Variation in Raven’s Progressive Matrices scores across time and place

Symen A. Brouwers a,⁎, Fons J.R. Van de Vijver a,b, Dianne A. Van Hemert c

a Tilburg University, The Netherlands
b North-West University, South Africa
c University of Amsterdam, The Netherlands

Abstract

The paper describes a cross-cultural and historical meta-analysis of Raven’s Progressive Matrices. Data were analyzed of 798 samples from 45 countries (N=244,316), which were published between 1944 and 2003. Country-level indicators of educational permeation (which involves a broad set of interrelated educational input and output factors that are strongly related to economic development), the samples’ educational age, and publication year were all independently related to performance on Raven’s matrices. Our data suggest that the Flynn effect can be found in high as well as low GNP countries, although its size is moderated by education-related sample and country characteristics and seems to be smaller in developed than in emerging countries.

© 2008 Elsevier Inc. All rights reserved.

Raven’s Progressive Matrices are a series of multiple-choice items of abstract reasoning. Each item depicts an abstract pattern in a two by two or three by three matrix; all cells contain a figure except for the cell in the right lower corner. Participants are asked to identify the missing segment that would best complement the pattern constituted by the other cells among a set of alternatives that are positioned beneath the matrix. John C. Raven published the first version of the test in 1938 and a revised version in 1956; the three versions of the test (Advanced, Colored, and Standard Progressive Matrices) have since been among the most widely-used intelligence tests. Its intuitively appealing question format and the use of figure stimuli have made the test attractive for cross-cultural comparisons. A meta-analysis of cross-cultural intelligence test scores showed that the Raven is the second most used test after the Wechsler Intelligence Scales (Raven, 2000; Rushton, Skuy, & Bons, 2004). ‘Culture-free’ (Cattell, 1940), ‘culture-fair’ (Cattell & Cattell, 1963), and ‘culture-reduced’ (Jensen, 1980) are all terms that have been proposed to describe the Raven or similar tests that do not seem to require much cultural knowledge for answering the items correctly. Particularly the first two labels are not undisputed. As early as 1966, Frijda and Jahoda argued that it is impossible to measure intelligence without the confounding influence of cultural factors, as both the definition of the concept and its expression are cultural. Nevertheless, the Raven tests are still considered to be measures of intelligence that show less influence of confounding cultural factors on the cross-national differences than any other intelligence test.

Both synchronic and diachronic evidence for variation of Raven test scores has been presented (Flynn, 1987, 2007; Lynn, 1982). The rise of intelligence test scores over time is commonly known as the Flynn effect and has been ascribed to various factors such as improved nutrition (Colom, Lluis-Font, & Andres-Pueyo, 2005), increased environmental complexity (Schooler, 1998), and socialization practices at home and at school (Williams, 1998). However, the bulk of research into the Flynn effect is based on individuals from high affluence countries. More recently, evidence begins to accumulate that the Flynn effect is not confined to high affluence countries or countries that invest strongly in education. Daley, Whaley, Sigman, Espinosa, and Neumann (2003) were the first to show a Flynn effect outside the twenty largest industrialized countries. In rural Kenya they found that performance on the Raven’s Progressive Matrices had undergone a strong increase across a fourteen-year interval. The latter study points to the potential cross-cultural generalizability of the Flynn effect. A cross-cultural meta-analysis of Raven test scores across a long period might help to examine this generalizability and to address the role of potentially moderating variables such as educational differences between countries.

