Learning and Individual Differences 22 (2012) 390–396

Contents lists available at SciVerse ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Item Response Theory analysis and Differential Item Functioning across age, gender and country of a short form of the Advanced Progressive Matrices

Francesca Chiesi [a,*], Matteo Ciancaleoni [a], Silvia Galli [a], Kinga Morsanyi [b], Caterina Primi [a]

[a] Department of Psychology, University of Florence, Italy
[b] University of Plymouth, UK

ARTICLE INFO

Article history:
Received 25 September 2011
Received in revised form 26 November 2011
Accepted 10 December 2011

Keywords:
Advanced progressive matrices
Short form
Item response theory
Differential item functioning
Fluid ability

ABSTRACT

Item Response Theory (IRT) models were applied to investigate the psychometric properties of the Arthur and Day's Advanced Progressive Matrices-Short Form (APM-SF; 1994) [Arthur and Day (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement, 54*, 395–403] in order to test if the scale is a reliable and valid tool to assess general fluid ability in a short time frame. The APM-SF was administered to 2264 high-school and university students. Once attested the one-factor structure of the scale, unidimensional IRT analyses for dichotomous data were applied to investigate the increases in item difficulty levels, Test Information Function, and Differential Item Functioning across age, gender, and country (comparing Italian and British respondents). Additionally, validity measures were reported. Findings attest that the Arthur and Day's APM-SF is a sound instrument for assessing fluid ability within a short time frame.

1. Introduction

Referring to very up to date data, a large survey conducted in nineteen European countries by several members of the International Test Commission (ITC) attested that the Raven's Matrices are in the fourth position among the ten most used tests in Europe (Evers, 2011). Among them, the Advanced Progressive Matrices (APM; Raven, 1962) are widely employed to assess fluid ability in adolescents and adults (Raven & Raven, 2008). Nonetheless, a potential limitation is the length of the test (the APM are composed of two sets: Set I consists of 12 items, generally used as a practice test, and Set II consists of 36 items) which might increase the influence of interfering variables, such as fatigue and boredom, decrease the motivation of respondents, and make it troublesome to use the instrument in a test battery.

To shorten the administration time, Hamel and Schmittmann (2006) proposed a timed version (20 min) of the APM. The main problem with the timed version is that it is not a pure measure of fluid ability as it is influenced by intellectual efficiency (Raven, Raven, & Court, 1993). Moreover, reducing administration time results in the exclusion of the most difficult items. In order to avoid these problems, two short forms of the APM have been proposed (Arthur & Day, 1994; Bors & Stokes, 1998).

Arthur and Day (1994) selected items by first dividing the APM-Set II into sections of three items each, and, for each section, the item with the highest item-total correlation was chosen. Ties were reconciled by selecting the most difficult item or the item with the largest drop in internal consistency if deleted from the test. A 12-item scale was obtained and administered to samples of university students in order to test its factor structure, the progressive item difficulty, and its reliability (Arthur & Day, 1994; Arthur, Tubre, Paul, & Sanchez-Ku, 1999; Day, Espejo, Kowollik, Boatman, & McEntire, 2007). A confirmatory factor analysis indicated that a single-factor model adequately represents the structure of the short form, and satisfactory indices of internal consistency and test–retest were obtained. Finally, the results attested that the overall progressive difficulty structure of the long form was maintained in the short form although three items appeared to be out of order.[1]

Bors and Stokes (1998) developed another 12-item short form of the APM. They argued that the Arthur and Day's short form presented potential redundancies in item characteristics, and that the first two items added noise to the total scores, since these were too easy. Administering the APM to university students, the items of Set II with relatively high item-total correlations and low inter-item correlations were selected. Following this procedure, they mainly selected items belonging to the middle-range of the ranking of the APM. The unidimensionality of the scale was attested applying explorative factor analysis, and adequate reliability indices (internal consistency, and test–retest) were reported. More recently, another study (Vigneau & Bors, 2005) tested the psychometric properties of the Bors and Stokes's short form obtaining results that do not fully

* Corresponding author at: Department of Psychology, via di San Salvi 12, Padiglione 26, 50135 Firenze Italy. Tel.: +39 0556237846; fax: +39 0556236047.
E-mail address: francesca.chiesi@unifi.it (F. Chiesi).

[1] Namely, Item 4 was harder than the following item, Item 23 and Item 25 were easier than the preceding item (numbers refer to the original form).

1041-6080/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.lindif.2011.12.007

replicate the previous ones. A modest internal consistency was found, and, whereas the principal component analysis supported the view that the scale measures one construct, applying Item Response Theory (IRT), specifically the Rasch model, the scale did not show a one-factor structure. Finally, the increasing levels of item difficulty were not fully confirmed.

