Generation Speed in Raven's Progressive Matrices Test

Tom Verguts
Paul De Boeck
University of Leuven, Leuven, Belgium

Eric Maris
University of Nijmegen, Nijmegen, The Netherlands

In this paper, we investigate the role of response fluency on a well-known intelligence test, Raven's Advanced Progressive Matrices (APM) test. Finding rules that govern the items is critical in solving this test. Finding these rules is conceptualized as sampling rules from a (statistical) rule distribution until the correct one is attained. Response fluency is then seen as *generation speed*, or the speed at which a person generates (samples) rules from this distribution. We develop a test that isolates this speed of sampling variable, and a method to check whether this variable was adequately isolated. The score on this test is then compared with performance on the APM test. It is found that the speed at which people sample from such distributions is an important variable in solving APM items.

The Advanced Progressive Matrices (APM) test has been widely used in psychological research. Since it correlates well with many other psychometric tests (Marshalek, Lohman, & Snow, 1983), it is often taken as a measure of *g*, and compared with other variables, such as brain measures (Haier, Siegel, Tang, Abel, & Buchsbaum, 1992; Reed & Jensen, 1992), elementary cognitive tasks (Jensen, 1987; Kranzler & Jensen, 1991) or more complex tasks (Larson, Merritt, & Williams, 1988; Larson & Saccuzzo, 1989).

The APM itself has also been the focus of some research. For example, some authors have factor analyzed the APM and argued that a certain number of factors underlies performance on this test (e.g., for one factor: Alderton & Larson, 1990; for two factors: Dillon, Pohlmann, & Lohman, 1981). Embretson (1995) has applied the conjunctive Rasch model (Embretson, 1984; Maris, 1995) to the APM and argued that working memory capacity and general control processes constitute the two main sources of individual differences. A detailed Raven (1962) analysis has been performed by Carpenter, Just, and Shell (1990). Using various measures of task performance (e.g., eye movements,

Direct all correspondence to: Tom Verguts, University of Leuven, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. E-mail: tom.verguts@psy.kuleuven.ac.be

INTELLIGENCE 27(4): 329–345
ISSN: 0160-2896

number correct scores), these authors also found two factors, which they called "ability to induce abstract relations" and working memory capacity.

The work presented in this paper is an attempt to clarify the first of Carpenter et al.'s (1990) factors, the ability to induce abstract relations. More specifically, we investigate the role of *response generation speed* (or response fluency), a variable which, in our view, has been relatively neglected in intelligence research. However, this idea goes back at least to Thorndike's (1898) insight that, in order to come up with a correct response to a certain task, one must generate a lot of possibilities; the correct one can then be retained. Essentially the same hypothesis has been stated in an (ontogenetical) evolutionary framework by Campbell (1956, 1960). In fact, the present theory was derived from the same metaphor linking biology and psychology (Dewitte & Verguts, in press).

Response fluency has been empirically studied before, but usually with verbal material (Ekstrom, French, & Harman, 1976; Janssen & De Boeck, 1997; Janssen, De Boeck, & Vander Steene, 1996; Sincoff & Sternberg, 1987; for an overview, see Carroll, 1993), while the current research is concerned with geometrical tests. Geometrical fluency tests have been devised by, e.g., Guilford (1956). Some of these are now part of the ETS French kit of reference tests (Ekstrom et al., 1979).

The remainder of this paper is organized as follows. First, the generation speed principle is described in detail, followed by a section on how the predictions derived from this principle are tested. Then, come the Method and the Results section. We conclude by discussing the relevance of our findings and by stating them in a broader perspective.

Generation Speed

A typical APM item is given in Fig. 1; participants are instructed to complete the lower right-hand cell of the 3×3 matrix with one of the eight answer alternatives at the bottom. They are told to find a logical rule governing the first as well as the second row of this matrix; once found, this rule has to be applied to row three in order to complete the item.

Essentially, the thesis of this paper is that a rule generation process plays a crucial role in solving the APM items. If (APM) rules are compared with balls in an urn, this means that people sample balls from an urn. Individual differences in the generation process can be thought of as sampling from different urns (qualitative differences) or at different rates (quantitative differences). We will concentrate on the latter. Given a limited time to solve the test, and given that the "different urns effect" is cancelled out, this implies that fast persons (fast in the sense of generating many possible rules in a limited time) have a higher probability to solve a particular item correctly. Usually, a 40 min time limit is given to solve the 36 Raven items, so one has, on average, 66.7 s for each item, which is reasonably severe. The idea that sampling and checking hypotheses plays a role in solving problems is not new. Milward and Wickens (1974) describe a whole set of theories based on this idea. The present paper concentrates on one such theory, applied to a specific domain (APM items) and describes a method to test whether individual differences in rule generation speed do indeed correlate with the APM. From introspection, it might seem that rule generation speed is not the most important factor in solving APM items. However, we do not claim that our hypothesis is also valid to introspective data, as the process may be partly or wholly beyond the awareness of the participant. Furthermore, we also do not claim that solving APM items is only but a matter of rule generation speed. What we will

