Items in Context: Assessing the Dimensionality of Raven's Advanced Progressive Matrices

François Vigneau and Douglas A. Bors

Educational and Psychological Measurement 2005 65: 109
DOI: 10.1177/0013164404267286

ITEMS IN CONTEXT: ASSESSING THE DIMENSIONALITY OF RAVEN'S ADVANCED PROGRESSIVE MATRICES

FRANÇOIS VIGNEAU
Université de Moncton

DOUGLAS A. BORS
University of Toronto at Scarborough

The problem of dimensionality with respect to Raven's Advanced Progressive Matrices (APM) specifically and, more generally, *g* or fluid intelligence, has been a long-standing issue. The present article reports two studies examining the dimensionality of both the original Set II of the APM ($n = 506$) and a short form ($n = 644$), using principal component analysis and Rasch analysis. Although the results from the principal component analysis were equivocal, results from the Rasch analyses more strongly suggested that both forms of the test are best described as being multidimensional. Furthermore, comparison of items common to both forms indicated a context effect, thus making adaptive testing versions of this test difficult.

Keywords: *dimensionality; item response theory; Rasch model; Raven's Progressive Matrices; intelligence*

The quest for the nature of general intelligence (*g*), as psychometrically defined, has led researchers down a long and winding road. At times, the search has been abandoned, only to be resumed at a latter time. The search as it pertains to Raven's Matrices has been no different. Beginning with Spearman (1939, 1946), Raven's Standard Progressive Matrices (SPM) and Raven's Advanced Progressive Matrices (APM) have been regarded by

This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC). Parts of the results were first presented at the XVth Journées de Psychologie Différentielle, Rouen, France, September 2002. Correspondence regarding this article should be addressed to F. Vigneau, Ecole de psychologie, Université de Moncton, Moncton, New Brunswick, Canada, E1A 3E9; e-mail: vigneaf@umoncton.ca.

Educational and Psychological Measurement, Vol. 65 No. 1, February 2005 109-123
DOI: 10.1177/0013164404267286

many as appropriate measures of *g* (Burke, 1958; Jensen, 1987). For almost as long, however, this contention of a unidimensional nature of the ability measured by the SPM and the APM has been debated, with no real signs of an end in sight (Jastak, 1949; Van der Ven & Ellis, 2000).

Being confident that the source of individual differences in performance on the SPM and the APM is unidimensional has consequences for both theoretical and applied work in psychology. An illustration of the impact of the adoption of a unidimensional model can be found in the literature on the relationship between *g*, typically estimated by either the SPM or the APM, and mental speed (cf. Neubauer, 1995; Vernon, 1987). In this area, because the matrices have been assumed to be unidimensional (*g*), any explanatory model assuming that one or more additional dimensions, common to both reaction time tasks and Raven's Matrices, could constitute the basis for the observed correlations has been implicitly excluded. The search has been for a single determinant with which to label *g*, be it working memory capacity, attention, or some other construct deemed to be elementary. Furthermore, what also appears to have been forgotten is that *g* itself, as defined psychometrically, might well be multiple correlated determinants just as easily as it could be a single determinant. The statistical behavior of these two scenarios could be identical.

One of the difficulties of trying to solve the problem of the dimensionality of the matrices is that the Raven items are dichotomously scored and range across a wide spectrum of difficulty levels (Carroll & Maxwell, 1979). Distributional normality of the variables (responses to each item) is an assumption of factor analysis, which has been the standard analytic technique used to evaluate test dimensionality. The fact that some items are relatively very easy, and others are comparatively difficult, results in the distributions of many Raven items being substantially skewed, some negatively and some positively. This produces an inconsistent pattern of attenuation across the interitem correlation matrix. This unavoidable property of Raven items has been held responsible for at least some of the ambiguity arising out of the factor analytic studies where additional dimensions are at least suggested by the factor patterns. For example, these additional dimensions are often understood to be statistical artifacts or, similarly, difficulty factors produced by the distributional properties of the items, and not additional relevant cognitive factors that influence performance on the items (Rost & Gebert, 1980).

