1.3 ($P=0.254$) | 8.2 ($P=0.004$) | 0.0665 | 1.01 | | −0.52 | −0.85 |
| 24 | | Movement | | 0.0 ($P=1$) | 5.1 ($P=0.024$) | 0.0523 | 0.94 | | 0.28 | 0.02 |
| *Both strategy items* | | | | | | | | | | |
| 19 | visual with condition | Superimposition | | 0.7 ($P=0.403$) | 0.1 ($P=0.752$) | 0.0000 | 0.69 | | −2.15 | |
| 20 | visual with condition | Superimposition | | 4.3 ($P=0.038$) | 4.6 ($P=0.032$) | 0.0674 | 0.63 | 0.29 | −1.44 | −3.15 |
| 25 | visual with condition | Movement | | 0.7 ($P=0.403$) | 1.8 ($P=0.180$) | 0.0000 | 0.83 | | −0.57 | |
| 26 | visual with Distrib. of 3 | Rotation | D2 | 0.1 ($P=0.752$) | 1.3 ($P=0.254$) | 0.0000 | 0.69 | | 0.16 | |
| *Either strategy items* | | | | | | | | | | |
| 2 | Pairwise, constant | Expansion | D2 | 2.1 ($P=0.147$) | 0.3 ($P=0.584$) | 0.0000 | 1.98 | | −2.17 | |
| 5 | Pairwise, constant | Continuation | D2 | 0.8 ($P=0.371$) | 1.0 ($P=0.317$) | 0.0000 | 1.23 | | −2.44 | |
| 6 | Pairwise, constant | Movement | | 0.8 ($P=0.371$) | 60.9 ($P=0.009$) | 0.0383 | 0.82 | | −3.2 | −3.83 |
| 14 | Pairwise, constant | Movement | | 0.8 ($P=0.371$) | 90.4 ($P=0.002$) | 0.0619 | 1.2 | | −1.56 | −1.95 |
| *Uncodable item 15* | | | | 0.2 ($P=0.655$) | 0.3 ($P=0.584$) | 0.0000 | 0.84 | | −1.66 | |

## 4. Discussion

The performed statistical analyses have shown that the Advanced Progressive Matrices Test is one-dimensional. It seems that this measure of reasoning ability does not require other cognitive abilities to a significant degree.

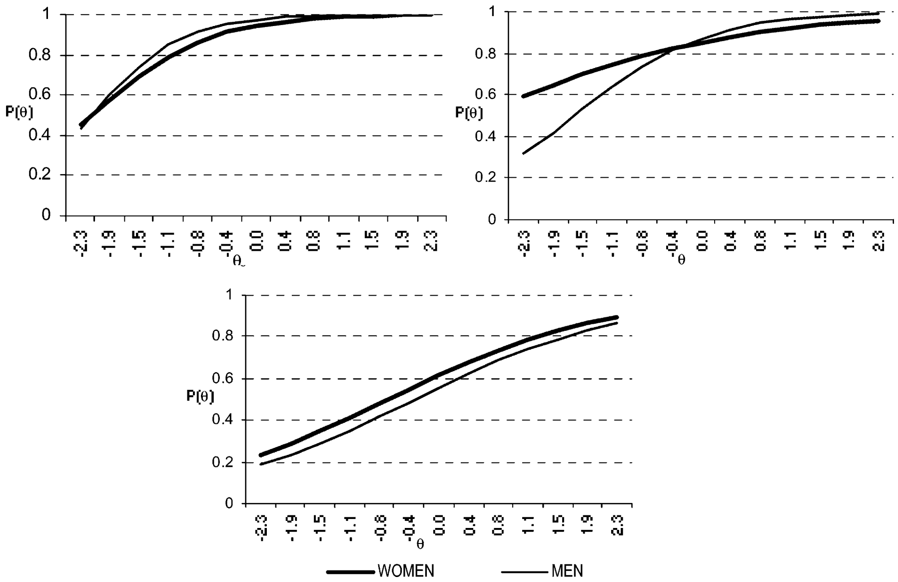


Fig. 1. Pure analytic items: (a.) item 4, item 8 and item 30.