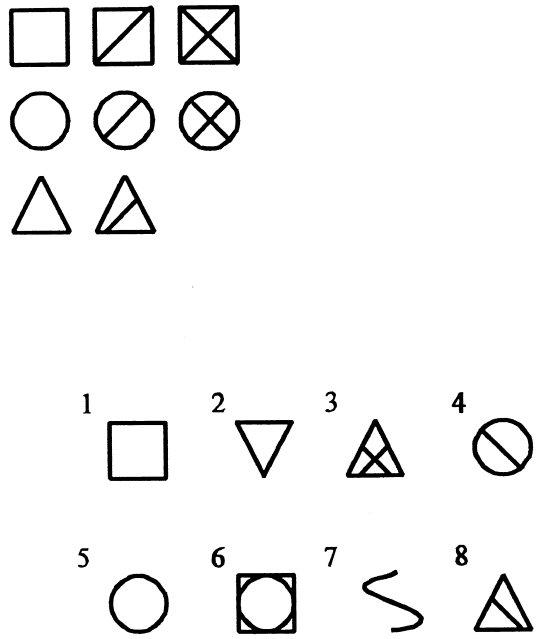

Figure 1. Analogy item (for copyright reasons, this item is invented, but it is similar to APM items).

test instead is whether individual differences in rule generation speed do exist and can be isolated in a separate generation test and whether these differences are correlated with the APM.

A Generation Task

The APM is a complex test, in which many variables other than generation speed may have an important influence. Separating these other variables from generation speed can be tackled by the method of *subtasks* (Embretson, 1980, 1984; Sternberg, 1977). This entails the construction of one or more tasks that are intended to measure one of the components under study, in our case generation speed. Comparison of such a subtask and the APM will be critical in our analysis. We have one subtask, consisting of five items, of which a typical item is given in Fig. 2.

This (sub)task will from now on be referred to as the *generation task*. Participants are asked to find a rule that might be applied to the two patterns presented. The two patterns are to be seen as an incomplete row that in its full length consists of three patterns (as in the APM test). Together with the rule they found, they are asked to draw the pattern consistent with their rule, thus completing the row. Participants have to do this, for each item, in as many ways as possible. One point is scored for every rule they found. Therefore, participants are assigned a score theoretically between 0 and infinity for each item. The score of each participant on this task (summed over the items) is the *generation score*. It is assumed that other variables are much less important here than in the APM test.

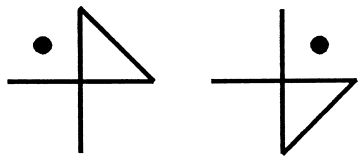

Figure 2. Generation task item.

One reason is that working memory load is much lower in this test than in the APM, in which it is an important variable of interindividual differences (Carpenter et al., 1990; Embretson, 1995). A second reason is that it was emphasized to the participants that the quality of the rules they generated would not be taken into account, whereas rule quality is important in the APM test, because the solution can be found only if the correct rule was sampled.

As noted above, we want to measure the number of rules someone can generate in a given amount of time. The generated rules will be grouped into equivalence classes in order to exclude the exploitation of one rule. For example, one rule in the item of Fig. 2 might be "add pattern 1 and 2 to obtain pattern 3." Another one is "take pattern 1, rotate 45°, and add pattern 2 to obtain pattern 3." These are counted as two separate rules. However, the rule "take pattern 1, rotate 90°, and add pattern 2 to obtain pattern 3" is not counted as a third rule since it is a simple variation of the second one. Interrater reliabilities for this procedure will be discussed later in the paper.

To summarize then, we use two tasks, one of which is a complex task, the APM test, while the second one is intended to measure generation speed only. Four hypotheses to be tested will now be described.

(1) Hypothesis 1 is that between participants, quantitative differences exist in the generation task. As noted before, the notion of "quantitative" in Hypothesis 1 refers to a speed (of sampling) variable. This first hypothesis refers to the fact that reliable variance exists in the generation test scores and is of a preliminary nature. Indeed, if this variance is too low, one cannot expect to find any correlation between the generation test and the APM.

(2) Hypothesis 2 is that between participants, no qualitative differences exist in the generation task. The term "qualitative" refers to the distribution that participants are assumed to sample from. In the urn metaphor, Hypothesis 1 describes the speed of sampling from the urn, while Hypothesis 2 assumes that urns do not differ between persons in the generation task. Otherwise stated, the second hypothesis refers to the fact that all participants sample the same types of rules in the generation task; a rationale for this assumption was given above. It is necessary that we find no, or few, qualitative differences in the generation task, otherwise the number of rules may be confounded with their quality, and hence the correlation between the APM and the generation test could no longer be interpreted as support for the role of generation speed. It should be stressed that the hypothesis of absence of qualitative differences in the generation task in no way implies that in the APM also there are no qualitative differences. One can reasonably assume that people do differ in the quality of rules they generate in the APM, but the question is to what extent purely quantitative differences are important as well. However, in order to isolate a generation speed variable, it must be guaranteed that in the generation

task under consideration there is no confounding of speed and quality. This is not a trivial issue, and it is a major challenge to succeed in such an isolation of speed from quality. In fact, the hypothesis of no qualitative differences in the generation task is far from evident, not to say counterintuitive.