Attempts to circumvent this problem of skewed item distributions have included various corrections to the correlations that are used as input for the analyses: phi coefficients, corrected phi coefficients, and tetrachoric correlations (see, for example, Rummel, 1970). The results of such procedures have been mixed. Rost and Gebert (1980), using phi coefficients, found three factors, which they attributed to a clustering based on item difficulties. When Rost and Gebert used corrected phi coefficients ($\text{phi}/\text{phi}_{\text{max}}$) derived from the

same data set they had used in the previous analysis, they found that a single factor offered the best fit. On the other hand, Dillon, Pohlmann and Lohman (1981), also using corrected phi coefficients, found two fairly distinct factors that were independent of difficulty level. Bors and Stokes (1998), using tetrachoric correlation coefficients, however, failed to replicate Dillon et al.'s pattern of factor coefficients and found the typical clustering pattern of easier and difficult items. Thus, even disregarding the technical problems often associated with attempting to use such coefficients or corrected correlations as input with standard statistical software packages, such corrections have not produced substantially clearer reliable factor solutions and interpretations. Confirmatory factor analysis is also sensitive to distributions and has not taken us beyond the dilemmas afflicting exploratory factor analysis (Bors & Stokes, 1998).

Recently, item response theory approaches to test construction, and Rasch model approaches particularly, have been used to examine the dimensionality of Raven's Matrices (Kubinger, Formann, & Farkas, 1991; Van der Ven & Ellis, 2000). Rasch model approaches have the advantage of being insensitive to the shape of item distributions, particularly with respect to dichotomous variables. In fact, the Rasch model was designed specifically for examining psychometric instruments containing just such items. Although Rasch analysis was not developed for the explicit purpose of investigating the dimensionality of psychometric instruments, given that unidimensionality is an essential assumption of the Rasch approach to item analysis and test construction, proponents of the approach have developed means of testing this postulate.

Kubinger, Formann, and Farkas (1991), in a study of schoolchildren between the ages of 8 and 14 (n = 527), examined the SPM and found that, when taken as a whole, the 60 items were not found to be Rasch homogeneous: That is, the assumption of unidimensionality could not be maintained. Using item-diagnostic statistics, however, they were able to identify 17 items that were Rasch homogeneous. When another comparable sample of participants was administered only those 17 items, the items were no longer found to be Rasch homogeneous, however. This suggested that performance on the individual items and the relations among the items were not independent of the context in which the items were administered. Van der Ven and Ellis (2000), however, in a study involving 901 students aged 12 to 15, analyzed the five subsets of the SPM separately. Three of the five subsets (A, C, and D) were found to be Rasch homogeneous. The other two subsets (B and E) were found not to be Rasch homogeneous.

Using both factor analytical and Rasch techniques, the present article presents analyses of the dimensionality of the APM based on two large data sets. Study 1 examines the performance of university students on all 36 items

of the APM. Study 2 examines a subset of these items, as constructed and administered as a short form of the test.

Study 1

Method

Participants

The timed version of the APM was administered to 506 first-year students (326 women, 180 men) at the University of Toronto at Scarborough who were given extra credit in an introductory psychology course for their participation. They ranged from 17 to 30 years in age (M = 19.96, SD = 1.83).

Procedure

Participants completed both Set I and Set II of the APM. The standard instructions were read aloud by the experimenter. The standard timings of 5 minutes for Set I and 40 minutes for Set II were allowed. Only the results from Set II are reported here.

Results

Given that the descriptive statistics and the details from a principal component analysis of these data have been conveyed elsewhere (Bors & Stokes, 1998), only a relevant précis is reported here. All 36 items were positively correlated with all other items, and the test as a whole yielded scores with strong internal consistency (alpha = .84). The corrected item-total correlations ranged from .07 to .52 (M = .32, SD = .08).

Factor Analyses

Table 1 shows the percentage correct and the corresponding skewness of the items. Accuracy rates ranged from 4% to 95% correct (M = 62%, SD = 28%), and skewness ranged from –4.08 to 4.61. Such a range in skewness would certainly have a considerable effect on the correlations between the items. Where the tetrachoric correlations between the items were previously used as the input for the analysis, the uncorrected Pearson product moment correlations were used here for the sake of comparison. Although there were negligible differences in some of the details between the two analyses, such as the factor coefficients being slightly greater when the tetrachoric correlations were used as the input, the overall factor structure and pattern of coefficients were found to be virtually indistinguishable.