The aim of the present study was to provide evidence that the Advanced Progressive Matrices-Short Form (APM-SF) by Arthur and Day (1994) might be employed as a sound assessment tool for measuring general fluid ability within a short time frame. The short form developed by Arthur and Day, as opposed to the Bors and Stokes's one, was chosen, based on the following issues. First, since one property of the Raven's Matrices is learning from experience during the test (Raven, Raven, & Court, 1998), long forms have been arranged such that each item should be progressively more difficult than the preceding item along a wide range of difficulty levels (from very easy to very difficult). Arthur and Day's strategy of sequentially dividing the test into sections should guarantee high similarity to the long form because this procedure specifically aimed to retain low, medium, and high difficulty items. Bors and Stokes, using the discriminative power of the items as a criterion, obtained a more uniform scale (i.e., all items belong mainly to the middle-range of the difficulty continuum of the APM) which does not entirely represent the original scale features. Second, Vigneau and Bors (2005) failed to confirm the psychometric properties of the short form of Bors and Stokes founding an unclear factor structure and low reliability.

The present work aimed to extend the study of the psychometric properties of the Arthur and Day's APM-SF applying the IRT models since until now the characteristics of the scale have been investigated using classical test theory (Arthur & Day, 1994; Arthur et al., 1999; Day et al., 2007). Moreover, this is in line with the International Test Commission's recommendations for the proper description and evaluation of existing and widely used psychological instruments (e.g., Muñiz, 2011).

Several studies on the full forms of the Raven's Matrices applied IRT models (Abad, Colom, Rebollo, & Escorial, 2004; Çikrikçi-Demirtaşli, 2000; Gallini, 1983; Georgiev, 2008; Raven, Prieler, & Benesch, 2005; van der Ven & Ellis, 2000). Nevertheless, there is no agreement regarding the most suitable IRT model. Some authors (Gallini, 1983; Raven et al., 2005) suggested that the three-parameter model (3PL) – difficulty, discrimination, and guessing – is preferable because there is a guessing component due to the multiple-choice format of the matrices. Some others (Çikrikçi-Demirtaşli, 2000; Georgiev, 2008) argued that guessing is irrelevant as each matrix has eight response options, and they opted for the two-parameter model (2PL). Starting from this disagreement, we first aimed to ascertain which model is the most suitable for examining the APM-SF.

Then, we tested the increasing order of item difficulty. Indeed, the IRT is extremely useful in ascertaining this test characteristic since, in contrast with the classical test theory, the IRT allows for obtaining item difficulty estimates that are metrically equivalent across samples. Consequently, it is possible to test increases in item difficulty levels, considering the item parameter as a unique property of the item which is independent from the characteristics of the respondents. Additionally, the IRT has another notable property which makes it suitable for short forms. When using a short form of a test, following the classical test theory, reliability levels are unavoidably reduced, as it relies heavily on the number of items of the test. The IRT allows for examining reliability through the Test Information Function (TIF) which basically tells how well the test is doing in estimating ability over the whole range of ability scores. That is, rather than a single value (e.g., coefficient alpha) for reliability, IRT recognizes that measurement precision can be different at different levels of the ability trait. The peak of the TIF is where measurement precision is greatest and, since the amount of information is defined at the item level, even a small number of good items might provide a good test reliability, at least for certain trait levels. In this particular case, a rather flat TIF is desirable indicating that the item set of APM-SF is highly discriminating within a broad range of ability.

Once we identified the properties of the APM-SF applying IRT, the equivalence of item parameter estimates across age, gender and country was investigated in order to provide further evidence for the soundness of the short scale. Within the framework of the IRT, Differential Item Functioning (DIF) is central to the investigation of the measurement equivalence of a scale at the item level. DIF makes it possible to test if the items measure the same trait dimension when administered to two or more distinct groups controlling for true group mean differences. Indeed, to compare groups of individuals with regard to their level on a trait one must be sure that the numerical values that quantify that trait are on the same measurement scale. For this reason, DIF testing is a procedure recommended by the International Test Commission for the proper evaluation of the psychometric characteristics of existing and widely employed psychological instruments (e.g., Muñiz, 2011).

With regard to age, since until now the APM-SF has only been tested with university students (the minimum age of respondents was 18 years) we aimed to check if the scale maintains the same psychometric properties when administered to younger respondents. Two groups were created dividing the whole sample into younger (from 14 to 17 years of age) and older (from 18 years of age) respondents. According to the norms of the APM (Raven et al., 1998), the former group is expected to display poorer performance than the latter, and thus, by testing measurement equivalence, we aimed to ascertain if the scale was equally suitable for people with both lower and higher levels of the measured trait. More specifically, DIF might be present in either the discrimination (a) or difficulty parameter (b). No DIF in the a-parameter would suggest that the items are equally related to theta for older and younger participants, whereas DIF in the b-parameter would suggest that some items were endorsed at different rates by older and younger participants, controlling for trait level.