The first two hypotheses are preliminary hypotheses, in the sense that if any of these two hypotheses turns out to be invalid, testing the other hypotheses is meaningless. For example, if the second hypothesis turns out to be invalid, then qualitative differences do exist in the generation task, so that obtaining a correlation between the generation task and the APM cannot be interpreted as evidence for the role of speed-of-sampling.

(3) Hypothesis 3 states that differences in rule generation speed (as isolated in the rule generation task) play a role in the APM. However, since the rule finding process is quite easy in some APM items, it may be expected that the generation task cannot correlate with items that are easy for rule finding, since it is impossible to find a correlation if one of the variables shows hardly any variance. Therefore, the third hypothesis should be modified slightly by saying that the generation test will only correlate with those items in which rule finding is (moderately) difficult. We will return to this nuance in the following.

(4) Hypothesis 4 is that a rule finding dimension can be found in the APM items. If rule finding is a process in solving APM items and if items differ with respect to this process, then it must be possible to isolate a rule finding dimension in the APM items. As noted before, generation speed is presumed to be an aspect of the rule finding we refer to here. The four hypotheses will be dealt with consecutively in the following section.

Hypothesis Testing

Hypothesis One

Testing of the quantitative differences hypothesis is done using a randomized blocks ANOVA design: each participant is treated as its own block, the five items are treated as the experimental conditions. Contrary to usual practice, no prediction is made about the experimental variable, but a strong effect is predicted for the blocking variable. It is useful to note at this point the relation between this procedure and a classical way to calculate test reliability based on ANOVA methods (Hoyt, 1941). This author presents a method to obtain the well-known coefficient α reliability estimate. Essentially, this method boils down to comparing total sum of squares with person and item specific sum of squares (formally, $SS_t - SS_p - SS_i$). The present procedure is the same, but we do not treat the result as the usual coefficient α estimate, since we are (here) not interested in the reliability of the test per se, but rather in whether reliable person variance in the generation task can be obtained.

Hypothesis Two

To describe our testing of the second hypothesis, let us consider an arbitrary item i of the generation task (so $i = 1, \ldots, 5$), and, say, R responses (rules) are possible for this item (R may vary over items). Each participant p generates a response vector

$$(x_1, x_2, \ \ldots \ x_{z_p}),$$

where z_p refers to the number of rules given by participant p. The x variables may take on a number from 1 through R, the total number of rules. The x-value indicates which of the R rules is generated as the first one (x_1), as the second one (x_2), etc., to the last response (x_{z_p}). Quantitative differences and, hence, the values z_p, are not studied in this section. The values z_p will be implicitly conditioned upon. With each rule r, a probability π_{pr} is associated for participant p, which gives the probability that p will generate rule r. For the generation task, our hypothesis then entails that

$$\pi_{pr} = \pi_r \qquad (p = 1, \ \ldots \ , N; \quad r = 1, \ \ldots \ , R) \tag{1}$$

in other words, π_r is constant across participants. A complication is that sampling occurs without replacement; participants never repeat the same rule. Therefore, the sampling probabilities are renormed after every sampling, in a way depending on the rule sampled in the step before. Therefore, the probabilities in (1) are the sampling probabilities only in the first step; we will refer to them as *the initial probabilities*. Formally, if X_{pj} is the variable indicating the jth response of participant p, its probability is written as

$$\text{Prob}(X_{pj} = k) = \gamma(x_{p1}, \ \ldots \ , x_{p(j-1)})\pi_k \tag{2}$$

The factor $\gamma(x_{p1}, \ \ldots \ , x_{p(j-1)})$ is introduced to make the probabilities add to one in every sampling round. Details concerning the precise form of γ and parameter estimation are provided in Appendices A and B. Two methods are used to test Hypothesis Two.

The first consists of estimating the parameters π using only the Low-group, that is, the participants scoring below the mean score per item. For each item, the mean score is between 2 and 3, so the Low-group in each case consists of participants giving 0, 1 or 2 rules. The High-group consists of the remaining participants (scores above 2). Similarly, parameters are estimated using the High-group only. Notice that participants may switch between (Low/High) groups across items. Then, per item, a two-factor ANOVA is performed on these estimated probabilities, with Low/High serving as the first factor and the R rules as the second (resulting in a design with $2 \times R$ cells, one observation per cell). Hence, we take one of the columns of the person $\times$ item datamatrix used in Hypothesis One, sort the datapoints into the different rule groups (1, ... , R; constituting the columns of the new matrix) and persons in two groups (Low/High, constituting the rows of the new matrix). The percentage of variance explained by the rules (= SS_{rules}/SS_{tot}) is taken as a measure of the adequacy of Hypothesis Two. Evidently, if the hypothesis is completely correct, then $SS_{rules}/SS_{tot} = 1$ for each item.