Much has been written about the relation between country characteristics and individual test scores (Ceci, 1991; Flynn, 2007; Luria, 1976; Lynn & Vanhanen, 2006; Rindermann, 2007). Cross-
cultural research has led to the need to make a distinction between intelligence and intelligence test scores (Vernon, 1979). Cross-cultural Piagetian research uses a similar distinction. Here, competence is taken to be rather distinct from performance (Daszak, 1977). The conceptual differentiation of competence and performance is meant to accommodate the influence of various, potentially biasing factors that might cause a disparity between ‘real’ and ‘observed’ intelligence. Examples of such factors are previous test exposure, cultural appropriateness of an instrument and its administration procedures, in addition to confounding sample characteristics. Van de Vijver and Leung (1997) coined the term ‘method bias’ to refer to the overall impact caused by these confounding factors and there is empirical evidence to suggest that they may contribute to actual Raven performance. For example, Ombrédane, Robaye, and Plumail (1956) showed that the predictive validity of Raven test scores became stronger by repeated administration in a group of illiterate, Congolese mine workers. Moreover, retest effects due to method factors are not restricted to non-Western participants alone and are known to prevail among Westerners (e.g., Blieszner, Willis, & Baltes, 1981; Wing, 1980). Te Nijenhuis, Van Vliet, and Van der Flier (2007) were able to show in a meta-analysis that gains on intelligence test scores after retesting or intervention tend not to be related to general intelligence (‘g’). These findings are in line with our notion of method bias.

Educational indicators are relevant country characteristics in the examination of cross-cultural differences in intelligence test scores. From an ontogenetic perspective, educational indicators relate to the frequency with which people have opportunities for cognitive stimulation. The Raven’s Progressive Matrices are measures of reasoning and in order to reason people need opportunities for learning how to transform given information into conclusions (Galotti, 1989). Vygotsky (1978) directly related education to the potential gap between competence and performance. He reasoned that performance only reflects one’s actual level of development and thus only the development that is already completed. Working in rural Tanzania, Sternberg and colleagues examined the utility of dynamic testing of school-attending children (Sternberg et al., 2002). They familiarized children with the skills and strategies that are thought to contribute to success on tests of cognitive ability. A significant gain in test scores after training was observed, which was not present in the untrained group (which received the same tests the same number of times). The relationship between educational indicators and test performance at country level cannot be solely interpreted as a simple consequence of increased intellectual functioning through schooling. The influence of test bias should also be taken into account; the Raven might contain elements that benefit people from one country more than people from another country.

Educational indicators such as expenditure per capita, educational level of teachers, and enrolment rates have been shown to predict country-level scores on cognitive instruments (Van de Vijver, 1997). Educational quality indicators are known to belong to a cluster of variables that denote economic development (Georgas, Van de Vijver, & Berry, 2004). Other variables in this cluster are enrolment into primary, secondary, and tertiary education, Gross National Product, percentage of population working in service industry, use of mass media, prevalence of telephones, and population growth (the last one with a negative relation). At country level these educational indicators would together denote “educational permeation”, which refers to the degree in which formal education has permeated society and might on average be encountered by the population of that society. Countries with a high educational permeation thus have many schools and these schools have high quality teaching materials and qualified teachers, a highly educated population, and a high demand for jobs that require higher education.

We present here a meta-analysis of studies that reported data on Raven’s Progressive Matrices, comprising samples of children and adults from 45 countries covering a period of 60 years. Publication year, educational permeation (measured by a broad set of interrelated educational input and output factors at country level), and educational age are the three most important variables that are examined in the analysis. Based on the literature, we expect Raven performance to increase with educational age (operationalized as the average number of years of schooling of the study sample) and indicators of educational quality (at country level), and we expect an increase of performance scores over time (Flynn effect).

1. Method

1.1. Sample

Studies that report data on Raven’s Progressive Matrices were located through PsycInfo (1887 to 2003), the Social Sciences Citation Index, the Researcher’s Bibliography for Raven’s Progressive Matrices and Mill Hill Vocabulary Scales (Court, 1995), and the catalogue of Dutch libraries. In addition, a request for data was sent to 200 authors around the world, plus mailing lists in relevant research areas. Other reports were found through snowballing on the basis of reference lists in studies already identified. Data that concerned Standard Progressive Matrices (SPM), Colored Progressive Matrices (CPM), and Advanced Progressive Matrices II (APM II) were included. Sample sizes and raw mean or median scores had to be available for all cases. Clinical populations, mentally retarded groups, and other samples selected solely on the basis of intellectual capacity were not included in the present study.

