International Journal of
Learning Management Systems

ISSN: 2785-9568 VOLUME 1, ISSUES 2, 2013 www.egyptfuture.org/ojs/

Psychometric properties of scores from an embedded and independently-administered short form of the Raven's Advanced Progressive Matrices

Ali Mohamed Ibrahim[1] and Ali Mahdi Kazem[2]

Sultan Qaboos University, Oman

Email: alimi@squ.edu.om[1],amkazem@squ.edu.om[2]

Received: 4 Nov. 2012 , Revised: 19 Dec. 2012 , Accepted: 18 Jan. 2013

Published online: 1 Jul. 2013

Abstract: The purpose of this study was to compare the psychometric properties of scores obtained from the Arthur and Day's short form of the Advanced Progressive Matrices (APM) when the items of the short form are administered within the full test (i.e., embedded short form), and when they are administered independently as a stand-alone test. The full test was administered to 433 students from two universities in the Sultanate of Oman, and the short form was completed by another sample of 179 students from one of the two universities from which the first sample was drawn. Results of the embedded short form were generally comparable to those reported by Arthur and Day (1994). But, the scores from the independently-administered short form had inadequate psychometric properties compared to those of the embedded short form and the full test. Hence, it was concluded that the short form under consideration is not suitable for predictive purposes due to the relatively low level of reliability of scores obtained from it when administered as a stand-alone test which results in a large standard error of measurement as compared to the full test.

Keywords: psychometric properties, short form of the Advanced Progressive Matrices, APM.

Introduction

In many situations short forms of standardized tests of cognitive abilities or other individual differences scales were developed in order to reduce administration time, and possibly to partly relieve examinees from exerting too much mental effort in the process of solving the highly demanding problems.

The Advanced Progressive Matrices Test (APM), is a well-known and widely-used nonverbal test of the "*g*" factor. It was developed by Raven (1965) to assess individual differences in observation, clear thinking and mental capacity. The test consists of 36 items which represent visual analogy problems. Each problem consists of a 3×3 matrix, in which the bottom right entry is missing and must be selected from a set of eight alternatives arranged below the matrix in the same page (Carpenter, Just, and Schell, 1990). The APM's manual specifies 40 minutes for the timed administration, and unlimited time for the untimed administration (Raven, Raven, and Court, 1998). Some researchers and practitioners who use the APM find this time limit somewhat prohibitive.

Hence, Arthur and Day (1994), in an attempt to reduce administration time of the test, selected 12 items from the 36 items of APM to form a short form of the test which they expected to represent the full-length test adequately, and hence, shorten administration time appreciably. In order to select the appropriate items to compose the short form, Arthur and Day adopted a number of decision rules. Since the APM items were designed to be progressively difficult such that there is nearly a monotonic increase in difficulty from the initial problems that have negligible error rates (i.e. very easy) to the last few problems that have error rates at or above 90% (Carpenter et al., 1990), Arthur and Day (1994) decided to select items so as to preserve this difficulty structure.

They first divided the test into 12 sections of three items each. Then of the three items of each section, the item with highest item-total correlation (i. e. discrimination) was selected. Most ties were reconciled by selecting the most difficult item. For the remaining ties the decision was to choose the item whose deletion resulted in the largest drop in Cronbach's *alpha* coefficient. Thus the 12 items selected items had the serial numbers 1, 4, 8, 11, 15, 18, 21, 23, 25, 30, 31, and 35 in the full version of the APM. Fifteen minutes were recommended for administering this short version of APM.

In a first study, the full test was administered to 202 university students (mean age = 21.40 years, *SD* =4.42). The Cronbach's alpha obtained for the short form was 0.72 (compared to 0.84 for the original APM). The correlation between the full APM and the short form was 0.90. But, the items of the short form were embedded in the full form. Hence, the authors emphasized that "there are several potential biases associated with evaluating the short form when embedded in the long (e.g. item dependency)" (p. 398). Therefore they decided to assess the short form's psychometric properties when considered separately from the full test. In other words, they attempted to assess the soundness of the short form by independently administering it as stand-alone test. Their second study accomplished this aim. The short form was completed by all 246 subjects comprising the sample of the second study (mean age = 21.81 years, *SD* = 5.35), while 161 of them completed the full test. The resultant alpha coefficient for the short form was reduced to 0.65 (n=246). Similarly, the correlation between the full APM and the short form was drastically reduced to 0.66 (n=161). In the third study, they added 215 subjects (with mean age = 19 years, *SD* = 1.22). All subjects completed the short form, and 111 of them were administered the test another time to assess the short form's test-retest reliability. The data for this third study were combined with second study data. An internal consistency of 0.69 was obtained (n = 461), and the test-retest for the short form was 0.75 (n=111).