The second test is based on the same dichotomizing principle and is an application of the recent *Posterior Predictive Check* (PPC) method in Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 1995; Gelman, Meng, & Stern, 1996; Meng, 1994;). A complete elaboration of these principles is beyond the scope of this paper; a brief statement is provided in Appendices A and B. Here, we suffice to mention that it allows a very flexible way to apply test statistics to data because, contrary to the

classical statistical paradigm, explicit reference distributions need not be derived. The statistic we will use is of the form

$$T = 1 - \frac{\mathrm{SS}_{\mathrm{Rules}}}{\mathrm{SS}_{\mathrm{Tot}}} = 1 - \frac{2\sum_{r=1}^{R}(p_{.r} - 1/R)^2}{\sum_{r=1}^{R}(p_{\mathrm{L}r} - 1/R)^2 + (p_{\mathrm{H}r} - 1/R)^2} \tag{3}$$

in which $p_{\mathrm{L}r}$ and $p_{\mathrm{H}r}$ denote the proportion of times that rule r is given in the Low- and High-group, respectively, and $p_{.r} = (p_{\mathrm{L}r} + p_{\mathrm{H}r})/2$. The value $1/R$ is the within groups as well as the between groups mean, since $\sum p_{.r} = \sum p_{\mathrm{L}r} = \sum p_{\mathrm{H}r} = 1$. As mentioned before, the Low-group are those participants who achieve a score of 2 or less on the particular item under study (of the generation task), the High-group are those who score higher. As can be seen, the statistic (3) measures the amount of variance not attributable to the rules in the observed response proportions. Evidently, our hypothesis implies that the statistic T evaluated in the observed data is low (i.e., close to 0).

A main effect of the grouping (Low/High) variable is not possible since proportions add to one. Qualitative differences between groups will show up in the interaction term.

It is clear that, if the rule generation probabilities do not differ depending on the number of generated rules (that is, between the Low-group and the High-group), there is no confounding between the quantitative measure of generation and the quality of the rules generated.

Hypothesis Three

Since generation speed is deemed to be a variable with at least some importance in the APM, a first prediction derived from Hypothesis Three is that the generation task is correlated with the APM.

However, for many Raven items, rule finding is very easy (i.e., the probability of sampling the correct rule is close to 1). Hence, individual differences in generation speed will not be visible in these items, implying a low correlation between performance on these items and the generation score. Similarly, items with correct-rule sampling probabilities close to 0 would not correlate with the generation score either. The proportion of success in the APM is never below 0.27 in our data, which implies that items of this second type are not present in the Raven test. Generation scores will therefore be compared with two subscales of Raven items, one in which rule finding is very easy and one in which rule finding is moderately difficult. The differentiation of these two scales allows us to test the hypothesis more clearly: Indeed, a correlation with scale 2 but not with scale 1 would rule out the alternative explanation of motivation as a mediating variable (assuming that the scales have about equal standard deviations), since motivation should not differentiate between these two scales.

Two mutually exclusive sets of Raven items are constructed. Two external observers are asked to rate on a 10-point scale the difficulty of *finding* (this is emphasized to them) the rule in a particular item. Using these data, we construct two

scales, the first of items with easy rule finding (items 22, 23, 24 and 26, *scale 1*) and items where rule finding is relatively difficult (items 32, 33, 35 and 36, *scale 2*). The two predictions of Hypothesis Three can then be stated succinctly as

$$\text{corr}(\text{Generating score},\ \text{Raven sum}) > 0$$

and

$$\text{corr}(\text{Generating score},\ \text{scale 1}) < \text{corr}(\text{Generating score},\ \text{scale 2}).$$

As a final point, one may question the division of items in two scales; for scale 2, one or two items might correlate with the generation task, resulting in a high overall correlation. On the other hand, some items of scale 1 might correlate negatively with the generation task, thus hiding the effect of the other items. To investigate this, correlations could be calculated for each item separately. Calculating correlations with binary items is awkward, however, since the proportion-correct values heavily influence the correlation. To circumvent this problem, we perform a (univariate) logistic regression for each item of scales 1 and 2 with generation score as a predictor.

Hypothesis Four

The fourth hypothesis concerns the APM items only, and entails that these items can be ordered on a rule finding difficulty scale. A common method of describing variables on an underlying dimension is to perform a factor analysis (e.g., Alderton & Larson, 1990; Dillon et al., 1981). For APM data, this procedure is problematic for two reasons. First, the variables involved are binary instead of continuous. Second, if other factors (beside generation speed) are important in solving APM items, we feel that it is more plausible that these factors cannot compensate one another. In a factor analysis, on the other hand, high scores on one factor can correct for low scores on other factors. A model that takes account of both problems is the *conjunctive Rasch model* (Embretson, 1980; Maris, 1995), which for current purposes, can be written as

$$\Pr(Y_{pi} = 1) = \frac{\exp(u_p - \beta_i)}{1 + \exp(u_p - \beta_i)}\,\mu_{pi}, \tag{4}$$

where the variable Y_{pi} indicates the score (1/0) on APM item i for person p. The first factor in the right hand side of Equation (4) denotes the probability that person p will find the correct rule for item i. The variable u_p indicates person p's ability. The values β_i denote the rule finding difficulties for items i. Only the items which are a member of scale 1 or 2 will be focused on here. Other items are either too easy (item numbers 1 through 20) or difficult to classify. Therefore, if $\beta_i^{(1)}$ and $\beta_i^{(2)}$ denote the scale 1 and scale 2 item difficulties, respectively (see Hypothesis Three), it is predicted that $\beta_i^{(2)} > \beta_j^{(1)}$ for all i and j.