The total sample consists of 193 studies; scoring all individual samples separately for age and gender resulted in a total number of 798 subsamples; the total sample size was 244,316. The data set involves 45 countries and covers the period from 1944 to 2003. Table 1 presents the distribution of the 798 subsamples across 45 countries and 60 years of publication. There is a clear bias in the distributions across country and year of publication. The United Kingdom, the United States of America and Poland have seen many studies, whereas countries as varied as Venezuela, Syria, Sweden, South Korea, Qatar, Norway and Mexico have all seen only one study (and most countries have never seen any study). The distribution of studies over time is skewed towards the present, with particularly high numbers for the period between 1984 and 1993. The same is true for the number of cultures per year. Data from many cultures were reported in the mid-1990s studies, but data from very few different cultures were reported until 1981. Of the different versions (APM, CPM, and SPM), the SPM is by far the most used (62.3% of 798 samples), followed by the CPM with 27.3% and the APM with only 10.4%.

Table 1

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Number of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Countries</td>
<td></td>
</tr>
<tr>
<td>Congo, France, Mexico, Norway, Qatar, South Korea, Sweden, Syria, Venezuela</td>
<td>1</td>
</tr>
<tr>
<td>Austria, Belgium, Brazil, Denmark, Egypt, Germany (East), Iceland, Ireland, Japan, Kenya, Nigeria, Singapore, Spain, Czechoslovakia, Ghana, Hong Kong, Israel, Italy, Netherlands, Romania, Taiwan, Tanzania, Yugoslavia</td>
<td>2 to 10</td>
</tr>
<tr>
<td>Argentina, Australia, Canada, China, Germany (West), India, Iran, New Zealand, Poland, Slovakia, South Africa, United Kingdom, United States of America</td>
<td>more than 20</td>
</tr>
<tr>
<td>Year of publication</td>
<td>Number of studies</td>
</tr>
<tr>
<td>1944–1953</td>
<td>61</td>
</tr>
<tr>
<td>1954–1963</td>
<td>26</td>
</tr>
<tr>
<td>1964–1973</td>
<td>25</td>
</tr>
<tr>
<td>1974–1983</td>
<td>126</td>
</tr>
<tr>
<td>1984–1993</td>
<td>408</td>
</tr>
<tr>
<td>1994–2003</td>
<td>152</td>
</tr>
</tbody>
</table>
In order to examine the dimensionality of the education-related characteristics at country level, illiteracy rate, enrollment into primary, secondary, and tertiary education, and the number of pupils per teacher were factor analyzed. A first factor with an eigenvalue of 2.83 was found to explain 56% of the variance. Illiteracy rate had a loading of −.83 on the factor, enrollment in primary education one of .09, enrollment into secondary education one of .93, enrollment into tertiary education a loading of .76, and the number of pupils per teacher a loading of −.83. The low loading of primary enrolment probably reflects the limited cross-country variability in this variable because of the universality of compulsory primary schooling. The factor covers a broad set of interrelated educational input and output factors and was labeled educational permeation.

2. Results

2.1. Descriptives

All scores were transformed from their raw mean to a 0–100 scale, depending on the number of items that were administered in the particular samples. Table 2 presents the mean scores on a single scale and the mean IQ scores by country. Visual inspection shows a large variation in country means, but no country shows any sign of a ceiling effect. Across the 798 samples, mean scores on Raven’s Progressive Matrices ranged from 10 to 97, with an overall mean of 61.88 and a standard deviation of 15.97. Standard deviations were available for 512 of the 798 samples; they ranged from 1.00 to 28.84, with a mean of 6.88 and a standard deviation of 3.09. Both chronological and educational age showed large ranges. Chronological age ranged from 3.00 to 82.50 years, with a mean of 13.94; educational age varied from 0 to 17.17 years, with a mean of 6.88 and a standard deviation of 3.09. Sex effects could not be addressed. Nine studies did not report participants’ sex, while 485 samples had some mixture of both males and females and could not be further broken down. Of the 288 remaining samples, 175 samples were entirely composed of males and 113 of females. Sex effects could not be addressed. Nine studies did not report participants’ sex, while 485 samples had some mixture of both males and females and could not be further broken down. Of the 288 remaining samples, 175 samples were entirely composed of males and 113 of females.