Having found little differences between the two analyses, we then repeated the procedure once again, this time using the corrected phi correlations as

Table 1
Percentage Correct and Skewness by Item

APM Item	% Correct	Skew	APM Item	% Correct	Skew
1	93	–3.29	19	73	–1.06
2	95	–4.08	20	66	–0.67
3	92	–3.23	21	67	–0.74
4	87	–2.15	22	50	–0.02
5	88	–2.37	23	60	–0.42
6	93	–3.53	24	40	0.39
7	91	–2.81	25	50	0.01
8	88	–2.28	26	44	0.22
9	90	–2.66	27	36	0.57
10	86	–2.05	28	32	0.79
11	90	–2.66	29	22	1.39
12	85	–1.94	30	35	0.64
13	70	–0.88	31	31	0.83
14	84	–1.84	32	21	1.40
15	79	–1.42	33	21	1.44
16	77	–1.29	34	18	1.66
17	74	–1.10	35	18	1.67
18	67	–0.74	36	04	4.61

Note. APM = Advanced Progressive Matrices.

input. The overall pattern of factor coefficients was no different from the previous two analyses. This being the case, only the details from the first two analyses are presented. This lack of difference in the factor patterns is an important finding, given the time and effort that has been expended on correcting the correlations used for factor analytic studies of such tests.

The principal components analysis (Pearson product-moment correlations) of the 36 Set II items produced 11 components with eigenvalues greater than one, with only 2 components having eigenvalues greater than two. All 36 items loaded positively on the first component. In the previously reported analysis based on the tetrachoric correlations, there were 12 factors with eigenvalues greater than one and 3 with values greater than two. In the present study, the first component (eigenvalue = 6.61) accounted for 18% of the variance; in the previously reported analysis, the first component accounted for 20%. In both analyses, the second component accounted for less than 6% of the variance. Qualitatively, both analyses offer us the picture of a weak single-factor model. Furthermore, the two patterns of factor coefficients following a (VARIMAX) orthogonal rotation of the first two components were very similar. Most of the first 28 items loaded the best, but only moderately or weakly, on one factor, whereas the last 7 items loaded best, but weakly, on a second factor. Thus, other than a possible clustering on the basis of item difficulty, a single-factor solution appears to be the most parsimonious interpreta-

tion in both the present and the previous analyses of that data set. Given the rather weak factor coefficients and corresponding small portions of the variance that this single factor explains, the acceptance of a one-factor solution is more of a default acquiescence than a compelling conclusion.

As mentioned above, Rasch analyses are designed to be free of the problems associated with dichotomous variables, and they make no assumption concerning the shapes of the distributions of the test items. It is thus possible to test for a single predominant factor by examining those assumptions necessary for a Rasch analysis. One such assumption is that the items differ only in terms of their level of difficulty. Thus, if the rotated clusters that were found in the principal component analyses were mere artifacts of differences in the items' difficulties, then finding that the APM items differ only in terms of their difficulty would provide support for the one-factor solution.

Rasch Analysis

All Rasch analyses reported in this article were performed using the conditional maximum likelihood (CML) estimation procedure as implemented in the Rasch Scaling Program (RSP; Glas & Ellis, 1994). Logits were used as initial estimates. The Q1 and the Q2 statistics are the most appropriate statistics for testing dimensionality (Glas & Ellis, 1993). As suggested by some authors (Glas, 1988; Van den Wollenberg, 1982; Van der Ven & Ellis, 2000), a distinction between two assumptions of the Rasch model might be needed. The distinction is between what has been called first-order and second-order realizations or, put differently, between unidimensionality in terms of difficulty and unidimensionality in terms of ability. The distinction is important because these two aspects are not coextensive. Whereas unidimensionality of ability implies unidimentionality of difficulty, the reverse is not true. There can be a monotone increase of item response functions without unidimensionality in terms of ability. The Q1 statistic tests the degree to which the item response functions are parallel (unidimensional difficulty). With the *Y*-axis being the probability of a correct response and the *X*-axis being the total score (ability level), if the items differed solely with respect to difficulty, that is, if they all reflected a single predominant factor, then the item response functions would be roughly parallel. A statistically significant Q1 indicates that the functions deviate significantly from this parallel structure. In such cases, one usually finds considerable crossing of the functions. Q2, as a statistic of dimensionality, is based on the assumption that if there is only one factor (difficulty) upon which the items are scaled, and if participants' abilities on that dimension are fixed, then the resulting correlations between the items' residuals should be zero, once the variance accounted for by the underlying factor has been partialled out (unidimensional ability). A statistically significant Q2 indicates that there remain substantial correlations between at least some of the items and that unidimensionality cannot be safely assumed.