The second factor (μ_{pi}) denotes the probability that person p will solve the item given that she has found the correct rule. This factor is of the same functional form as

the first one, but since only rule finding difficulty is studied here, we will not elaborate on this factor, that is we will not introduce a second θ and β.

Method

Participants

Participants are 127 undergraduate psychology students of Leuven university who received course credit for their participation. Of the 82% of participants of whom sex is known (they were allowed to use an alias), 18% are male.

Procedure

In the first session, the APM test was taken, for which 40 min were allowed. One week later came, the second session in which the generation task was administered. In the instructions, participants are encouraged to give as many rules as possible for each item. Here, a 20 min time limit was given. Scoring of the generation task proceeded as follows. First, per item, all generated rules were grouped into equivalence classes (resulting in R classes per item, see above). Of course, different persons seldom gave exactly the same rule; everyone formulates the rules in her own way. Then, it was determined which rules each person had given. Finally, external raters were explained the purpose of the study and asked to determine how many rules each person had generated; this allowed calculation of an interrater reliability score.

Results

The generation task and the APM test have a split-half reliability of 0.86 and 0.80, respectively. The generation task has an interrater reliability of 0.90. Other reliabilities will be mentioned throughout the text.

Hypothesis One

The ANOVA indicates clearly that participants (blocks) have a large effect ($p < 0.001$). Equivalently, we may note that coefficient $\alpha = (MS_{pers} - MS_{int})/MS_{pers} = 0.796$, and a formal test (Feldt, 1965; Verhelst, 1998) indicates that correspondingly, $p < 0.001$.

The item (experimental) variable had virtually no effect ($p > 0.05$). The rather low ω^2 implies that much of the variance is to be attributed to an item–person interaction, meaning that generation speed differences between individuals also depend on the item or that they partially reflect random variation. Nevertheless, since the blocking variable has a strong effect, as indicated by the ANOVA analysis, one may conclude that quantitative differences do exist between participants on the generation task. Given that $\alpha = 0.796$, the differences seem to be captured in a sufficiently reliable way.

Hypothesis Two

A second question is whether participants also differ in a qualitative manner on the generation task. For each item, parameters are estimated using only the Low-group and using only the High-group. Estimates for items 1 to 5 are shown in Table 1. The number of rules varies from 16 to 32, so some rule probabilities are not shown. The probabilities not

Table 1. Rule Probabilities Estimated for Low- and High-Group Separately

		Rule Numbers									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Item 1	Low	0.27	0.31	0.29	0.04	0.02	0.01	0.01	0.00	0.01	0.06
	High	0.41	0.27	0.19	0.04	0.01	0.01	0.01	0.02	0.02	0.04
Item 2	Low	0.01	0.30	0.01	0.08	0.21	0.13	0.01	0.08	0.00	0.00
	High	0.00	0.16	0.00	0.06	0.28	0.28	0.02	0.05	0.01	0.04
Item 3	Low	0.02	0.43	0.08	0.08	0.03	0.01	0.00	0.00	0.02	0.01
	High	0.01	0.17	0.11	0.20	0.14	0.01	0.01	0.01	0.03	0.00
Item 4	Low	0.11	0.17	0.03	0.16	0.28	0.00	0.07	0.03	0.07	0.00
	High	0.14	0.07	0.02	0.22	0.33	0.01	0.08	0.00	0.01	0.00
Item 5	Low	0.62	0.00	0.00	0.17	0.06	0.01	0.00	0.05	0.01	0.00
	High	0.35	0.04	0.03	0.26	0.19	0.01	0.02	0.05	0.01	0.01

displayed are always close to zero (max = 0.07). Across items, rules cannot be compared. Some rules do not appear in one of the groups; if this happens, its estimator is 0 and we write a 0 in the corresponding cell.

The number of participants in each group is always around 60. As can be seen, only on items 3 and 5 non-negligible differences between Low- and High-group occur. This is reflected in the percentages of variance accounted for by the rules: These are 96%, 93%, 83%, 98% and 91%, respectively, with a mean of 92.2%. These percentages indicate that differences between rules are far more important than pattern differences between groups. Even for the most deviant items, the ordering of probabilities is almost the same across groups. Since the probabilities in each group add to one, the main group effect is zero, and the remaining differences will show up in the group $\times$ rule interaction term. These remaining differences are of a qualitative kind and their "percentages of variance accounted for" are the complements of the percentages just mentioned.