Arthur and Day (1994), concluded that their short form is "a much more efficient tool for researchers and practitioners to use than the more time-consuming full test" (p. 402). But the authors cautioned that "more normative data are needed on the short form…since there is no normative data on this form as a stand-alone test. As such, incorrect inferences might result from using long-form norms to interpret short-form scores" (p. 402). Finally, the authors believe that the short form encourages other researchers to provide additional psychometric data.

Therefore, in a further study Arthur, Tubre, Paul and Sanchez-Ku (1999) sought to present additional psychometric and normative data of Arthur and Day's short form of APM. They attempted to accomplish this objective by obtaining already available data from multiple researchers who had used the short form in a number of studies which provided a larger sample than the one used by the original developers of the short form. Subsequently, Arthur et al. (1999) obtained psychometric properties (difficulty levels, internal consistency and test-retest reliability of test scores, convergent validity, factor structure, and percentile scores) from responses to the APM short form obtained from 11 researches in a large U.S. university (n=1506, 50.9% were females, 41.3% were males, and the rest did not indicate their sex). The mean age of respondents in the multiple sample was 19.38, *SD* = 2.29).

The results obtained by Arthur et al. (1999) confirmed those of Arthur and Day (1994). The researchers subsequently concluded that the short form of APM is a useful alterative for researchers who would like to use the full test but have limited administration time available to them. The researchers, realizing that "one of the primary advantages of the APM is its apparent low level of culture-loading" (Arthur et al., 1999, p. 360), recommended that future research should address racial differences on the APM short form. Hence, the current study is a contribution in this direction since it attempted to provide further evidence from a non-western culture.

Bors and Stokes (1998) also attempted to construct another short form of 12 items from APM using a somewhat different procedure for selecting items than that adopted by Arthur and Day (1994). Their purpose was to compare the psychometric properties of the scores obtained from their form to those which could be obtained using Arthur and Day's short form. After reviewing studies which investigated the factor structure of APM, Bors and Stokes (1998) concluded that a single factor underlies performance on the APM, and therefore, the researchers conclude that there was no need for them to sample items from different subsets of the test in order to construct a short form as Arthur and Day (1994) did in constructing their short form. They first rank-ordered the items by their item total correlations (i.e. discriminations), and they checked the inter-item correlations to remove any redundancies. An item was not selected if it had a high item-total correlation but also substantially correlated with another item with a similar high item-total correlation, thus adding little to the predictive power of the test. Five of their 12 selected items were in common with Arthur and Day's items. The strategies for selecting the items for these two short forms were somewhat different. But, in test development (or construction) strategies both characteristics of items (difficulty and discrimination levels) should be taken into consideration when selecting appropriate items from the targeted pool of items. Similarly, it is expected that the same strategies apply to the development of a shorter test from a longer one. But, Stanton, Sinar, Blazer, & Smith (2002) stated that "important differences exist between the two activities. For an example, an existing scale have known validity relations that must be preserved in its shortened version" (p. 169). Therefore, they suggested three criteria for developing a short form from a long scale. These criteria were: Internal, external, and judgmental. Internal qualities of items refer to their properties that can be assessed with reference to other items or to the sum of scores (such as corrected item-total correlations). External qualities of items refer to their relations with external constructs. Judgmental qualities are subjective (e.g.

item's clarity of expression, its relevance to a particular group of respondents, its 'face' validity, etc). Most of these judgmental qualities are applicable in nonverbal tests such as the APM.

Internal consistency of the scores of the short form of Bors and Stokes (1998) was 0.73, and the correlation of scores with the full-length test scores was 0.92. It should be noted, however that this high correlation is inflated since the short item score is part of the full test total score. In other words, the short form in this case was "embedded". Arthur and Day reported a comparable coefficient of 0.90. When Bors and Stokes applied Arthur and Day's short form to their data, the resulting correlation with the full-length test was slightly lower (0.89 compared to 0.92). Unlike Arthur and Day, Bors and Stotes did not administer their short form of APM as an independent scale.