As pointed out by Glas and Ellis (1993), Q1 and Q2 are highly sensitive to the participant sample size. With large numbers of participants, such test statistics can have small probability (p) values even though the violation of the model may be relatively small. For this reason, they recommend testing the statistical significance of the Rasch model with very small p levels, such as .001. All tests of Q1 and Q2 in the present article will use this prescribed level when interpreting the results.

When all 36 items from APM Set II were included in the analysis, the scale was not found to be homogeneous. That is, it appeared that at least some of the items differed on a dimension other than difficulty, which was illustrated by the fact that functions varied across performance-level groups. This was indicated by the intersecting item response functions (Q1 = 325.69, df = 175, $p < .001$). Multidimensionality was further suggested by substantial residual interitem correlations, reflected in a high Q2 (12,366.45, df = 3,564, $p < .001$).

Furthermore, it was not just a matter of a few poorly fitting items. Manipulations based on various indicators (item Q statistics, U statistic) indicated that roughly half of the items should be dropped from the APM to attain non–statistically significant levels of Q1 and Q2. Interestingly, most of the remaining items are positioned in the middle portion of the scale. When Items 4, 8, 11, 13, 15 to 18, 22 to 29, 31, 32, and 34 are retained (n = 503), Q1 = 86.71 (df = 72, p = .14) and Q2 = 843.37 (df = 760, p = .13). Additionally, we compared the item probability-of-success functions for men and women. APM items do not seem to be biased with respect to gender (Q1 = 40.2192, df = 35, p = .25).

As shown above, selective subsets of APM items can be found to be Rasch homogeneous. More interesting, but perhaps not surprisingly in light of the principal component analyses, after accommodating for a few deviant items, we found that APM items grouped according to their position in the test (1-12, 13-24, and 25-36) were independently homogeneous, both in terms of monotone increasing parallel item response functions (Q1) and in terms of unidimensional ability (Q2) (see Table 2).

Thus, the clustering discovered in the principal component analyses may not be a simple artifact of increasing difficulty. The present Rasch analysis illustrates, beyond the issues of difficulty and skewness, that the three sections of the APM (beginning, middle, and end) can be regarded as rather independent subtests. If this is the case, then the increasing difficulty, across the 36 APM items, may not represent merely a quantitative change, but it also may reflect a qualitative change. That is, it seems that as participants proceed through the matrices, they arrive at points where something different from what they were doing to solve the matrices is then required for them to be successful, not just more of the same.

Table 2
Q1 and Q2 by 12 Item Subsets

	Statistic	*df*	*p* Value
Items 1 to 12			
Q1	11.5027	11	.4022
Q2	56.7704	108	1.0000
Items 13 to 24			
Q1	45.0860	36	.1425
Q2	129.1801	175	.9962
Items 25 to 36			
Q1	38.9429	27	.0641
Q2	173.4807	140	.0287

Finally, a nonlinear factor analysis of the 36 items was conducted. This was carried out with the NOHARM software package (Fraser, 1993; Fraser & McDonald, 1988). The resulting pattern of factor coefficients was similar to those reported from the standard principal component analyses. All items loaded positively on the first unrotated factor with coefficients ranging from .21 to .73. The rotated factor coefficients pattern (VARIMAX) of two dimensions was also similar to the rotated coefficients pattern that resulted from the principal components analyses. The first 28 items loaded moderately to strongly on the first factor. Items 29 to 36 loaded moderately to strongly on the second factor. The only observable difference between the two analyses was the degree of factor definition: the relative strength of the coefficients of the items on the two factors. In comparison to their counterparts in the principal components analyses, in the nonlinear analysis, the stronger coefficients of most variables were stronger and their weaker coefficients were weaker. Where the Rasch analyses suggested two qualitative shifts in the matrices (three factors), the nonlinear factor analysis indicates that there may be a single qualitative change (two factors).

Study 2

For purposes of research with an undergraduate university population, Bors and Stokes (1998) developed a short form of Set II of the APM. The 12 items chosen from the original 36 (Items 3, 10, 12, 15, 16, 18, 21, 22, 28, 30, 31, and 34) were selected on the basis of their correlation with total score and their relative independence from other items. The first of these criteria resulted in the elimination of a large number of the easiest and most highly skewed items. An exploration of this short form of the APM affords two relevant opportunities with respect to our question of dimensionality. First, an examination of this short form will allow for a test of the impact of skewness

on the factor structure. Will a principal component analysis of the short form reveal a more homogeneous instrument? This would be reflected in a stronger first unrotated component and weaker rotated clusterings. Should this be the case, it would indicate the presence of a more unidimensional test embedded within a larger set of items, at least when the items are administered to an undergraduate population. It could be that the subscores for this more unidimensional subset produce the correlations between the APM and simple cognitive and noncognitive tasks.