Second, we perform a Bayesian PPC analysis based on the statistic T in Equation (3). The observed values T for each item are 0.041, 0.068, 0.171, 0.044 and 0.114, respectively, a result similar to that obtained with the previous method. For example, both methods indicate item three to be the worst fitting item. Note, however, that proportions accounted for by the rules are not completely similar for both methods, since the first method is based on estimated probabilities (i.e., formula (2)) and the second on observed proportions (p_{Lr} and p_{Hr}).

Applying the PPC procedure results in p-values of 0.060, 0.013, 0.000, 0.174 and 0.000, for items 1 through 5, respectively, which shows that items 3 and 5 (items which were also most deviant according to the first method) do not fit the model. Yet, proportions of variance accounted for are high for all items. So, although the model does not strictly fit the data, we can conclude nevertheless that response sampling probabilities do not vary strongly between persons. We will come back to this result in the following paragraphs.

Hypothesis Three

The question whether the (generation task) speed variable is relevant in the APM task, our main research question, will now be investigated. Table 2 shows that the first

Table 2. Task Intercorrelations

	Scale 1	*Scale 2*	*Generation Score*	*Reduced Generation Score*
APM Score	0.624*	0.657*	0.428*	0.402*
Scale 1		0.349*	0.147	0.158
Scale 2			0.370*	0.317*
Generation Score				0.944*

*$p < 0.001$, two-tailed.

Table 3. Regression Weights of Generation Score on APM Items of Scales 1 and 2

APM Item Number	*Scale Number*	*Regression Weight*	*p-Value*
22	1	0.029	0.529
23	1	0.057	0.309
24	1	0.069	0.118
26	1	0.047	0.269
32	2	0.094	0.035
33	2	0.206	0.000
35	2	0.086	0.049
36	2	0.081	0.092

prediction is confirmed. The correlation between the APM and the generation score is 0.428 ($p < 0.001$). Concerning the second prediction, about scales 1 and 2 of the APM, it turns out that only scale 2 correlates with the generation task, as is clear from Table 2. The difference between 0.147 and 0.370 is significant ($p < 0.01$). The two scales do not really differ as to their variance, with standard deviations of 1.11 and 1.21, respectively, with an equal number of items and, hence, the difference cannot be explained by a difference in interindividual variance. Moreover, scale 2 correlates higher with the generation task than with the other set of APM items (scale 1).

The testing of Hypothesis Two indicated that items 3 and 5 did not fit the requirement of no qualitative differences. We therefore also present the results with these two items excluded from the analysis. This creates a new variable, denoted *reduced generation score*, and it is given in the last column of Table 2. Correlations are slightly diminished, but the main results, however, remain valid. The shrinkage can be explained by the lowered reliability (Cronbach's $\alpha = 0.80$ and 0.68 for generation score and reduced generation score, respectively), due to a reduction from five to three items.

Finally, it is needed to check whether all items in one scale behave similarly with respect to generation score; more specifically, we wish to ascertain that all items are either related (scale 2) or not related (scale 1) to the generation score. We therefore perform a logistic regression for each item on this variable. The regression weights and corresponding *p*-values of the regression weights (slopes) for items of both scales are given in Table 3. This Table 3 shows that all items in the scale behave appropriately, that is, items of scale 1 are not (or weakly) related to generation score, while items of

Table 4. Conjunctive Rasch Item Difficulties (One Dimension)

APM Item Number	*Scale Number*	*Item Difficulty* β
22	1	−1.957
23	1	−1.018
24	1	−2.066
26	1	−1.261
32	2	−0.099
33	2	−0.390
35	2	−0.272
36	2	−0.220

scale 2 are. Given that the result is consistent over items, it may be considered a robust finding.

Hypothesis Four

The final hypothesis, then, states that the generation dimension can be identified in APM data alone. We applied a conjunctive Rasch estimation program (Maris, Verguts, & Tuerlinckx, 1997) to the APM data that provides us with estimated item parameters. The item difficulty parameters β of scales 1 and 2 for the conjunctive Rasch model as described above are given in Table 4. These results show that one of the APM dimensions can be identified as a rule finding difficulty parameter, since all item difficulties β are higher for scale 2 than for scale 1. That all values are negative is not important here: The values are not to be considered weights but tresholds that are identified only up to an additive constant.

As a conclusion then, we can state that the rule generation is an important factor in finding rules in APM items, and that mere speed of generation, independent of the quality of the rules one tends to generate, can explain part of the variance in the APM. Not all variance was accounted for, since we found a second dimension as well (see Hypothesis Four) and since the correlation of generation speed with the APM items is far from perfect.

Discussion

In the past, solving analogy problems (like the APM items) has been the focus of a wide array of research. One line of research has stressed the importance of working memory capacity (e.g., Kyllonen & Christal, 1990; Mulholland, Pellegrino, & Glaser, 1980; Simon & Kotovsky, 1963). Others have argued for the importance of metacognition or control processes (Embretson, 1995; Sternberg, 1985). Some authors have constructed new Raven-type items in order to facilitate model testing (Embretson, 1995; Hornke & Habon, 1986).