Therefore, Vigneau and Bors (2005) suggested that an analysis of Bors and Stokes (1998) short form of the APM offered an opportunity to test the effect administering some items of a test as an independent scale. They referred to this effect as being similar to the "context effect" which was reported by Kubinger et al. Hence, Vigneau and Bors raised an important question: "Do these short form items behave any differently when they are administered separately than they behave when they are administered within the context of the entire 36-item scale?"(p. 117). Kubinger et al. (cited in Vigneau and Bors, 2005) found that when 17 items which were identified as Rasch homogeneous in the Standard Progressive Matrices Test (which consists of 60 items) no longer were found to be homogeneous when administered as an independent 17 item scale. Vigneau and Bors (2005) concluded from their study to assess the dimensionality of the APM that items can behave differently depending upon which other items are also administered. In other words, items presented in one context may measure a somewhat different ability when presented in another context.

Hamel and Schmittmann (2006) used a 20-minute timed version of APM and compared it to the untimed APM among first year psychology students. They found a correlation of 0.75 between the 20-minute timed version scores and scores on the whole APM, but no reliability coefficients were reported. Nevertheless, these authors concluded that administering APM with a time limit of 20 minutes is better than administering 12 items only of the test, because this selection represents a different task for the examinees. In other words, these short versions might differ from the original APM in a qualitative way. Moreover, Hamel and Schmittmann added that the validity of the APM as a power test bears heavily on learning from experience gained by the examinee during the process of completing the test. Although these comments are sound, learning from the test requires that the average examinee is allowed sufficient time to try all (or at least most) of the items. As far as APM is concerned, 20 minutes are not adequate for most students to accomplish this objective. Therefore, it seems that the assumption of learning from the test within 20 minutes time limit is not tenable.

Furthermore, if scores from a short form of the APM or from administering the test with 20 minutes time limit (as suggested by Hamel and Schmittmann, 2006) are correlated (and in fact they do) with scores from completed test, does this necessarily mean that the short form or the 20-minute version of the test, can be used in lieu of the full-length test to measure individual or group performance on the test? Any observed correlation, represents the relationship of a part to a whole. In other words, does any of these two options (i. e. a short form or 20 minute time limit) give accurate estimation (or prediction) of the level of the trait measured among individuals and groups? In a related answer to this question, Arthur et al. (1999) cautioned against the use of the short form of APM or any other short form in intensive clinical assessments, and stated explicitly that "…it should be noted that where the length of administration is not a major concern, using all APM items will provide a better assessment"(p. 360).

Another related question regarding accuracy of prediction of short forms of tests also arises here. Does a short form of a test measure the intended hypothetical “true score” of an individual as the full-length test does? Neither Arthur and Day (1994) nor Bors and Stokes (1998) who constructed short forms of APM attempted to answer this question. However, Bors and Stokes (1998), explicitly stated that "Given that the predictive utility of a test is limited by score reliability, and that the reliability of test scores is somewhat related to its length, the challenge is to produce a considerably shorter version of the APM without a substantial reduction in reliability"(p.385). Furthermore, we recall that in classical test theory, the accuracy of prediction of "true score" from observed score is achieved by constructing a confidence interval or band for the true score using the standard error of measurement, which is function of the standard deviation of the scores and the reliability coefficient (Nitko and Brookhart, 2007). Hence, it is important to compare the standard error of measurement calculated from the full-length test and that calculated from the short form or the 20-minute timed version which was suggested by Hamel and Schmittmann (2006).

The purpose of the present cross-cultural investigation was to compare Arthur and Day’s short form of APM (first as an embedded test and then as independently administered test) and the full-length test with regard to: item difficulties, item discrimination indices, reliability, convergent validity, standard error of measurement, and levels of performance of examinees in the two conditions.

Method

1- Sample

Participants comprised two sub-samples. The first sub-sample was drawn from two universities (one governmental and the other private) in the Sultanate of Oman. (n=433, 152 males, 281 females, Mean age (21.17, SD = 1.45). Intact classes were randomly selected from various colleges (158 from scientific colleges, 275 from humanities colleges).

The other sub-sample (n=179, 105 males, 74 females, Mean age= 21.77, SD = 1.46) was drawn from the governmental university only in order to administer the short form of APM as stand-alone test.

2- Procedure

Members of the first sample were administered the full-length APM during the Fall semester of the academic year 2010-2011 as a preliminary standardization of the APM among university students in the Sultanate of Oman. On the other hand, students of the second sample were administered the Arthur and Day's short form of APM the during the Summer session of 2011 and the Fall semester of the academic year 2011-2012. From the data of the first sample, items of the short form of Arthur and Day were analyzed separately as an embedded short form in order to compare the psychometric properties of the responses to these items with those obtained from the test of full-length and from the independently-administered short form USA data provided by Arthur and Day (1994).

Results and discussion

1- The embedded short form:

1-1- Difficulty levels of the items:

Table (1) presents the item difficulties (proportions of correct answers) from the present sample and those from Arthur and Day’s sample. Since the short form in the present study is embedded, its item difficulties are identical with item difficulties of the full test.