Second, Kubinger et al. (1991) reported a significant context effect, in which 17 items identified as Rasch homogeneous in the 60-item SPM no longer were found to be homogeneous when administered as an independent 17-item scale. An analysis of the Bors and Stokes (1998) short form of the APM offered a similar opportunity to test the effect of context. Do these short-form items behave any differently when they are administered separately than they behaved when they were administered within the context of the entire 36-item scale?

Method

Participants

The short-form version of the APM (Bors & Stokes, 1998) was administered to 644 first-year students (418 women, 226 men) at the University of Toronto at Scarborough who were given extra credit in an introductory psychology course for their participation. Participants ranged from 17 to 51 years in age (M = 19.85, SD = 3.24). Reflecting the cultural diversity of the student population at the University of Toronto at Scarborough, 429 participants reported English as their first language, whereas 215 reported one of 38 languages other than English as their first language.

Procedure

For instructional purposes, as developed in Bors and Stokes (1998), participants completed the first 2 items from Set I. They were then administered the 12-item short form. The standard instructions were read aloud by the experimenter. Participants were allotted 15 minutes to complete the test items.

Results

Scores on the short form ranged from 0 to 12 (M = 7.16, SD = 2.23). There was a small but statistically significant difference in the performance of women (M = 7.02, SD = 2.32) and men (M = 7.42, SD = 2.34), $F(1,642)$ = 4.40, MSE = 5.417, $p < .05$.

Table 3 shows the percentage correct and the skewness for each of the 12 items. Item difficulties, as expressed as the percentage of participants answering the item correctly, ranged from 20% to 92% (M = 60%, SD = 25%). As would be expected, the range of skewness was considerably less than that found in the 36-item version and reported in Study 1. The interitem correlations ranged from .02 to .34, with a mean of .14 (SD = .07). The test was found to yield scores with modest internal consistency (alpha = .65). The corrected item-total correlations ranged from .20 to .43 (M = .30, SD = .07).

Factor Analysis

Using the Pearson correlation matrix as input, three components with eigenvalues greater than one were derived: 2.588 (21.6% of the variance), 1.18 (9.83% of the variance), and 1.11 (9.26% of the variance). All items loaded positively on the first unrotated component. The coefficients were all moderate, ranging from .33 to .64. As can be seen from Table 4, with the exception of short-form Item 6, the (VARIMAX) rotated component matrix for a three-factor solution illustrates a pattern of beginning-items, middle-items, and end-item clustering, with the strongest coefficients being moderate. Again, as was reported in Study 1, save for a clustering related to increasing levels of difficulty, a one-factor solution would likely be the most economical interpretation. Reducing the range of skewness, however, failed to produce factor analytic results more supporting of a single-factor solution.

Rasch Analysis

For the purposes of these analyses, it was necessary to remove 31 participants because of their perfect or null performance. As with the 36-item test, the 12-item short version did not appear unidimensional; although here the results were not significant in term of the difficulty functions, they were with respect to the correlations between the residuals: Q1(df = 33) = 53.5692, p = .0132; Q2 (df = 216) = 558.8469, $p < .001$.

Similar to the situation found in Study 1 using all 36 APM items, groups of items on the 12-item short-form test, defined in terms of their position in the scale, were found to be Rasch homogeneous (see Table 5).

The Context Effect: Conjoint Rasch Analysis of Both Test Forms

As described above, Kubinger et al. (1991) reported a significant context effect for 17 SPM items. A conjoint analysis of the 12 items of the short form in both samples (Study 1 data, n = 488; Study 2 data, n = 631) produced statistically significant values of Q1 (140.99, df = 77, $p < .001$) and Q2 (844.94, df = 432, $p < .001$). (For this analysis, 18 participants were dropped from the Study 1 data [original n = 506] sample because of perfect or null total scores.) These results are not surprising, given that the data from the 12-item short form alone was not found to be unidimensional.