2.2. Initial analyses

In order to estimate the effect of country on Raven performance, a univariate ANOVA was conducted with performance as the dependent variable and country as grouping variable. The effect of country on performance is significant, $F(44, 753) = 4.79$, $p < .001$, partial $\eta^2 = .22$. Cohen (1988) proposed boundary values for small, medium, and large effects of .01, .06, and .14, respectively. The effect size observed here is thus large. In order to estimate the effect of year of publication on Raven performance, a

![Fig. 1. Performance on Raven’s Progressive Matrices plotted against year of publication.](Image 303x66 to 552x261)
performance, a univariate ANOVA with performance as the dependent variable and year of publication as the grouping variable was carried out. The effect of year of publication is significant, $F(40, 757) = 3.55, p < .001$, and large, partial $r^2 = .16$. Fig. 1 presents the pattern of performance over time. A visual inspection does not suggest a clear patterning despite the large effect size; mean performance does not look different for the 1950s than for the 1990s. 

The effect of year of publication is significant, $F(40, 757) = 3.55, p < .001$, and large, partial $r^2 = .16$. Fig. 1 presents the pattern of performance over time. A visual inspection does not suggest a clear patterning despite the large effect size; mean performance does not look different for the 1950s than for the 1990s. 

Figs. 2 and 3 visually present the change of performance on Raven’s Progressive Matrices across chronological and educational age, respectively. The relationship between chronological age and performance corresponds to what is typically found in the literature (e.g., McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; Salthouse, 1996). There is a sharp increase of performance across childhood, adolescence, and early adulthood, which is followed by a gradual decline until old age. The lower scores among the older cohorts appear to be common across the three versions, although the APM shows the strongest effect (not further documented here). To what extent this finding is due to the relative small sample of APM studies will have to remain open. Another source of the lower scores seems to be the lower educational age of older cohorts, as explored below in more detail. Fig. 3 shows the relation between educational age and performance. A positive association of test scores and educational age is clearly visible.

### 2.3. Correlational and regression analyses

Correlations between relevant sample, study, and country characteristics are presented in Table 3. Educational age correlated significantly with year of publication, $r(514) = -.19, p < .001$. The direction of the relation between educational age and year of publication is striking. More recent studies apparently sampled participants with on average lower educational ages than earlier studies. Until the age of twenty chronological age and educational age correlated almost perfectly, $r(444) = .96, p < .001$, but the correlation was (almost significantly) negative for people over the age of twenty, $r(54) = -.24, p = .08$. The lack of significance of this latter correlation is probably due to the small number of people aged older than twenty in the study samples. The relations between these characteristics and Raven performance are addressed in the next section.

### Table 3

Correlations between sample-, country-, and study-level characteristics.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Chronological age $\leq$ 20 yr</th>
<th>Chronological age $&gt; 20$ yr</th>
<th>Educational age</th>
<th>GNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational age</td>
<td>.96***</td>
<td>-.24</td>
<td>.04</td>
<td>.16</td>
</tr>
<tr>
<td>Country characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gross National Product</td>
<td>-.01</td>
<td>-.00</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>Educational permeation</td>
<td>.05</td>
<td>.10</td>
<td>.03</td>
<td>.79***</td>
</tr>
<tr>
<td>Study characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year of publication</td>
<td>-.13**</td>
<td>.21**</td>
<td>-.19***</td>
<td>.08*</td>
</tr>
</tbody>
</table>

*p < .05. ** p < .01. *** p < .001.