Table 1: Item difficulty indices for embedded short form of present data and Arthur and Day’s (1994) data.

APM item #	Present study	Arthur & Day
1	85	89
4	73	78
8	75	87
11	75	84
15	67	77
18	43	67
21	58	58
23	51	44
25	42	64
30	32	44
31	23	38
35	19	31

The Pearson's correlation coefficient between item difficulties of the embedded short form of APM reported by Arthur and Day (1994) and those obtained from the sample of the present study was 0.93. This high level of association indicates that the progressive difficulty of the items was preserved. But, it is observed that most items in Arthur and Day’s data were easier (mean difficulty level was 63) than in the current data (mean difficulty level was 47). This may be attributed, partly, to differences between samples since the sample of the present study was more heterogeneous than Arthur and Day’s sample.

1-2- Item discrimination indices:

Table 2 displays item discriminations (i.e. corrected item-total correlation) for the short form of the present data and those reported by Arthur and Day.

Table 2: Item discrimination indices for embedded short form of present data and Arthur and Day's (1994) data.

APM item #	Present data	Arthur & Day
1	.41	.37
4	.51	.34
8	.46	.40
11	.50	.49
15	.39	.36
18	.38	.36
21	.36	.54
23	.42	.42
25	.36	.37
30	.29	.42
31	.22	.36
35	.21	.33

The correlation between these two sets of discrimination indices is only 0.25 (not significant, p = 0.44). This weak correlation indicates that the order of the items with regard to discrimination power in the embedded short form in the current study is not similar to that obtained by Arthur and Day for their embedded short form. However, the mean discrimination index of the items of the embedded short form in the two studies were 0.38 and 0.40, respectively. These two means are almost the same indicating that the average discrimination power of the items remained intact in the embedded short form of the current study.

1-3- Reliability of the scores:

The reliability coefficient of student scores in the embedded short form of the present study (0.72) was equal to that obtained by Arthur and Day. The reduction in alpha coefficient of scores of the short form in both cases (compared to that of the total test) is due to the reduction of number of items. If the Spearman-Brown prophecy formula is used for a test three times as long, we get an estimated alpha coefficient= 0.90 which is comparable to the actual alpha obtained from the scores of the full test in the present study (which was 0.87). Hence, shortening the test (as expected) will reduce the reliability of the scores.

1-4- Convergent validity:

The only external criterion available to the present study was self-reported GPA. The correlation of GPA with scores of the full-length test was.152 (*p*=.005, *n*=336), while for the embedded short test, this correlation was.141 (*p*=.01, *n*=336). These two correlations were in fact equal.

1-5- Average performance:

In order to compare between performance of students in the full APM test and its short form, the short form for each student was multiplied by three in order to equate the full test and the short form total scores. The results revealed that the mean performance in the short form (mean= 19.29, SD = 8.30) and the mean performance in the full test (mean= 18.59, SD = 6.75) were almost identical (both approximately = 19 out of 36).

1-6- Standard error of measurment:

For the full test, the standard error of measurement was 2.43, while for the embedded short from it was 4.39. This difference indicates (as expected) that error of the measurement associated with short form scores are much larger than those associated with the scores obtained from the full test. Hence, estimated true scores of the embedded short form are less accurate (i.e. wider confidence intervals for observed scores) than those of the full test.

2- The independently-administered short form:

2-1- Difficulty levels of the items:

The difficulty levels of the items of the short form of APM when administered as a stand-alone test were compared with their difficulty levels when administered within the full test in order to see whether these difficulty levels are preserved in both cases. Table 3 displays these difficulty levels.

Table 3: Item difficulty levels for independently-administered (stand-alone) short form and full test of the present study.

Original APM item #	Stand-alone short form (n=179)	Full test (n=433)
1	96	85
4	55	73
8	85	75
11	59	75
15	73	67
18	43	43
21	57	58
23	35	51
25	47	42
30	28	32
31	42	23
35	13	19

The Pearson's correlation coefficient between item difficulties of the independently-administered short form of APM and the full test for the current study was (0.87). This high level of association indicates that the progressive difficulty of the items was preserved. Moreover, it is observed that the two means of difficulty levels in both cases were equal (53 and 54). Therefore, it is concluded that the item difficulty levels of the short form is not affected by administering them as stand-alone test or within the full test.

2-2- Item discrimination indices:

The discrimination indices of the items of the short form of APM when administered as a stand-alone test were compared with their discrimination indices when administered within the full test in order to see whether these discrimination indices are preserved in both cases. Table 3 displays these discrimination indices.