Table 3
Percentage Correct and Skewness by Item: Advanced Progressive Matrices (APM) Short Form

APM Short-Form Item Number	Original Item Number	% Correct	Skew
1	3	92	–3.05
2	10	86	–2.14
3	12	86	–2.10
4	15	73	–1.04
5	16	76	–1.19
6	18	72	–0.98
7	21	64	–0.58
8	22	50	0.00
9	28	30	0.88
10	30	37	0.52
11	31	30	0.86
12	34	20	1.49

Table 4
VARIMAX Rotated Principal Component Matrix: Advanced Progressive Matrices (APM) Short Form

		Component		
APM Short-Form Item Number	APM Original Item Number	1	2	3
1	3	.50	.07	–.06
2	10	.67	.09	.08
3	12	.55	.41	.09
4	15	–.08	.69	.06
5	16	.37	.49	.02
6	18	.63	.01	.24
7	21	.18	.61	–.01
8	22	.17	.56	.15
9	28	–.11	.33	.47
10	30	.21	.01	.54
11	31	.22	.07	.61
12	34	–.11	.16	.69

Once again, these results were not caused by a few misfitting items only; a sizable portion of the scale must be dropped to reach non–statistically significant values of Q1 and Q2. An example of a homogeneous subset is the cluster of Items 1, 4, 6, and 11. This small subset had a Q1 of 28.8788 ($df = 32$, $p = .6253$) and a Q2 of 166.3040 ($df = 135$, $p = .0348$).

When considered simultaneously, the two data sets can be seen as a manipulation of context in which the 12 short-form items that both samples

Table 5
Two Homogeneous Subsets From the 12-Item Advanced Progressive Matrices (APM)

	Statistic	*df*	*p* Value
Items 1-6 (excluding Item 3; *n* = 377)			
Q1	9.0509	4	.0598
Q2	13.0046	10	.2234
Items 7-12 (*n* = 559)			
Q1	18.1742	15	.2536
Q2	38.0526	36	.3761

had in common have been administered either embedded in the full 36-item scale or as parts of the self-standing 12-item short form. When data on the 12 short-form items were analyzed simultaneously and context was used as splitting variable, Q1 was not statistically significant (Q1 = 21.5388, $df = 11$, $p = .0282$). This indicates that being administered as part of Set II or as part of the short form did not affect the relative difficulties of the 12 items selected for the short form. Only 1 item (short-form Item 4) showed a slight bias; it was relatively easier to solve when administered as part of the 36-item Set II.

Although the results obtained by Kubinger et al. (1991) were not replicated here with the APM, there were several examples across the present data sets indicating that the dimensionality of a scale and the performance on subsets of items were at least in part context dependent. That is, subsets of items that were identified as unidimensional (according to the Q2 statistic) in one study were not found to be unidimensional in the other study. For example, when administered as part of the 36-item Set II, Items 2 through 11 (short-form numbers) were found to be unidimensional (Q1 = 53.1890, $df = 36$, $p = .0324$; Q2 = 190.9499, $df = 175$, $p = .194$). However, when the same items were administered as part of the 12-item short form, they were not found to be homogeneous (Q1 = 49, $df = 27$, $p = .0058$; Q2 = 243.5263, $df = 140$, $p < .001$). Conversely, a group of items (short-form Items 1, 4, 6, 7, 8, 9, 10, and 11) found to be unidimensional when administered within the context of the short form (Q1 = 28.8788, $df = 32$, $p = .6253$; Q2 = 66.3040, $df = 135$, $p = .0348$) was found not to be unidimensional when administered as a part of the 36-item Set II (Q1 = 53.1825, $df = 32$, $p = .0107$; Q2 = 210.7367, $df = 135$, $p < .001$). In both of these examples, the statistical significance of Q2 was not merely the consequence of one or two poorly fitting items. Also, in both these examples, using $p < .001$ as our criterion, we found the change in Q2 but not the change in the Q1 to be substantial enough to lead to a change in the conclusions regarding the dimensionality of targeted groups of items.

Discussion

As with previous findings exploring the dimensionality of the APM, the results of the present study are somewhat ambiguous. That is, the findings of the present study will not allow us to safely assume that a single ability underlies individual differences in performance on the APM. Furthermore, if the APM is used as an index of fluid intelligence, then we cannot assume that fluid intelligence is dominated by a single underlying ability (*g*). Both of these assumptions have accompanied most of the research attempting to identify a single simple cognitive or noncognitive correlate of *g*. Thus, interpreting correlations between simple tasks, such as inspection time, and the APM may be more complicated than once considered.