### Table 4

Correlations between performance on Raven’s Progressive Matrices and sample-, country-, and study-level characteristics.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample characteristics</td>
<td></td>
</tr>
<tr>
<td>Chronological age $\leq$ 20 yr</td>
<td>.58***</td>
</tr>
<tr>
<td>Chronological age $&gt; 20$ yr</td>
<td>-.14</td>
</tr>
<tr>
<td>Educational age</td>
<td>.56***</td>
</tr>
<tr>
<td>Country characteristics</td>
<td></td>
</tr>
<tr>
<td>Gross National Product</td>
<td>.16***</td>
</tr>
<tr>
<td>Educational permeation</td>
<td>.25***</td>
</tr>
<tr>
<td>Study characteristics</td>
<td></td>
</tr>
<tr>
<td>Year of publication (original)</td>
<td>.07</td>
</tr>
<tr>
<td>Year of publication (partial)$^b$</td>
<td>.22***</td>
</tr>
</tbody>
</table>

$^a$ All scores were transformed from their raw mean to a 0–100 scale and then averaged across the Advanced, Colored, and Standard Versions.

$^b$ The partial correlation between year of publication and performance was corrected for educational age and educational permeation.
The Flynn effect would be observed if test performance and year of publication are positively associated. As can be seen in Fig. 1, the relation is weak (though marginally significant), $r(798) = .07, p = .05$. The weakness of the relation could be a consequence of moderators not accounted for. More specifically, the educational situation of samples may be crucial, both in terms of participants’ educational age as in terms of countries’ educational permeation. Educational age was positively related to performance, but as shown in Table 3, educational age was negatively related to year of publication. When educational age and educational permeation were included as control variables in the estimation, the correlation between Raven performance and year of publication became $-.22 (p < .001)$. Thus, after controlling for sample- and country-related educational characteristics, we observed the expected Flynn effect in performance on Raven’s Progressive Matrices.

The importance of sample and country characteristics in moderating the Flynn effect is further underscored in a regression analysis. Performance on the Raven was the dependent variable, while year of publication, educational age, and educational permeation were predictors. The proportion of explained variance in performance is large, $R^2 = .41, p < .001$. The relation between educational age and performance is strong and significant ($\beta = .59, p < .001$). Educational permeation has a somewhat smaller effect on performance, but the effect is still significant ($\beta = .26, p < .001$). A small, though salient Flynn effect can be derived from the positive relation between performance and year of publication ($\beta = .18, p < .001$). When converted to IQ points, this effect corresponds to an increase in IQ of 2.01 points per decade. The regression analysis demonstrates that while the zero-order correlations of our predictors with Raven performance are not significant, the regression coefficients (which might be viewed as partial correlations) are significant.

The regression analysis implicitly assumes the universality of the Flynn effect. A final analysis addressed this assumption in more detail by testing the presence of the Flynn effect in individual countries. A country was included in the analysis if data from this country met three criteria: The country should be present in the dataset with at least 20 samples; data of the country should be collected on at least two independent occasions; data of the country should have a minimum dispersion of 14 years from the earliest to the latest occasion. Eight out of the 45 countries in the dataset met all criteria (namely Australia, Canada, the former West-Germany, India, Iran, Poland, United Kingdom, and the United States). For each country, a separate regression analysis was conducted with year of publication and educational age as predictors and the Raven score as dependent variable. Results are presented in Table 5.

Canada showed a significantly negative regression coefficient for year of publication. This might signify a reversed Flynn effect. The United Kingdom was the only affluent country with a salient Flynn effect. The largest Flynn effects were found in India, Iran, and Poland. A closer examination of the raw country means confirmed that variation in the size of the Flynn effect is not caused by ceiling effects in the data, indicating that the present findings resemble actual variations in the Flynn effect. The size of the Flynn effect showed a significantly negative correlation with the Gross National Product of the country, $r(8) = -.74, p < .05$. Unfortunately, countries from Africa and South America were measured only once and hence, we do not know whether the negative correlation extends to developing countries. It may be concluded that our data suggest a temporal patterning in the Flynn effect. The effect was first observed in Western countries, but here it seems to have reached its ceiling. Countries with a lower though increasing level of economic development show a more pronounced Flynn effect.