Evidence for the importance of speed variables for complex tasks such as the APM has been established also. Sternberg (1977) and Sternberg and Gardner (1983), for example, found correlations between speed of performing elementary components and performance on complex induction tasks. In the cognitive correlates approach, Jensen (e.g., Jensen, 1987; Jensen & Munro, 1979; Reed & Jensen, 1991, 1992) has extensively studied relations between complex tasks and elementary speed tasks. The

book of Vernon (1987) is dedicated to the subject. The idea goes back at least to Galton (1883).

We have hypothesized that sheer speed of sampling plays a role in finding a correct rule and, hence, in solving APM items. The generation task that was constructed to measure speed independently of quality, was shown to correlate with the APM and especially with items where rule finding is important. However, speed of generation does not seem to tell the whole story. Other variables may play a role: Participants may differ in the quality of rules sampled in the APM, as we did not test the no-qualitative-differences hypothesis for the APM but only for the generation test. Also, participants may differ in the accuracy of applying generated results. This last factor is probably related to working memory capacity which was found to be important in previous research (e.g., Embretson, 1995), as one has to keep (intermediate) generation results in mind while applying other generation results.

That we were able to construct a test with only minor differences in the quality of the rules generated between participants with a high generation score and participants with a low generation score means that speed as such is an important generation aspect. The correlation of this variable with the APM indicates that generation speed helps the quality of the result in a complex task (with time constraints, as in most tests) like the APM. Chances of finding a good rule do increase with generation speed even without differences in availability of rules: Trying out more rules helps to find the solution.

Our results cannot be interpreted as a correlation due to overall speed, given the time constraints of the test, as the correlation was shown to be specific for items with a rule finding challenge. This makes an explanation in terms of general speed implausible and it supports the importance of generation speed. Carpenter et al. (1990) have identified two processing components: ability to induce abstract relations and ability to handle goals. Our paper focuses on the first of these: It was shown that productivity (or speed) of sampling from a distribution of rules is correlated with rule finding and hence with the ability to induce abstract relations. Therefore, it may be concluded that the APM measures, among other variables, the speed at which different solutions are generated by a testee.

Acknowledgements: The authors wish to thank Siegfried Dewitte, Michel Meulders, Gert Storms, Francis Tuerlinckx and Iven Van Mechelen for their fluency in providing many useful comments.

Appendix A. ML Estimation

Equation (2) can be written in the form

$$\text{Prob}(X_{pj} = k) = \frac{\pi_k}{1 - \sum_{m=1}^{j-1} \prod_{r=1}^{R} \pi_r^{I_{(r)}(x_{pm})}} \tag{A1}$$

Table A1. Simulation Study: Means and RMSDs

π	*Mean*	*RMSD*
0.09	0.089	0.019
0.09	0.087	0.014
0.09	0.087	0.013
0.09	0.090	0.012
0.09	0.084	0.014
0.09	0.094	0.013
0.09	0.092	0.010
0.09	0.091	0.017
0.09	0.091	0.017
0.09	0.088	0.019
0.01	0.012	0.004
0.01	0.011	0.005
0.01	0.010	0.003
0.01	0.013	0.005
0.01	0.010	0.002
0.01	0.012	0.005
0.01	0.010	0.003
0.01	0.009	0.002
0.01	0.008	0.004
0.01	0.010	0.003

In Equation (A1), $I_{\{A\}}(B)$ is an index function defined as 1 if $A = B$ and 0, otherwise. A summation running from 1 through 0 is defined as 0. The denominator of Equation (A1) sees to it that in every sampling round, the probabilities add to one. Parameters of the model (Equation A1) are estimated according to the Maximum Likelihood (ML) principle. A slightly modified version of the steepest ascent method is implemented to perform the ML estimation (Gill, Murray, & Wright, 1995). To demonstrate accuracy of estimation (technically, goodness of recovery), a small simulation study is conducted. The number of rules generated by a person is drawn from a Poisson distribution with rate $\lambda = 4$. Number of participants is equal to 100 and number of parameters (i.e., rule probabilities π) equals 20. The first 10 rule probabilities are set to 0.09 and the last 10 to 0.01 (so that $\sum_r \pi_r = 1$). Ten datasets are generated, with person rates (from the Poisson distribution) fixed across datasets. In Table A1, we report mean estimated values, that is,

$$\pi_{r.} = \frac{1}{10} \sum_{j=1}^{10} \pi_{rj}$$

where π_{rj} denotes the jth estimate for parameter r. Furthermore, we report root mean squared deviation (RMSD), which is defined as

$$\text{RMSD} = \sqrt{\frac{1}{10} \sum_{j=1}^{10} (\pi_{rj} - \pi_{r.})^2}$$

The RMSD provides an estimate of the stability of estimation across datasets. One may note from Table 5 that both measures (mean estimate and RMSD) are satisfactory for all parameters.