Similarly, knowing that the same norms are applied for male and female respondents (Raven et al., 1998), we aimed to test the equivalence of the scale across gender. Contradictory empirical evidence has been reported regarding gender differences on the Raven's Progressive Matrices. Some claimed that there is a difference[2] (for a review, Lynn & Irwing, 2004), while others stated that there is not (e.g., Colom & Garcìa-Lopez, 2002; Court, 1983; Jensen, 1998). Despite this controversy, which goes beyond the scope of the present paper, we expected to find the items of the Arthur and Day's short form metrically equivalent for males and females. Gender DIF analysis was performed on the full form of the APM (Abad et al., 2004) finding that males have an advantage on some items, and females on others. The authors stated that this represents a potential bias in measuring general fluid ability, so we aimed to ascertain whether Abad et al. 's (2004) findings were confirmed performing gender DIF on the APM-SF.

Finally, knowing that the Raven Matrices are considered to show less influence of cultural factors and cross-national differences than other intelligence tests (see Brouwers, Van der Vijver, & Van Hemert, 2009 for a brief review), we aimed to ascertain if the APM-SF items were metrically equivalent across different countries. To do this, we administered the scale to Italian and British respondents.

Concerning validity, the core goal of any short form test should be to replicate the pattern of relationships established for the construct as measured by the long form of the test (Smith, McCarthy, & Anderson, 2000). For this reason, we explored the relations of the APM-SF scores with several variables chosen as follows.

[2] Gender differences have been related to and explained by the distinguishable processes that some authors deem measured by Raven's Matrices (see Footnote 3) (e.g., Colom, Escorial, & Rebollo, 2004; Mackintosh & Bennett, 2005).

First, as the relationship between *g* and working memory (WM) has been widely demonstrated (see Yuan, Steedle, Shavelson, Alonzo, & Oppezzo, 2006, for a review), we expected to find a positive correlation between APM-SF scores and WM measures.

Second, to gain further insight into the validity of the scale, relationships with different aspects of reasoning were taken into account knowing that fluid ability has been found to be correlated with mechanical reasoning, (e.g., Colom, Juan-Espinosa, & Garcia, 2001; Haavisto & Lehto, 2004), mathematical reasoning (e.g., Lynn & Irwing, 2004; Waltz et al., 1999), and probabilistic reasoning (e.g., Kahneman & Frederick, 2002; Stanovich & West, 2000).

Third, several studies investigated the relationship between intelligence and scholastic achievement (e.g., Furnham & Chamorro-Premuzic, 2004; Jensen, 1998; Luo, Thompson, & Detterman, 2003; Pind, Gunnarsdottir, & Johannesson, 2003; Saccuzzo & Johnson, 1995) that is considered a good criterion for validating intelligence tests (Deary, Strand, Smith, & Fernandez, 2007). Nonetheless, successful school performance depends on many personal characteristics others than intelligence such as persistence, interest in school, and willingness to study. Moreover, what students learn in school depends not only on their individual abilities but also on teaching practices and on what is actually taught. Then we explored the relation between APM-SF scores and grades as achievement measures knowing that its strength might be reduced by the educational framework heterogeneity in our sample.

2. Method

2.1. Participants

Participants were 1956 Italian high school and university students (55% male, age range from 14 to 40 years, M = 19.31, SD = 4.08) distributed by gender and age as follows: 398 14-to-16-year-olds (male = 64%), 462 17-to-18-year-olds (male = 57%), 494 19-to-20-year-olds (male = 46%), 246 21-to-24-year-olds (male = 50%), 243 25-to-40-year-olds (male = 46%). Additionally, aiming to test the cross-country national equivalence of the scale, 308 English university students (22% male, age range from 18 to 39 years, M = 20.77, SD = 3.43) participated to the study. University students from both countries attended psychology schools.

2.2. Measures

The Advanced Progressive Matrices-Short Form (APM-SF, Arthur & Day, 1994) is composed of items 1, 4, 8, 11, 15, 18, 21, 23, 25, 30, 31 and 35 of the APM-Set II. The respondent's task is to determine which of eight possible alternatives fits into the missing space so that row and column rules are satisfied. Consistently with the long form (composed of Set I which is used as a practice test, and Set II representing the actual test), 3 items derived from Set I of the APM were used for practice before completing the APM-SF.

Working memory was measured through the Digit Span scale of the WAIS-R (Wechsler, 1997). Participants were presented with sequences of numbers and asked to repeat some of them forward and others backward. The series begin with two digits and keep increasing in length, with two trials at each length.

To measure probabilistic reasoning ability we used tasks derived from the heuristics and biases literature (Gilovich, Griffin, & Kahneman, 2002) and employed in previous studies (Chiesi, Primi, & Morsanyi, 2011; Morsanyi, Primi, Chiesi, & Handley, 2009). Mathematical reasoning was measured through the Prerequisti di Matematica per la Psicometria scale (PMP; Galli, Chiesi, & Primi, 2011), a test constructed applying IRT analyses to measure basic mathematical abilities. To measure mechanical reasoning ability, the Bennett Mechanical Comprehension Test (BMCT; Bennett, 1969) was used. For each scale a composite score was calculated summing correct answers.