### 3. Discussion

We examined the associations between performance on Raven’s Progressive Matrices with various education-related country characteristics and year of publication. A total of 193 publications were included in our meta-analysis, which contained 798 independent samples from 45 countries and covered a period of 60 years. This considerable variation in countries and years of publication is crucial for testing the cross-cultural generalizability of the Flynn effect. A number of results emerged that would not have been evident when looking at the Flynn-effect as an isolated measure of individual differences.

There were two results that carry important conceptual implications. First, the regression analysis showed that year of publication has a relation with Raven performance independent of individuals’ educational age and countries’ educational permeation; Raven performance increases by 2.01 IQ points per decade. Moreover, educational age was the best predictor of Raven performance. These analyses suggest that The Flynn effect is not an artifact of the on average higher levels of education in countries where the economy is growing (that tend to invest more and more in education). The current study suggests that an increase in Raven performance is independently associated with three factors: educational permeation, educational age, and publication year. Two of these factors, educational age and educational permeation, will often act in concert: economic growth over an extended period will often lead to more educational permeation and to an increase of the average educational age of a population. If the Flynn effect would be observed in a country with a substantial economic development in the period of observation, the size of the Flynn effect may have been boosted by that economic development. This pattern of results suggests a more complex relationship between intelligence and wealth at country level than suggested by Lynn and Vanhanen (2002, 2006) and shows that explaining this relationship requires much caution (Hunt & Wittmann, 2008).

Second, the Flynn effect seems to be present in all countries represented in our meta-analysis, with variation in the effect confined to its size; yet, the generalizability of this second finding requires closer examination. One question that emerges after our analysis is whether our data set includes sufficient temporal and cross-cultural variation in order to assert the universality of the Flynn effect. It could be argued that a sample of 45 countries is sizeable; however, the cultural variation in the sample is not optimal. An inspection of Table 1 suggests that affluent Western countries and developing countries are overrepresented and it is only for some, mainly Western, countries that a sizeable variation in years of publication is available. As a consequence, one could argue that variability in our data set is limited. Still, our data suggest that Flynn effect is not linked to Western societies alone and is independent of individual-level and country-level education-related factors.

We found that the size of the Flynn effect is related to country affluence, with more affluent countries showing a smaller IQ increase. These findings suggest that the Flynn effect is a function of earlier levels of performance, in which new elements of information connect with already available elements of information. This finding has implications for current views on cross-cultural differences in abstract...
thinking. Researchers tend to employ a distinction between information and processor when interpreting cross-cultural differences in intelligence scores. Information is seen by various researchers as the raw material that feeds in the mental processor (e.g., Luria, 1976; Rindermann, 2007). Alternatively, some consider information to constitute the stimulus that motivates access to the mental construct of abstract thinking (e.g., Ceci, 1991; Van de Vijver, 2002). The present findings question the validity of a sharp distinction between information and processor when interpreting cross-cultural differences in intelligence scores, since cross-cultural differences are confined neither to the information, nor to the processor. The finding of a gradual decline of the Flynn effect with increased affluence is more compatible with a view of a cognitive system in which new information builds on existing knowledge and procedures already available than with a view in which either the information or processor capacity create the Flynn effect.

The present findings suggest that pervasive cognitive variability is best thought of in terms of changing distributions of the ways in which people approach a problem, rather than stable differences between individuals or between cultures (Siegler, 1994). Each Raven item really is a task of inductive reasoning, for every individual after a certain age, but the method, strategies, and heuristics that people use in order to solve a problem is known to change from situation to situation, even for the same individual (Kahnebens, Slovic, & Tversky, 1982; Siegler, 1994).

References

Blijsnzer, R., Willis, S. L., & Baltes, P. B. (1981). Training research in aging on the intelligence scores, since cross-cultural differences are constitutes the stimulus that motivates access to the mental construct and processor when interpreting cross-cultural differences in


Studies used in the meta-analysis


Cortical glucose metabolic rate correlates of abstract reasoning and attention for psychometric tests of g.


Klauer, K. J. (1997). Läßt sich die Strategie des induktiven Denkens auf schulisches Lernen transferieren? [Can the strategy to reason inductively be taught such that it transfers to learning of school-type material?] Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 29, 225–241.