Our traditional principal component analyses and nonlinear factor analyses both provided qualified support for both the multidimensional and the unidimensional models. The results from the Rasch analyses were more supportive of the multidimensional model. Neither the original APM Set II nor the Bors and Stokes (1998) short form was homogeneous in terms of either difficulty or ability. Further Rasch analyses also indicated an apparent context effect for subgroups of items, despite the minimal manipulation of context in the present two studies.

As suggested by Van den Wollenberg (1982) and others, a distinction needs to be made between the unidimensionality of difficulty and the unidimensionality of ability when testing the Rasch model. As the results of the present two studies illustrate, the distinction is important because these two aspects are not coextensive. Our data revealed a unidimensionality of difficulty for some sets of items without a concurrent unidimensionality of ability.

The Rasch model, a one-parameter logistic model, is arguably the item response theory model best suited to test the unidimensionality of psychometric instruments made of dichotomously scored items. Other item response theory models exist, however, that allow for more than one parameter, such as the two-parameter logistic model, which accommodates items with different slope values, or the three-parameter model, which introduces a guessing parameter. Although adding parameters to the model provides the researcher with more flexibility when attempting to fit a model to the data, the contribution of multiple-parameter models to the evaluation of the dimensionality of psychometric instruments, such as the APM, is not clear. Allowing for more than one parameter is to create multidimensionality. When items differ with respect to the magnitude of their correlations with the total score, it is either because they contain the latent trait in various amounts, or because the latent trait they measure is multidimensional, in which case it does not make sense to use a single total score to estimate items' discriminabilities in the first place.

These findings, particularly the fact that items can behave differently depending upon which other items are also administered, have implications for the applied use of the APM. Given the effect of context suggested by our results and those of others, the best strategy relative to estimating *g* with the APM would not be Rasch-based adaptive testing. This is the case because results suggest that items presented in one context may measure a somewhat different ability when presented in another context. This also has implications for the construction and use of short forms of such tests. Removing easy items to develop short forms for targeted populations will likely change the nature of what is measured by the more difficult items and make comparisons difficult.

References

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.

Carlson, J. S., & Wiedl, K. H. (1979). Toward a differential testing approach: Testing-the-limits employing the Raven matrices. *Intelligence*, *3*, 323-344.

Carroll, J. B., & Maxwell, S. E. (1979). Individual differences in cognitive abilities. *Annual Review of Psychology*, *30*, 603-640.

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.

Fraser C. (1993). NOHARM87 [Computer software]. Available from the University of North Texas Web site: http://www.unt.edu/rss/class/rich/5840/mcdonald/Downloads.htm

Fraser C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267-269.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525-546.

Glas, C. A. W., & Ellis, J. L. (1993). *RSP user's manual: Rasch scaling program*. Groningen, the Netherlands: iec ProGamma.

Glas, C. A. W., & Ellis, J. L. (1994). *RSP: Rasch scaling program* [Computer software]. Groningen, the Netherlands: iec ProGamma.

Jastak, J. (1949). Problems of psychometric scatter analysis. *Psychological Bulletin*, *46*, 177-197.

Jensen, A. R. (1987). Individual differences in the Hick paradigm. In P. A. Vernon (Ed.), *Speed of information processing and intelligence* (pp. 101-175). Norwood, NJ: Ablex.

Kubinger, K. D., Formann, A. K., & Farkas, M. G. (1991). Psychometric shortcomings of Raven's Standard Progressive Matrices, in particular for computerized testing. *Revue européenne de psychologie appliquée*, *41*, 295-300.

Neubauer, A. C. (1995). *Intelligenz und Geschwindigkeit der Informationsverarbeitung* [Intelligence and speed of information processing]. Vienna, Austria: Springer-Verlag.

Rost, D. H., & Gebert, A. (1980). Zum Problem der Faktoreninterpretation bei Raven's Coloured Progressive Matrices [On the problem of factor interpretation in Raven's Coloured Progressive Matrices]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *1*, 255-273.

Rummel, R. J. (1970). *Applied factor analysis* (2nd ed.). Evanston, IL: Northwestern University Press.

Spearman, C. (1939). Intelligence tests. *Eugenics Review*, *30*, 249-254.
Spearman, C. (1946). Theory of general factor. *British Journal of Psychology*, *36*, 117-131.
Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123-139.
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.
Vernon, P. A. (Ed.). (1987). *Speed of information processing and intelligence*. Norwood, NJ: Ablex.