Appendix B. Bayesian and PPC Details

In the Bayesian paradigm, all parameters in the model are assumed to have a *posterior distribution* given the data, denoted by $p(\pi|\text{data})$, where π refers to the complete set of parameters. A posterior distribution $p(\pi|\text{data})$ for a vector of probabilities π can be constructed as

$$p(\pi \mid \text{data}) \propto p(\text{data} \mid \pi)p(\pi)$$

where $p(\pi)$ denotes the prior distribution for π. This prior is taken to be a Dirichlet distribution with all parameters equal to 1.1, which is a rather flat prior. The data part $p(\text{data}|\pi)$ has formula (A1) as its building block, and equals the likelihood function in the data. To perform the PPC process, we sample parameters from this posterior using a procedure called the Gibbs sampler (Gilks, Richardson, & Spiegelhalter, 1996, Chap. 1). When the Gibbs sampler has converged (i.e., when parameter values are a true sample from the posterior distribution), a parameter sample of size M is generated, denoted by $\pi^1, \ldots, \pi^M$. Each such set π^m generates a new (replicated data set), denoted by $\text{data}^{\text{rep},m}$. Next, we evaluate the statistic T in each replicated data set, denoted $T^{\text{rep},m}$, and we compute the same statistic T as a function of the observed data, denoted T^{obs}. Finally, we compare the proportion of times that $T^{\text{rep},m} \geq T^{\text{obs}}$, $m = 1, \ldots, M$. This proportion is our (Bayesian) p-value. Low proportions indicate that T^{obs} is rather high compared with T-values generated under the model and, hence, that the model does not adequately fit to the data with regard to the statistic T.

References

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

Campbell, D. T. (1956). Adaptive behavior from random response. *Behavioral Science, 1*, 105–110.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review, 67*, 380–400.

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

Carroll, J. B. (1993). *Human cognitive abilities*. New York: Cambridge Univiversity Press.

Dewitte, S., & Verguts, T. (in press). Behavioral variation: A neglected aspect in selectionist thinking. *Behavior and Philosophy.*

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

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Embretson, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479–594.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49*, 175–186.

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

Feldt, L. S. (1965). The approximate sampling distribution of Kuder–Richardson reliability twenty. *Psychometrika, 30*, 357–370.

Galton, F. (1883). *Inquiries into human faculty and its development*. London: Macmillan.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D.B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1–19). London: Chapman & Hall.

Gill, P. E., Murray, W., & Wright, M. H. (1995). *Practical optimization*. London: Academic Press.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin, 53*, 267–293.

Haier, R. J., Siegel, B., Tang, C., Abel, L., & Buchsbaum, M. S. (1992). Intelligence and changes in regional cerebral glucose metabolic rate following learning. *Intelligence, 16*, 415–426.

Hornke, L. F., & Habon, M. W. (1986). Item construction and evaluation with the linear logistic model. *Applied Psychological Measurement, 10*, 369–380.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153–160.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21*, 37–50.

Janssen, R., De Boeck, P., & Vander Steene, G. (1996). Verbal fluency and verbal comprehension abilities in synonym tasks. *Intelligence, 22*, 291–310.

Jensen, A. R. (1987). Process differences and individual differences in some cognitive tasks. *Intelligence, 11*, 107–136.

Jensen, A. R., & Munro, E. (1979). Reaction time, movement time and intelligence. *Intelligence, 3*, 121–126.

Kranzler, J. H., & Jensen, A. R. (1991). The nature of psychometric *g*: Unitary processes or a number of independent processes? *Intelligence, 15*, 397–422.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence, 14*, 389–433.

Larson, G. E., & Saccuzzo, D. P. (1989). Cognitive correlates of general intelligence: Toward a process theory of *g*. *Intelligence, 13*, 5–31.

Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: Some implications of task complexity. *Intelligence, 12*, 131–147.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523–547.

Maris, E., Verguts, T., Tuerlinckx, F. (1997). *Estimation of the marginal conjunctive Rasch model.* Manuscript in preparation.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*, 107–127.

Meng, X. L. (1994). Posterior predictive *p*-values. *Annals of Statistics, 22*, 1142–1160.

Millward, R. B., & Wickens, T. D. (1974). Concept-identification models. In D. H. Krantz, R. C. Luce, R. D. Luce, & P. Suppes (Eds.), *Contemporary Developments in Mathematical Psychology, Vol. 1: Learning, Memory, and Thinking* (pp. 45–100). San Francisco, CA: W. H. Freeman and Company.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*, 252–284.

Raven, J. C. (1962). *Advanced Progressive Matrices. Set II*. London: H. K. Lewis.

Reed, T. E., & Jensen, A. R. (1991). Arm nerve conduction velocity (NCV), brain NCV, reaction time, and intelligence. *Intelligence, 15*, 33–47.

Reed, T. E., & Jensen, A. R. (1992). Conduction velocity in a brain nerve pathway of normal adults correlates with intelligence level. *Intelligence, 16*, 259–272.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70*, 534–546.

Sincoff, J. B., & Sternberg, R. J. (1987). Two faces of verbal ability. *Intelligence, 11*, 263–276.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge Univiversity Press.

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology, 112*, 80–116.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplements, 2*.

Verhelst, N. D. (1998). *Estimating the reliability of a test from a single test administration.* Manuscript submitted for publication.

Vernon, P. A. (Ed.). (1987). *Speed of information processing and intelligence*. Norwood, NJ: Ablex.