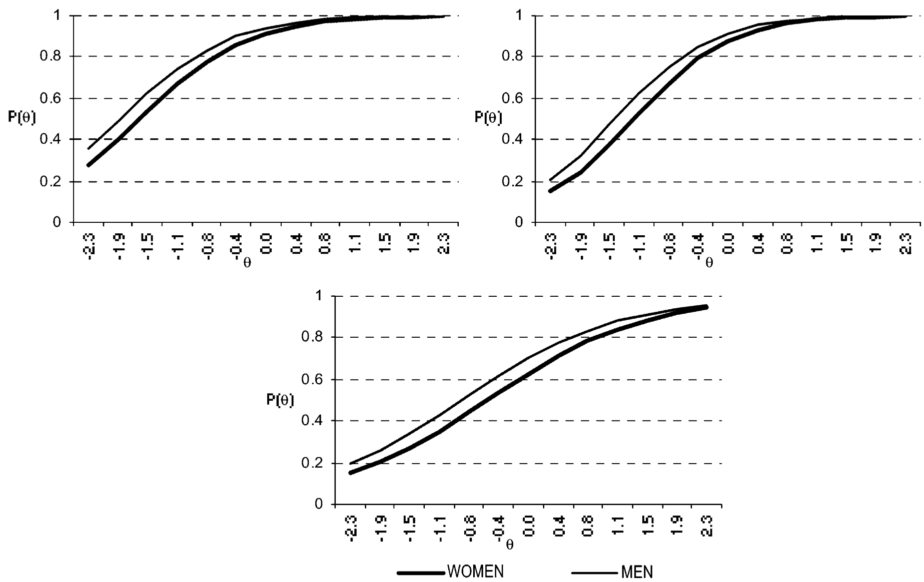


Fig. 2. Pure visuo-spatial items: item 7, item 10 and item 23.

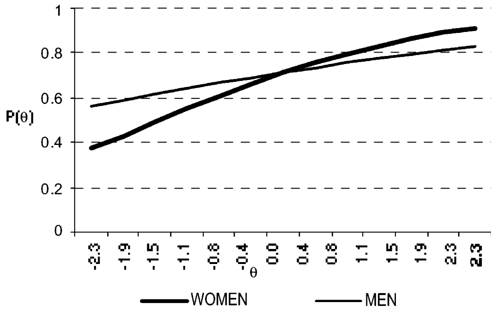


Fig. 3. Both strategies item: item 20.

However, the one-dimensional nature of the APM does not imply that some items could not be biased against female performance. There is a sex difference in the PM Test (Colom and García-López, 2002; Lynn, 1998). However, given that this test is based on abstract figures and that males have on average a higher spatial ability than females (Voyer et al., 1995), we predicted that some items may be easier for males. Thus, males might solve some items due to their visuo-spatial nature. This could be considered as a threat for bias.

The findings showed that although there are several unbiased items, some are clearly biased against female performance. The bias will be higher when male–female differences in spatial performance are specially pronounced. We found that DIF matters for several APM visuo-spatial items. Deleting the biased items reduces the sex difference from 4.06 (4.53 after correcting by the reliability of the test of 30 items, 0.806) to 3.32 IQ points (3.96 corrected by reliability of the test of 19 unbiased items, 0.707). Thus, the male advantage is still present and significant (the correlation between APM scores and sex was −0.110, $P < 0.000$).

However, the male advantage can be a by-product of the analysis performed. The DIF analysis assumes that the anchor test comprises unbiased items measuring the relevant construct. Through DIF analyses we can only demonstrate the former assumption, but not the latter. It could happen that the 19 remaining unbiased items still measure visuo-spatial ability to some extent. This hypothesis cannot be tested with the present dataset. It will be necessary to apply the APM unbiased items with other ability tests in order to analyze its position in the factor space.

However, this suggestion is supported by Colom et al.'s (submitted) findings. Those researchers administered the APM and the spatial rotation subtest from the Primary Mental Abilities Battery to a sample of psychology undergraduates. Males outperformed females on both tests. However, the male advantage on the APM turned out to be non significant when differences in spatial ability were statistically controlled. Furthermore, Lim's (1994) results strongly agree with this statement; he found that females showed an additional APM loading in a spatial factor. Their results should be replicated with the 19 unbiased APM items found in the present study.

In summary, researchers have investigated the visuo-spatial basis of the PM Test. The male advantage on this Test could come from its visuo-spatial nature. The present study has rejected the idea that this advantage comes from particular biased items likely to be solved by visuo-spatial strategies. These findings suggest additional analyses to test the position of the APM unbiased items in the factor space. The prediction is that they will still show loadings on spatial factors, especially for females.

## References


Alderton, D. L., & Larson, G. E. (1990). Dimensionality of Raven's Advanced Progressive Matrices items. *Educational and Psychological Measurement*, *50*, 887–900.