In order to measure scholastic achievement, the students had to report their final high school grades (range 60–100), and their final high school math grades (range 0–10). The final examination grade (range 0–30) in an introductory statistics course attended by some of the university students participating to the study was used as an indicator of academic achievement.

2.3. Procedure

Each participant individually completed the APM-SF in a self-administered format. The administration time ranged from 10 to 20 min with an average time of 15 min. Additionally, 966 participants completed the probabilistic reasoning tasks, 202 the Bennett Mechanical Comprehension Test, and 151 the Prerequisti di Matematica per la Psicometria scale. For all these tests answers were collected in a paper-and-pencil format, each one was briefly introduced to the students and instruction for completion was given. The Digit Span scale was individually administered to a subsample of 652 students. Finally, measures of achievement were registered for 1157 high-school students and for 115 university students.

2.4. Data analyses

As a first step, we tested the dimensionality of the APM-SF in order to select the most suitable IRT analysis. The one-factor structure of the scale was tested with a Confirmative Factor Analysis (CFA) for dichotomous data using Mplus 3.0 software (Muthén & Muthén, 2004) that implemented the Weighted Least Squares Means And Variance Adjusted (WLSMV) estimation method. This method is recommended for categorical variables (Flora & Curran, 2004; Muthén & Muthén, 2004) on the basis of simulation studies (Muthén, duToit, & Spisic, 1997).

Then, IRT analyses were performed, following three steps. First, to ascertain whether the 2PL or the 3PL model better describes the data, the differences of -2loglikelihood (-2log) for nested models were computed using Marginal Maximum Likelihood estimation method implemented in Multilog software (Thissen, 1991). The difference between the statistics for hierarchical models is distributed as chi-square and may be used to test the significance of additional parameters. A significant difference implies that the model with more parameters provides a superior fit to the data. Second, to evaluate the items fit for both IRT models the statistics $S-\chi 2$ were computed using the GOODFIT software (Orlando, 1997). This statistic is based on joint likelihood distributions for each possible total score (Lord & Wingersky, 1984; Thissen, Pommerich, Billeaud, & Williams, 1995). The observed proportions of individuals responding with a particular response (e.g., correct) for each total score group are then compared to expectations based on joint likelihood distributions that consider the likelihood for each total score *t* across all possible response patterns for *t*. Because large samples lead to a greater likelihood of significant chi-square differences the critical value at $p = .01$, rather than the one at $p = .05$, was used (Stone & Zhang, 2003). Third, to provide a picture of the performance of the single items and the global scale, item parameters and the Test Information Function (TIF) were reported.

For Differential Item Functioning (DIF), the Item Response Theory Likelihood Ratio test approach (IRTLR; Thissen, Steinberg, & Wainer, 1988) was used. The first step was to identify a group of DIF-free anchor items that could be used to link the two participants groups in terms of their ability. Anchor items were generated in an iterative manner using IRTLRDIF software (Thissen, 2001). IRTLRDIF uses maximum likelihood to test a model in which all parameters are constrained to be equal for a studied item, to a model in which one or more parameters are freely estimated. Under an iterative procedure,

Table 1
Parameters (a = discrimination, b = difficulty, c = guessing) under the 3PL model, factor loadings (all significant at $p<.001$), and Maximum Item Information Function under 3PL model for each item of the APM-SF.

Item[a]	a	b	c	Factor loading	$IIF_{MAX}(\theta)$
1	.97	−1.23	.00	.67	.68 (−1.20)
4	1.22	−.38	.16	.69	.78 (−.20)
8	.86	−1.04	.00	.63	.54 (−1.00)
11	1.50	−.42	.03	.81	1.53 (−.40)
15	.80	−.24	.02	.62	.44 (−.20)
18	.91	−.04	.02	.67	.58 (.00)
21	1.35	.77	.11	.63	1.07 (.80)
23	.72	.30	.00	.60	.37 (.40)
25	.62	.40	.02	.53	.27 (.40)
30	.87	.87	.07	.58	.47 (1.00)
31	.92	1.20	.08	.53	.52 (1.20)
35	.75	.93	.01	.60	.41 (1.00)

[a] Numbers refer to the original APM-Set II.

all items are initially evaluated for DIF by sequentially testing the two models. The difference between the log-likelihood statistics of the two models is distributed as a chi-square statistic. A significant chi-square statistic at $p<.05$ level indicates that at least one of the parameters differs between groups and is assumed to demonstrate DIF. After removing items with potential DIF, the procedure was repeated as many times as necessary until purification of the anchor set was achieved. Next, the remaining set of test items was sequentially evaluated for DIF against the set of purified anchor items. To adjust for multiple comparisons, the Bonferroni's correction was used. Finally, in order to test the magnitude of the DIF, the Non Compensatory DIF index (NCDIF, Raju, 1999) was used. NCDIF reflects differences in the conditional probabilities of response to an item randomly selected individuals form the two groups. Items demonstrating DIF will have values of NCDIF higher than the cutoff .006 (Raju, 1999). Raju's (1998) DFIT4 program was used to compute NCDIF indices.