Table 4: Item discrimination indices of full-length test and stand-alone (independently-administered) short form of present data.

Original APM item #	Full test	Stand-alone Short form
1	.50	.17
4	.54	.19
8	.52	.27
11	.61	.19
15	.44	.28
18	.40	.36
21	.41	.27
23	.45	.17
25	.41	.24
30	.27	.09
31	.25	.13
35	.20	.26

Table 4 reveals that discrimination indices for the same items when they were independently-administered as a stand-alone test were clearly lower than when they are administered within the full test. The correlation between these two sets of discrimination indices was only 0.13. This low and insignificant correlation indicates that the order of the items with regard to discrimination power in the stand-alone short was not similar to that of the full test. Moreover, the mean discrimination index of the items of the stand-alone short form was 0.22 while this mean for the same items within the full test was 0.40. These two means were different which indicated that the average discrimination power of the item did *not* remain intact in the stand-alone short form where the same items were less discriminating when administered as stand-alone test. Arthur and Day (1994) did not report discrimination indices of their data for independently-administered short form.

2-3- Reliability of the scores:

Cronbach's alpha for the independently-administered short form was found to be 0.54 (n=179), an inadequate level of internal consistency. This could be attributed to the appreciable shortening of the test. Hence, shortening the test (as expected) reduces the reliability of scores. Similarly, the test-retest coefficient of reliability after two weeks was inadequate (0.51, n=55).

2-4- Convergent validity:

The correlation of the total short form scores and self-reported GPA was 0.24 (*n*=137, *p*=.005). We recall this level of convergent validity of the independently-administered short form is slightly greater than that of the full form.

2-5- Average performance:

The mean of the total scores of the independently-administered short form (after equating with the full test score by multiplying total scores by 3) was 19.00 (SD = 6.52). This value was identical to the value obtained from the full test which was administered to the first sub-sample of the present study.

2-6- Standard error of measurement:

The standard error of measurement of the total scores of the independently-administered short form was 4.42. Thus, the result is similar to that of the embedded short form scores. (i.e. errors of measurement were much larger than those associated with the scores obtained from the full test). This is attributed to the lower reliability coefficient of the short form. Hence, estimated true scores of the independently-administered short form are less accurate since wider confidence intervals for "true" scores are obtained compared to those obtained from observed scores the full test.

Conclusion

The current study attempted to provide further evidence from a different cultural setting, regarding the feasibility of developing a short form of the APM which could result in psychometrically sound scores as the full test does. Specifically, the purpose of this study was to compare the psychometric properties of scores obtained from the Arthur and Day's (1994) short form of the Advanced Progressive Matrices Test (APM) when the items of the short form are administered within the full test (i.e., embedded short form) and when they are administered independently as a stand-alone test (as they should be). The full APM was administered to 433 students from two universities in the Sultanate of Oman, and the short form was completed by another comparable sample of 179 students from one of the two universities from which the first sample was drawn. Results of the embedded short form were comparable to those reported by Arthur and Day (1994). The 12 items of the short form (when administered within the full test) were progressively difficult (as they should be) and the internal consistency reliability coefficient was 0.72. Furthermore, (as expected) a high correlation coefficient was found between scores of the embedded short form and the full APM scores. But, the scores from the independently-administered short form had inadequate psychometric properties (e.g. reliability coefficient = 0.54) compared to those of the embedded short form and the full test. Hence, it was concluded that the short form under consideration is not suitable for predictive purposes due to the relatively low level of reliability of scores obtained from it which results in a large standard error of measurement as compared to the full test. It seems reasonable to conclude that, the APM short form which was developed by Arthur and Day (1994) does not provide accurate estimation of an individual’s level of general intelligence as measured by the APM.

References

Arthur, W. Jr., and Day, D.V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, *54,* 394-403..

Arthur, W. Jr., 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 Progressive Matrices Test. *Journal of Psychoeducational Assessment*, *17,* 354-361.

Bors, D.A., & Stokes, T. (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.

Carpenter, P.A., Just, M.A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven's Progressive Matrices Test. *Psychological Review*, 97, 404-431.

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

Nitko, A. J., & Brookhart, S. (2007). *Educational assessment of students*. Upper Saddle River, N.J.: Pearson Education, Inc.

Raven, J. C. (1965). *Advanced Progressive Matrices, Sets I and II*. London: H. K. Lewis.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven Manual section 4: Advanced Progressive Matrices*. San Antonio: Harcourt.

Stanton, J. M., Sinar, E. F., Blazer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*, 167-194.

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