Arthur, W., & Woehr, D. J. (1993). A confirmatory factor analytic study examining the dimensionality of the Ravens' advanced progressive Matrices. *Educational and Psychological Measurement*, *53*, 471–478.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*, 382–398.

Burke, H. R. (1958). Raven's progressive matrices: A review and critical evaluation. *Journal of Genetic Psychology*, *93*, 199–228.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, *98*, 404–431.

Carroll, J. B. (1997). Psychometrics, intelligence, and public perception. *Intelligence*, *24*, 1 25-52.

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high-school graduates. *Personality and Individual Differences*, *32*, 3 445-451.

Colom, R., Juan-Espinosa, M., Abad, F., & García, L.F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*, 57–68.

Colom, R., García, L. F., Abad, F. J., & Juan-Espinosa, M. (2002). Null sex differences in general intelligence: evidence from the WAIS-III. *Spanish Journal of Psychology*, *5*, 1 29-35.

Colom, R., Escorial, S., & Rebollo, I. (submitted). *Sex differences in the Raven are influenced by sex differences in spatial ability*.

Colom, R., Palacios, A., Rebollo, I., & Kyllonen, P.C. (submitted). *Working memory is (almost) perfectly predicted by g*.

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence*, *21*, 135–155.

Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement*, *41*, 1295–1302.

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence*, *20*, 169–189.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, *43*, 95–103.

Gustafsson, J. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179–203.

Gustafsson, J. (1988). Hierarchical models of individual differences in cognitives abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, vol. 4*. Nueva Jersey: LEA.

Holland, W. H., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hunt, E. (1974). Quote the Raven? nevermore!. In G. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Erlbaum.

Hyde, J. (1981). How large are cognitive gender differences? A meta-analysis using $w^2$ and d. *American Psychologist*, *36*, 892–901.

Hyde, J., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: a meta-analysis. *Psych. Bull*, *107*, 139–153.

Hyde, J., & Linn, M. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53–69.

Jensen, A. R. (1998). *The g factor*. London: Praeger.

Joreskog, K. G., & Sorbom, D. (1989). *LISREL 7: user's reference guide*. Mooresville, IN: Scientific Software.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 51–66.

Lim, T. K. (1994). Gender related differences in intelligence: application of confirmatory factor analysis. *Intelligence*, *19*, 179–192.

Linn, M., & Petersen, A. (1985). Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child development*, *56*, 1479–1498.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lynn, R. (1998). Sex differences in intelligence: a rejoinder to Mackintosh. *Journal of Biosocial Science*, *30*, 529–532.

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

Maller, S. J. (2001). Differential item functioning in the WISC-III: item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, *61*, 793–817.

McLaurin, W., Jenkins, J., Farrar, W., & Rumore, M. (1973). Correlations of IQ on verbal and non-verbal tests of intelligence. *Psychological Reports*, *22*, 821–822.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: item analysis and test scoring with binary logistic models [computer program]*. Mooresvilles, IN: Scientific Software.

Neisser, U., Boodoo, G., Bouchard, T. J., Boyking, W., Brody, N., Ceci, S., Halpern, D., Loehlin, J., Stenberg, R., & Urbina, S. (1996). Intelligence: knowns and unknowns. *American Psychologist*, *51*, 2 77-101.

O'neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In H. Wainer, & H. I. Braun (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Paul, S. M. (1985). The advanced Raven's Progressive Matrices: normative data for an American university population and examination of the relationship with Spearman's g. *Journal of Experimental Education*, *54*, 95–100.

Raven, J.C, Court, J. H., & Raven, J. (1996). *Manual del test de matrices progresivas de Raven [Handbook of the Matrices Progressives Tests]*. Madrid: TEA.

Stark, S., & Chernyshenko, S. (2001). Methods for detecting differential item/test. In F. Drasgow (Ed.), *A Practical Guide To IRT: introduction to item response theory analyses and applications*. San Diego, CA.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

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

Thissen, D. (1991). *MULTILOG: Multiple Categorical item analysis and test scoring using item response theory (version 6)*. Chicago: Scientific Software.

Verguts, T., Boeck, P. D., & Maris, E. (2000). Generation speed in Raven's progressive matrices test. *Intelligence*, *27*, 329–345.

Voyer, D., Voyer, S., & Bryden, M. (1995). Magnitude of sex differences in spatial abilitites: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270.