3. Results

3.1. Dimensionality, increasing item difficulty, and reliability

In line with the theoretical framework which guided the development of the Raven Matrices (i.e., the unidimensional theory of intelligence by Spearman, 1927), the APM-SF was expected to have a one-factor structure, just as the long forms[3] (see Raven & Raven, 2008 for a review). Confirming previous studies (Arthur & Day, 1994; Arthur et al., 1999), the data indicated that a single-factor model adequately represents the structure of the APM-SF. Specifically, the CFI = .98 and the TLI = .98 indicated a very good fit (Bentler, 1990) as well as the RMSEA = .03 (Browne & Cudeck, 1993). Factor loadings were all significant ($p<.001$), ranging from .53 to .81 (Table 1).

Given the one-factor structure of the scale, unidimensional IRT analyses were performed and the results showed that the 3PL model had a better fit ($\Delta\text{-}2\log_{2PL-3PL}(12) = 37.9$, $p<.001$). All items had a non- significant $S-\chi^2$ under the 3PL model whereas under the 2PL model one item showed a poor fit ($p=.004$). Thus the 3PL model appeared to be the preferable one.

Parameters estimated under the 3PL model were reported in Table 1. The discrimination parameters were between 0.62 and 1.50. That is, the items showed medium to large discrimination levels (Baker, 2001). The item difficulty parameters ranged between −1.23 and 1.20. Three items appeared out of order, i.e. they were more difficult than expected following their ranking. Namely, Items 4 and Item 21 were harder than the following items, and Item 31 was harder than Item 35 (numbers refer to the original form). Finally, the contribution of guessing to the probability of correct response was greater than zero for some items. Nonetheless, c values were under .35, cut-off above which the amount of guessing is not considered acceptable (Baker, 2001). Namely, one item had a guessing parameter of .16, the guessing parameter was around .10 for four items, and very close to 0 for five items.

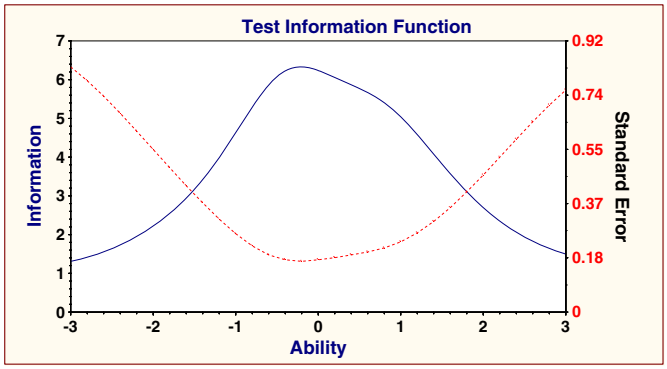

Fig. 1. Test Information Function of APM-SF under the 3PL model.

The TIF attested that the instrument was sufficiently informative and that the item set of APM-SF is well discriminating within a broad range of ability (see Fig. 1). The standard error was particularly low for medium levels of trait, and the maximum value of the amount of test information was approximately 6.4 (corresponding to an approximate ability level of −0.2[4]). Within the range of ability from −1.0 to +1.0, the amount of test information was greater than 4. Within the wider range from −1.5 to +1.5, it was greater than 3. In sum, whereas under the three-parameter model guessing parameters greater than zero lower the amount of test information, and, despite the small number of items, the TIF attested a satisfactory amount of precision for the measures obtained by administering the APM-SF.

3.2. Measurement equivalence across age, gender and country

Preliminarily, the unidimensionality of the scale and the item fit under the 3PL model in all groups (younger, older, male, female, British) were tested in order to verify the possibility of using unidimensional IRT and the 3PL model for the DIF analysis. The results showed that the unidimensionality assumption was met. Specifically, across the five groups CFI and TLI ranged from .98 to .99 indicating a very good fit (Bentler, 1990), just as the RMSEA that ranged from .04 to .03 (Browne & Cudeck, 1993). Factor loadings (Table 2) were significant ($p<.001$). Finally, for each group all items had an acceptable fit under the 3PL model.

The DIF analysis across age revealed that one item displayed a significant difference. This item was designated as a study item for the DIF analysis. The eleven remaining items were identified as anchor items. Item 23 showed DIF for the difficulty parameter (b) as indicated by the significant difference between the -2loglikelihood (Table 3). The NCDIF index was .02. Its magnitude confirmed that Item 23 showed non-ignorable DIF. Specifically, parameters indicated that it was easier for older respondents. Nonetheless, since only one item exhibits DIF (less than 10% of the total number of items that composed the scale; Budgell, Raju, & Quartetti, 1995), the APM-SF can be considered equivalent across age.

[3] Recently it has been proposed that the Raven's Matrices tests measure at least two distinguishable processes, usually called perceptual and analytic (see Kunda, McGregor, & Goel, 2010; Mackintosh & Bennett, 2005 for a brief review). In answer to this assumption, Raven and Raven (2008) stated that the heterogeneity of item difficulty characteristics led to extract factors that simply reveal "dimensions" of difficulty, but the matrices measure the *same* underlying continuum of ability.

[4] Theta (measured on the same scale of b) has a mean of zero and a SD of 1.0, e.g. an ability level of −0.2 can be understood to mean 0.20 SD below the mean.

Table 2
Standardized factor loadings (all significant at $p<.001$) of APM-SF for each age, gender, and country group.

Item	Age		Gender		Country	
	Younger	Older	Male	Female	England	Italy
1	.56	.63	.67	.66	.60	.67
4	.61	.65	.73	.62	.73	.69
8	.55	.60	.65	.59	.53	.63
11	.77	.79	.80	.82	.77	.81
15	.54	.59	.62	.62	.57	.62
18	.54	.66	.70	.63	.47	.67
21	.42	.64	.63	.64	.58	.63
23	.45	.63	.67	.52	.38	.60
25	.43	.51	.56	.49	.33	.53
30	.39	.57	.55	.59	.49	.57
31	.30	.53	.50	.58	.25	.53
35	.52	.57	.65	.53	.62	.60

Table 3
Differential Item Functioning analyses of APM-SF across age (14–17 year-olds vs. 18–40 year-olds), gender (males vs. females), and country (British vs. Italian).

	Age		Gender		Country	
Item	Δ-2log$_a$	Δ-2log$_b$	Δ-2log$_a$	Δ-2log$_b$	Δ-2log$_a$	Δ-2log$_b$
1	Anchor item		1.1 ($p=.294$)	4.3 ($p=.038$)	0.0 ($p=.999$)	3.1 ($p=.078$)
4	Anchor item		Anchor item		Anchor item	
8	Anchor item		Anchor item		Anchor item	
11	Anchor item		Anchor item		0.0 ($p=.999$)	15.2 ($p=.001^*$)
15	Anchor item		Anchor item		Anchor item	
18	Anchor item		Anchor item		Anchor item	
21	Anchor item		0.7 ($p=.403$)	3.2 ($p=.074$)	Anchor Item	
23	3.3 ($p=.069$)	7.9 ($p=.005^*$)	Anchor item		Anchor item	
25	Anchor item		Anchor item		Anchor item	
30	Anchor item		Anchor item		Anchor item	
31	Anchor item		Anchor item		Anchor item	
35	Anchor item		Anchor item		Anchor item	

* $p<\alpha=.05/4=.013$ (Bonferroni's correction).

The DIF analysis across gender revealed that no items displayed a significant difference, attesting the measurement equivalence of the APM-SF for males and females (Table 3).

The DIF analysis across country revealed that one item displayed a significant difference. Item 11 showed DIF for the difficulty parameter (b) (Table 3). Nonetheless, the NCDIF of .005 attested that DIF was negligible.

In conclusion, the APM-SF can be considered metrically equivalent across age, gender and country.

3.3. Validity measures

Pearson product-moment correlations attested that all the investigated relations were significant. Values appear to be in line with the values reported in previous studies employing the APM, and they are adequate measures of validity following the cut-offs proposed by the European Federation of Psychologists' Association (EFPA) (Muñiz, 2009).[5] Specifically, the correlation between the APM-SF and digit span scores was $r(N=653)=.42$, $p<.001$.The correlations between the APM-SF score and the reasoning measures were $r(N=921)=.35$, $p<.001$ for probabilistic reasoning, $r(N=202)=.27$, $p<.001$ for mechanical reasoning, and $r(N=151)=.47$, $p<.001$ for mathematical reasoning. The correlations with the achievement measures indicated that the APM-SF was related to high school final grades ($r(N=126)=.25$, $p<.01$), final mathematics grades ($r(N=1157)=.24$, $p<.001$), and final introductory statistics grades ($r(N=115)=.27$, $p<.01$).

4. Discussion

Due to the wide use of the Raven's Progressive Matrices to assess fluid ability (for an up to date report, see Evers, 2011), the aim of the present study was to provide evidence that the short form of the APM developed by Arthur and Day (1994) might be employed for a sound assessment of general fluid ability within a short time frame. Once the one-factor structure of the scale had been attested, IRT analyses suggested that the 3PL model was preferable over the 2PL model in line with previous studies that indicated that the three-parameter model was the most suitable to analyze the full form of Raven's Matrices (Gallini, 1983; Raven et al., 2005). This means that the 12 items of the APM-SF are characterized by and display differences in difficulty, discrimination, and guessing.

Guessing needs to be taken into account for some items in order to adjust observed scores for the impact of chance. This is probably due to the fact that although the matrices have eight response options, in some cases the choice might be between fewer alternatives (e.g., the chance of guessing "correctly" might be not one over eight, but one over three or four). Nonetheless, since the estimated guessing parameters were low (the maximum values was .16), all the items can be considered adequate. The discrimination parameters attested that the items are informative and precise in assessing the respondents' ability. The difficulty parameters indicated a substantial variation among the items, in accordance with the long version of the test whereas the ranking does not correspond to a perfect increase in difficulty levels since three items (Item 4, Item 21, and Item 31) were harder than the following items in the scale. Nonetheless, these findings, with the exclusion of Item 31, are in line with those reported by Arthur and Day (1994) applying classical test theory (see Footnote 1), and IRT analyses confirmed the need to modify their rank.

Concerning reliability, the Test Information Function, taking into account difficulty, discriminative power, and guessing, revealed that the 12-item APM scale is adequately accurate in measuring the latent trait across a quite broad range of ability.

Additionally, the scale was proved to be age, gender, and country metrically equivalent. The analysis of age DIF attested the equivalence of the APM-SF items across younger and older respondents. That is, the test allows for accurate discrimination between respondents with different levels of the trait regardless of their age. As expected, the analysis of gender DIF attested the equivalence of the APM-SF items across male and female respondents. That is, the test allows for discrimination between respondents with different levels of the trait regardless of whether they are males or females, and, in contrast with previous results (Abad et al., 2004), no gender-related advantages or disadvantages were found (i.e., no item appeared to be easier or harder for males than females). In line with the claim that Raven's Matrices can be considered to be mostly unaffected by cultural factors or cross-national differences (Brouwers et al., 2009), the analysis of country DIF attested the equivalence of the APM-SF item parameters across Italian and British respondents.

Finally, the results provide evidences to the validity of the scale as a short form of the Raven's APM replicating the relationships previously reported for the long form. Scores obtained administering the APM-SF were found to be positively correlated with the working memory (see e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Fry & Hale, 1996), different kind of reasoning (see e.g., Colom et al., 2001; Haavisto & Lehto, 2004; Stanovich & West, 2000; Waltz et al., 1999), and scholastic/academic achievement (see e.g., Brody, 1992; Pind et al., 2003; Saccuzzo & Johnson, 1995).

[5] For the criterion-related validity values between .20 and .35 are deemed adequate, values between .35 and .50 are good, and values higher than .50 are excellent.

Further studies might confirm and extend the present findings. First of all, following the ranking empirically identified in the present study, future investigations should be conducted administering the reordered scale in order to stress the potential effect of that reordering on the test performance. Then, applying multivariate analyses, such as Structural Equation Modeling, the nomological validity of fluid ability as measured by the short form of the APM might be explored. Additionally, Differential Item Functioning across other countries might also be investigated, as well as the adequateness of the scale when used with clinical populations.

In sum, the present study supports the adequacy of the Arthur and Day's short form of the APM as a measure of general fluid ability. Given the reduced administration time, the scale provides a suitable and efficient tool for researchers and practitioners.

References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences*, *36*, 1459–1470.

Arthur, W., & Day, D. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, *54*, 395–403.

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Matrices Test. *Journal of Psychoeducational Assessment*, *17*, 354–361.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). USA: ERIC Clearinghouse on Assessment and Evaluation.

Bennett, C. K. (1969). *Bennett mechanical comprehension test.* San Antonio: The Psychological Corporation.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for the first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*(3), 382–398.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego: Academic Press.

Brouwers, S. A., Van der Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in Raven's progressive matrices scores across time and place. *Learning and Individual Differences*, *19*, 330–338.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, *19*, 309–321.

Chiesi, F., Primi, C., & Morsanyi, K. (2011). Developmental changes in probabilistic reasoning: The role of cognitive capacity, instructions, thinking styles and relevant knowledge. *Thinking and Reasoning*, *17*, 315–350.

Çikrikçi-Demirtaşli, N. (2000). A study of Raven Standard Progressive Matrices Test's item measures under Classical and Item Response Model. *Paper presented at 31st European Mathematical Psychology Congress, Austria: Graz.*.

Colom, R., Escorial, S., & Rebollo, I. (2004). Sex differences on the Progressive Matrices are influenced by sex differences on spatial ability. *Personality and Individual Differences*, *37*(6), 1289–1293.

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

Colom, R., Juan-Espinosa, M., & Garcia, L. F. (2001). The secular increase in test scores is a "Jensen effect". *Personality and Individual Differences*, *30*, 553–559.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163–183.

Court, J. C. (1983). Sex differences in performance on Raven's Progressive Matrices. *Alberta Journal of Educational Research*, *29*, 54–74.

Day, E. A., Espejo, J., Kowollik, V., Boatman, P. R., & McEntire, L. E. (2007). Modeling the links between need for cognition and the acquisition of a complex skill. *Personality and Individual Differences*, *42*, 201–212.

Deary, I. J., Strand, S., Smith, P., & Fernandez, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13–21.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology. General*, *128*, 309–331.

Evers, A. (2011, July). Testing practices and attitude towards tests and testing: The results of a global survey. *Paper presented at the 12th European Congress of Psychology, Istanbul.*

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*, 237–241.

Furnham, A., & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and Individual Differences*, *37*, 943–955.

Galli, S., Chiesi, F., & Primi, C. (2011). Measuring mathematical ability needed for "non-mathematical" majors: The construction of a scale applying IRT and differential item functioning across educational contexts. *Learning and Individual Differences*, *21*, 392–402.

Gallini, J. K. (1983). A Rasch analysis of Raven item data. *The Journal of Experimental Education*, *52*(1), 27–32.

Georgiev, N. (2008). Item analysis of C, D and E series form Raven's Standard Progressive Matrices with Item Response Theory two-parameter logistic model. *Europe's Journal of Psychology* http://www.ejop.org/archives/2008/08/item_analysis_o.html

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment.* New York: Cambridge University Press.

Haavisto, M., & Lehto, J. E. (2004). Fluid/spatial and crystallized intelligence in relation to domain specific working memory: A latent variable approach. *Learning and Individual Differences*, *15*, 1–21.

Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, *66*(6), 1039–1046.

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Kunda, M., McGregor, K., & Goel, A. (2010). Taking a look (literally!) at Raven's intelligence test: Two visual solution strategies. *Proceeding of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1691–1696)..

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating". *Applied Psychological Measurement*, *8*, 453–461.

Luo, D., Thompson, L. A., & Detterman, D. K. (2003). The casual factor underlying the correlation the psychometric g and scholastic performance. *Intelligence*, *31*, 67–83.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: a meta-analysis. *Intelligence*, *32*, 481–498.

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices by African and White university students in South Africa. *Intelligence*, *28*, 251–265.

Morsanyi, K., Primi, C., Chiesi, F., & Handley, S. J. (2009). The effects and side-effects of statistics education. Psychology students' (mis-)conceptions of probability. *Contemporary Educational Psychology*, *34*, 210–220.

Muñiz, J. (2009, July). The role of EFPA in setting standards for tests and test use. *Paper presented at the 11th European Congress of Psychology, Oslo*.

Muñiz, J. (2011, July). International strategies to improve tests and testing. *Paper presented at the 12th European Congress of Psychology, Istanbul.*

Muthén, B. O., duToit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*, *75*.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus: The comprehensive modeling program for applied researchers. User's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

Orlando, M. (1997). *Item fit in the context of item response theory. Doctoral dissertation, University of North Carolina.* Dissertation Abstracts International, 58/04-B, 2175.

Pind, J., Gunnarsdottir, E. K., & Johannesson, H. S. (2003). Raven's standard progressive matrices: New school age norms and a study of the test's validity. *Personality and Individual Differences*, *34*, 375–386.

Raju, N. S. (1998). *DFITD4: A Fortran program for calculating dichotomous DIF/DTF [computer program].* Chicago: Illinois Institute of Technology.

Raju, N. S. (1999). *Some notes on the DFIT framework.* Chicago: Illinois Institute of Technology.

Raven, J. C. (1962). *Advanced progressive matrices.* London: Lewis & Co. Ltd..

Raven, J. C., Prieler, J., & Benesch, M. (2005). *A replication and extension of the item-analysis of the Standard Progressive Matrices Plus, together with a comparison of the results of applying three variants of Item Response Theory.* WebPsychEmpiricist http://wpe.info/papers_table.html

Raven, J. C., & Raven, J. (Eds.). (2008). *Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics*. Unionville, New York: Royal Fireworks Press.

Raven, J., Raven, J. C., & Court, J. H. (1993). *Raven manual section 1: General overview.* Oxford: Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual section 4: Advanced Progressive Matrices.* Oxford: Oxford Psychologists Press.

Saccuzzo, D. P., & Johnson, N. F. (1995). Traditional psychometric tests and proportionate representations: An intervention and program evaluation study. *Psychological Assessment*, *7*, 181–194.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short form development. *Psychological Assessment*, *12*(1), 102–111.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*, 645–665.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331–352.

Thissen, D. (1991). *MULTILOG user's guide.* Chicago: SSI.

Thissen, D. (2001). Computer software. Retrieved from. http://www.unc.edu/~dthissen/dl.html

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item response theory scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 149–169). Hillsdale, NJ: Lawrence Erlbaum.

van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's Standard Progressive Matrices. *Personality and Individual Differences*, *29*, 45–64.

Vigneau, F., & Bors, D. A. (2005). Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement*, *65*(1), 109–123.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*, 119–125.

Wechsler, D. (1997). *Scala d'Intelligenza Wechsler per Adulti - Riveduta*. Wechsler Adult Intelligence Scale—Revised. Firenze: Organizzazioni Speciali.

Yuan, K., Steedle, J., Shavelson, R., Alonzo, A., & Oppezzo, M. (2006). Working memory, fluid intelligence, and science learning. *Educational Research Review*, *1*, 83–98.